Does the rich demand more large houses:

Evidence from housing sales in King county, US

Group III: Zhiyu Chen, Wenxin Feng, Yuxin Zhang

1. Introduction

In real estate business, house pricing constitutes an essential part and stirs the dynamics among buyers, sellers and brokers. Key attributes such as square footage of the living space, year when the house was built, the overall house condition are widely acknowledged by the market. However, how house price per square foot adjusts in sync with the increase in square footage across groups with different income levels remains a puzzle to us.

This paper uses data from "House Sales in King County, USA" and endeavors to investigate the question of our interest. Commonly applied Hedonic regression is incorporated as the theoretical foundation of preliminary model specification. The pricing model is constructed by house characteristics of 1) sqft_living15 2) bedrooms 3) bathrooms 4) sqft_lot15 5) floors 6) waterfront 7) sqft_above 8) sqft_renoliving 9) time_sold 10) time_renovated 11) up 12) middle 13) up_sqft_living 14) mid_sqft_living 15) grade. Detailed explanations are provided later.

Methodology of this research paper covers data processing, model specification and diagnostics, multivariable regression, hypothesis testing.

1.1 Research Question

Linking people's different income levels to the per unit housing price, it's natural to presume that people from higher classes are more willing to pay for bigger houses.

Based on this presumption, our research question narrows down to the point of answering: when the square footage of total house living space increases by unit, does the unit price of house per square foot change more in upper class regions than that in middle and lower class regions?

By delving into this question, we hope to shed light on the demand and supply relationship of houses for people from different income classes. The point being, we would like to seek for certain logics behind the market dynamic for regular homebuyers to reference. Real estate market involves various dynamics and is oftentimes found to be information asymmetric for outsider buyers. Albeit they may bargain with sellers or brokers, regular buyers tend to be put in a passive situation and end up price-takers. Meanwhile, for real estate developers, this question may also contribute to understand the profit margin from square footage of houses. In this sense, they can align different sizes of houses with their target clients from specific income class.

1.2 Literature Review

House pricing entails cautious examination upon the background of the research question to provide a reasonable and appropriate model. Invoking the discourse Abdulai and Awusu-

Ansah have covered in their paper "House Price Determinants in Liverpool, United Kingdom", we can draw on hedonic pricing model to better build the house pricing model. As these two authors state, "[h]ousing as a heterogeneous good can be viewed as a package of inherent characteristics...in the housing sector, the hedonic pricing model is applied to estimate the marginal contribution of each property and neighbourhood characteristic to the house price" (5-6). Key attributes, namely independent variables in our regression model, represent various characteristics of housing given available data set. From regressing unit price of housing on these key attributes, Abdulai and Awusu-Ansah believes the coefficients reflect "the willingness to pay for the attributes...[and it's the willingness that] determines housing price" (6). This argument allows our research to break down the house pricing model into feasible components for empirical study purpose.

The legitimacy of using unit price of housing per square foot is supported by recent behavioral finance findings. Christopher J Mayer suggests in his paper "US Housing price dynamics and behavior finance" that what really concerns people when buying houses is not the total price but the unit housing price per square foot, based on which many judge "if the house is at a perfect price". We think this finding also helps real estate developers when they are looking into new housing projects. It's natural and easy to think that the unit housing price in rich regions is naturally higher than that in poorer areas regardless of the living square. This is due to the higher benchmark unit price, which is income level related. But this is does not guarantee the profitability of building large houses in rich areas because it is also very intuitive to think that it costs higher to build per unit square in rich areas. But the unit cost is usually considered fixed. That is to say, it's actually critical for the constructors to consider the size of the house to build as he wants as high marginal profit per unit square as possible. If they find that building larger square houses results in higher marginal unit price, they would tend to build larger houses in that region. That's why the empirical analysis of how income levels have impact on per living square influence of the unit housing price.

1.3 Variables

I. **Dependent variable:** unitprice = price / sqft_living

In light of the preceding, we identify house price per square foot as the dependent variable.

Fig. 1 denotes the distribution of $y_1$as well as the bell curve distribution of the full set of observations. The data appears to be skewed to right with skewness 2.364101. The distribution of the dependant variable shows characteristics of dispersion, which indicates the randomness in data selection procedure.

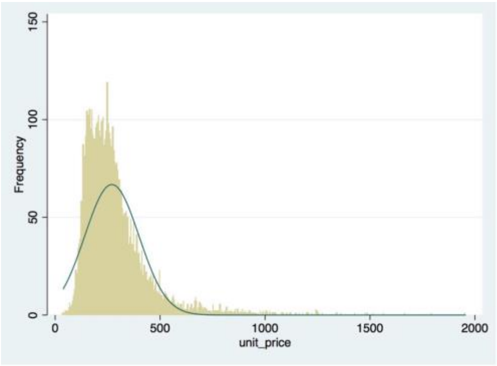


Figure 1

II. **Independent variables drawn from data set**

A. Incorporated non-dummy variables

- bedrooms: number of bedrooms per house

This variable affects unit house price towards a direction that still remains unknown to us. Previous research work has shown that different buyer groups respond to increase in bedrooms differently, in a either positive, negative, or neutral way. In one way, more bedrooms indicates large house, which may add to the per unit price. In another way, more bedrooms may indicates smaller free space, which may decrease the per unit price.

- bathrooms: number of total bathrooms in house
  By the same token as bedrooms, this variable affects unit house price in a possible similar mechanism.

- floors: total floors(levels) in house
  Total floors in house often suggest more square footage and therefore affect the unit house price.

- grade: overall grade given to the housing unit, based on King County grading system
  Different grades suggest various house conditions and therefore affect unit house price. Also, a systematic grading standard from King County serves better in this research.

- sqft_above: square footage of house apart from basement (in square foot)
  Square footage of house apart from basement suggest the pragmatic living space for family within house. Thus, this variable is incorporated to better research result.

- sqft_living15: living room area of the house in 2015 (renovation implied)
  The change in living room area of the house by the time house was sold puts weight on the current condition of house, and therefore affects unit house price.

- sqft_lot15: lot size area of the house in 2015 (renovation implied)
  The change in lot size area of the house by the time house was sold puts weight on the current condition of house, and therefore affects unit house price.

B. Dropped variables

- ~~sqft_living~~: square footage of the house (in square foot)
  This variable reflects the original square footage of the house before any renovation; thus, we assume the actual effect of square footage to be more correlated with variable sqft_living15. This variable is dropped from the model.

- ~~sqft_lot~~: square footage of the lot (in square foot)
  Same as variable sqft_living

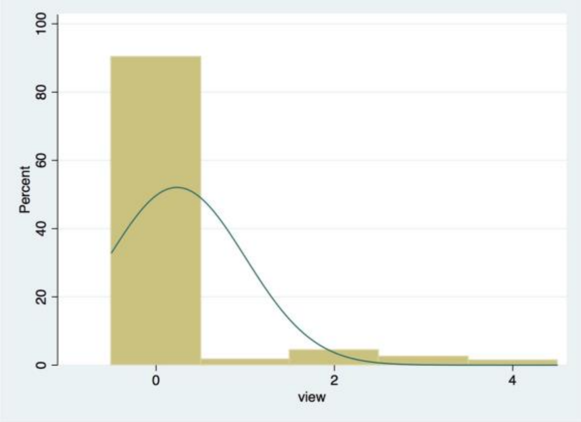


Figure 2

- ~~view~~: number of times that the house has been viewed by potential buyers
  Only approximately 10% out of total data contains information in this category. Fig. 2 is the distribution graph of it. This variable has a narrow distribution with value ranging from 1 to 4, meaning less significant to our research.

- ~~condition~~: how good the condition is on the whole
  This variable represents different overall conditions of house with relatively arbitrary terms. Information overlap also occurs since the following variable grade also conveys the message of house conditions and is measured in a more systematic way. The grade variable is more disperse and rated in detailed. We therefore choose to drop condition variable.

- ~~yr_built~~: the year when the house was built
  The age of house indicates the condition of house to certain extent. However, straightforward year number does not mean much. We're more interested to figure out the time gap between when the house was built and when the house was sold, instead of directly including this variable in the model. Therefore, yr_built is dropped from the model.

- ~~yr_renovated~~: the year when the house was renovated
  Same as variable yr_built.

- ~~sqft_basement~~: square footage of the basement (in square foot)
  This variable is removed from the model because of multicollinearity problem.

- ~~zipcode~~: zip code to the region that the house is located
  Information contained in zipcode cannot be directly retrieved. Therefore our treatment accordingly is to translate zipcode number to corresponding regional income level. Detailed process will be discussed later in 1.3.III

- ~~lat~~: latitude coordinate to the house location
  Precise location coordinates overhang the objectives of this research. This empirical study is more interested to learn the income level across different areas, for which the information can be derived from zip codes. Hence, the variable lat is dropped from the model.

- ~~long~~: longitude coordinate to the house location
  Same as latitude coordinate.

II. Variables derived from existing data

- time_sold: yr_sold - yr_built (in years)
  This variable measures how long it is between the year when the house was built and when it was sold. This variable definitely has influence on the per unit price but in an unclear way. On the one hand, new houses with better decoration might have higher per unit price. On the other hand, some old houses are as valuable, if not more.

- time_renovated: yr_sold - yr_renovated (in years)
  This variable measures how long it is between the year when the house was renovated and when it was sold. Usually, the more recent the renovation happened, the higher per unit price it can be sold at.

- sqft_renoliving: sqft_living15 - sqft_living (in square foot)
  This variable measures how many square foot of the house living space has changed after renovation.

- sqft_renolot: sqft_lot15 - sqft_lot (in square foot)
  This variable measures how many square foot of the house lot has changed after renovation.

III. Dummy variables

- up: for group from upper class with high income
  - Take value of 1 if they come from upper class and 0 if not

- middle: for group from middle class with medium income
  - Take value of 1 if they come from middle class and 0 if not

- waterfront: house which has a view to a waterfront
  - Take value of 1 if the house has view to waterfront and 0 if not

  View to a waterfront brings add-value of surroundings to the house and therefore affect the house price.

*** How we obtain the categorical data about the average income class?**

According to the Federal data, if we categorize the whole population in American into three classes based on their income, we have 20% population in Upper class, 51% population in the middle class, and 29% population in the lower population (Fig. 3).

In the housing price data set, we have the prices of the houses sold from May 2014 to May 2015 from King county in washington state in America. If we pose the assumption that the class components in King county the same as it is nationwide, we can classify different income groups by ranking the median income of each cities in King county. Hence, we have the top 20 percentile as the upper class, 20 to 71 percentile as the middle class, and the rest as the lower class based on their income. And because of the regional features of different classes, we can mainly use the zipcode to tag each house sold based on the income class they come from. Fig. 4 gives us information of median household income in King County.

<u>Note: assumptions we make in this data sorting procedure</u>

1. The income distribution in these cities follows an asymptotic pattern of the national income distribution.
2. The population distribution on the city level does not vary much from each other.

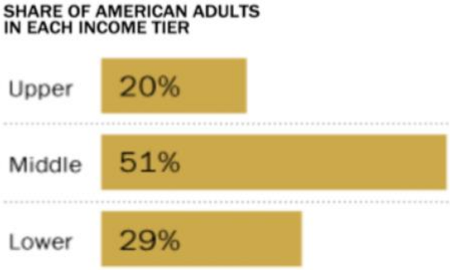


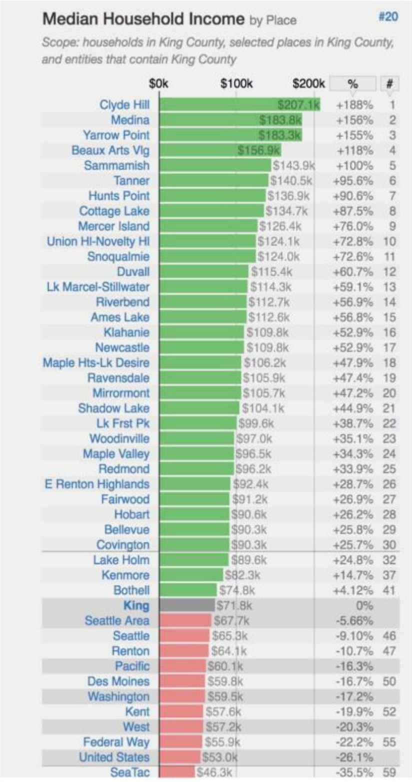
Figure 3

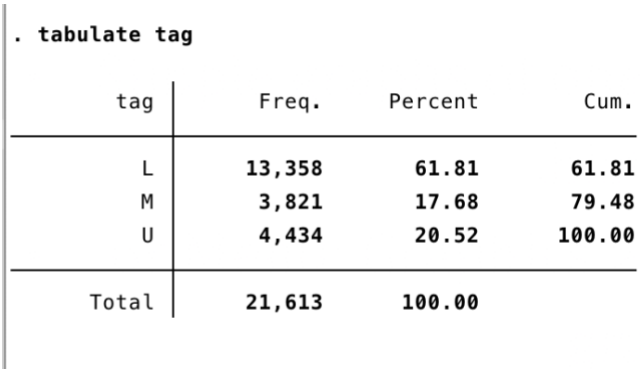
*** Findings and explanations of income categorization:**

After labelling each house by the tag according to the income level revealed from the zipcode, we check if we have enough data for each group to study. To do this, we derive the frequency table in STATA.

The most astonishing and apparent finding is that more houses are sold in the low income regions in between May 2014 and May 2015, and it shows not just in numbers, but more astonishingly in the percentage considering it's smaller portion in the average income distribution.

Figure 4





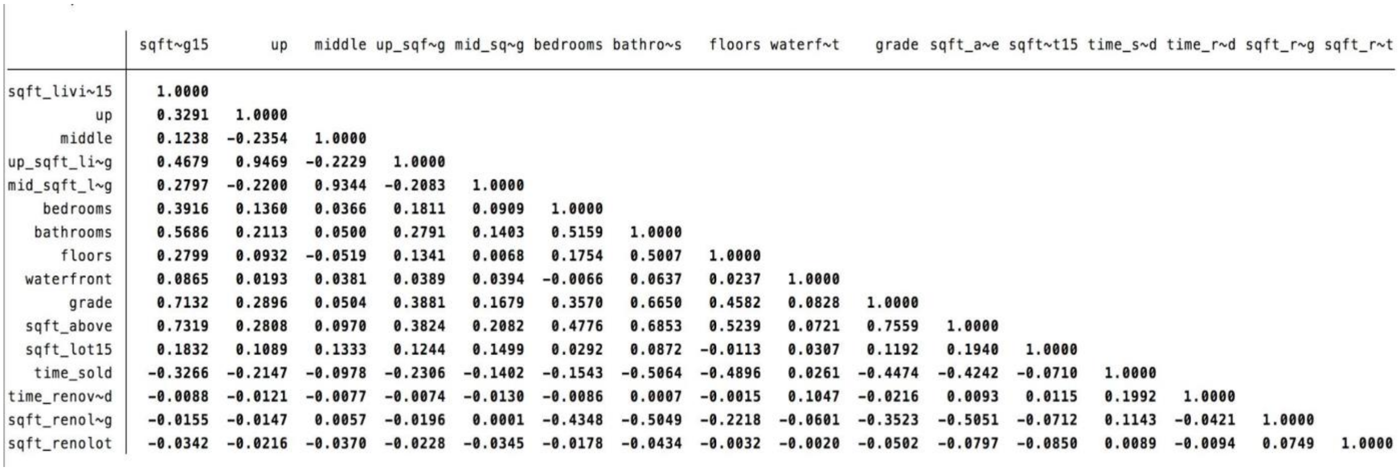
```
. tabulate tag

        tag |      Freq.     Percent        Cum.
------------+-----------------------------------
          L |     13,358       61.81       61.81
          M |      3,821       17.68       79.48
          U |      4,434       20.52      100.00
------------+-----------------------------------
      Total |     21,613      100.00
```

One explanation might be that poor people might suffer more from the economic crisis in 2008, which almost crashed the housing market in America. During that time, many people default on their mortgages, and the ability to pay back borrowed from the bank is low, especially for those with lower average income. Because of that, in the after wave of the economic crisis, those houses would go under frequent transactions in between the buyers. Another reason is that rich people from other regions, they make use of the low interest rate advantage supported by the government to settle the market to borrow money from the bank and buy houses at a relatively lower price in expectation of the housing market coming back in the future. And they would target those houses in poorer regions owners of which were in default situation and would like to sell at a lower price.

IV. Dummy interaction forms

- up_sqft_living = up (dummy) * sqft_living15
  - A measurement for per living square at upper income groups' extra contribution to the housing price per square foot
  - Base category: lower income class
- mid_sqft_living = middle (dummy) * sqft_living
  - A measurement for per living square at middle income groups' extra contribution to the housing price per square foot
  - Base category: lower income class

* up_sqft_living, mid_sqft_living would be the variables under major study to critically give inference of our research question, since they measure the class difference as for the contribution of per square foot living for the unit housing price.

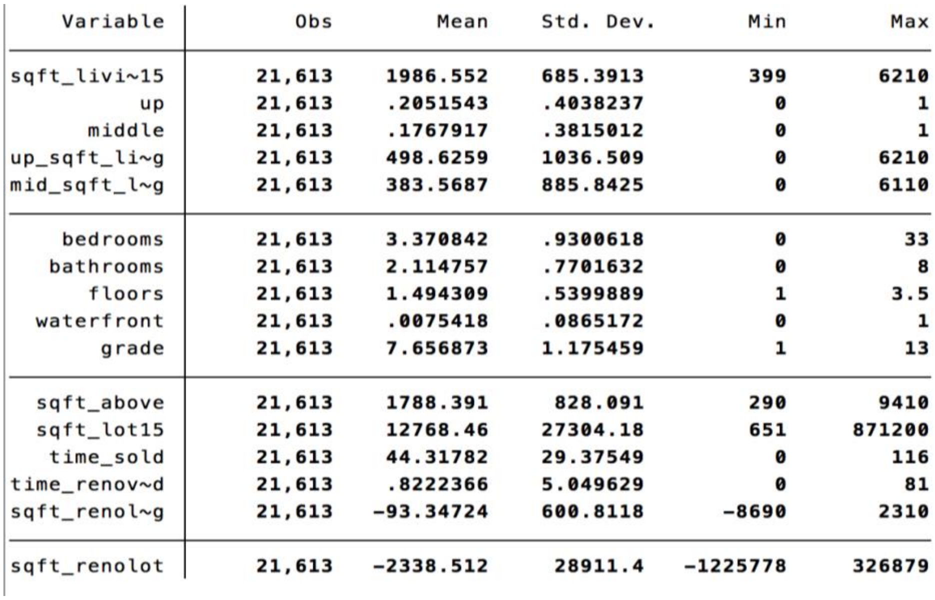


```
             | sqft~g15       up   middle up_sqf~g mid_sq~g bedrooms bathro~s   floors waterf~t    grade sqft_a~e sqft~t15 time_s~d time_r~d sqft_r~g sqft_r~t
-------------+------------------------------------------------------------------------------------------------------------------------------------------------
sqft_livi~15 |   1.0000
          up |   0.3291   1.0000
      middle |   0.1238  -0.2354   1.0000
up_sqft_li~g |   0.4679   0.9469  -0.2229   1.0000
mid_sqft_l~g |   0.2797  -0.2200   0.9344  -0.2083   1.0000
    bedrooms |   0.3916   0.1360   0.0366   0.1811   0.0909   1.0000
   bathrooms |   0.5686   0.2113   0.0500   0.2791   0.1403   0.5159   1.0000
      floors |   0.2799   0.0932  -0.0519   0.1341   0.0068   0.1754   0.5007   1.0000
  waterfront |   0.0865   0.0193   0.0381   0.0389   0.0394  -0.0066   0.0637   0.0237   1.0000
       grade |   0.7132   0.2896   0.0504   0.3881   0.1679   0.3570   0.6650   0.4582   0.0828   1.0000
  sqft_above |   0.7319   0.2808   0.0970   0.3824   0.2082   0.4776   0.6853   0.5239   0.0721   0.7559   1.0000
  sqft_lot15 |   0.1832   0.1089   0.1333   0.1244   0.1499   0.0292   0.0872  -0.0113   0.0307   0.1192   0.1940   1.0000
   time_sold |  -0.3266  -0.2147  -0.0978  -0.2306  -0.1402  -0.1543  -0.5064  -0.4896   0.0261  -0.4474  -0.4242  -0.0710   1.0000
time_renov~d |  -0.0088  -0.0121  -0.0077  -0.0074  -0.0130  -0.0086   0.0007  -0.0015   0.1047  -0.0216   0.0093   0.0115   0.1992   1.0000
sqft_renol~g |  -0.0155  -0.0147   0.0057  -0.0196   0.0001  -0.4348  -0.5049  -0.2218  -0.0601  -0.3523  -0.5051  -0.0712   0.1143  -0.0421   1.0000
sqft_renolot |  -0.0342  -0.0216  -0.0370  -0.0228  -0.0345  -0.0178  -0.0434  -0.0032  -0.0020  -0.0502  -0.0797  -0.0850   0.0089  -0.0094   0.0749   1.0000
```

IV. Variables correlation matrix

1.4 Basic Descriptive Statistics

Table 1: summary statistics of basic independent variable



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
sqft_livi~15 |     21,613    1986.552    685.3913        399       6210
          up |     21,613    .2051543    .4038237          0          1
      middle |     21,613    .1767917    .3815012          0          1
up_sqft_li~g |     21,613    498.6259    1036.509          0       6210
mid_sqft_l~g |     21,613    383.5687    885.8425          0       6110
-------------+---------------------------------------------------------
    bedrooms |     21,613    3.370842    .9300618          0         33
   bathrooms |     21,613    2.114757    .7701632          0          8
      floors |     21,613    1.494309    .5399889          1        3.5
  waterfront |     21,613    .0075418    .0865172          0          1
       grade |     21,613    7.656873    1.175459          1         13
-------------+---------------------------------------------------------
  sqft_above |     21,613    1788.391     828.091        290       9410
  sqft_lot15 |     21,613    12768.46    27304.18        651     871200
   time_sold |     21,613    44.31782    29.37549          0        116
time_renov~d |     21,613    .8222366    5.049629          0         81
sqft_renol~g |     21,613   -93.34724    600.8118      -8690       2310
-------------+---------------------------------------------------------
sqft_renolot |     21,613   -2338.512     28911.4   -1225778     326879
```

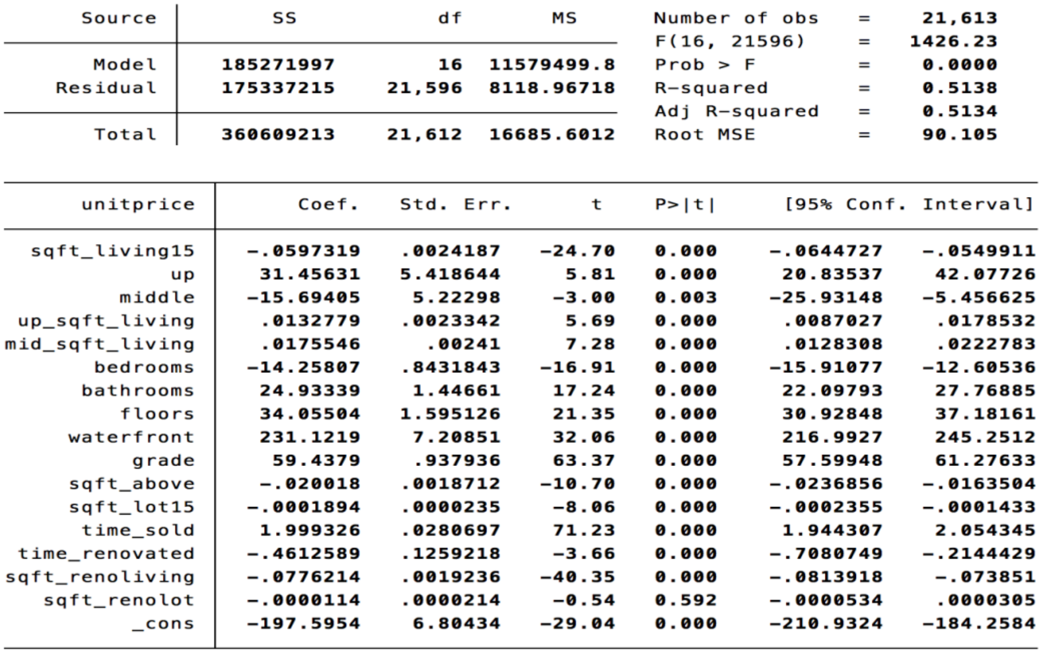
2. Inferential analysis

2.1 Model specification

2.1.1 Preliminary model construction

*** Preliminary regression model based on economic reasoning**

$$\begin{aligned} unitprice = {} & \beta_0 + \beta_1\text{sqft\_living15} + \beta_2\text{bedrooms} + \beta_3\text{bathrooms} + \beta_4\text{floors} \\ & + \beta_5\text{sqft\_above} + \beta_6\text{sqft\_lot15} + \beta_7\text{time\_sold} + \beta_8\text{time\_renovated} \\ & + \beta_9\text{sqft\_renoliving} + \beta_{10}\text{sqft\_renolot} + \beta_{11}\text{grade} + \delta_0\text{up} + \delta_1\text{middle} \\ & + \delta_2\text{up\_sqft\_living} + \delta_3\text{mid\_sqft\_living} + \delta_4\text{waterfront} + u \end{aligned}$$

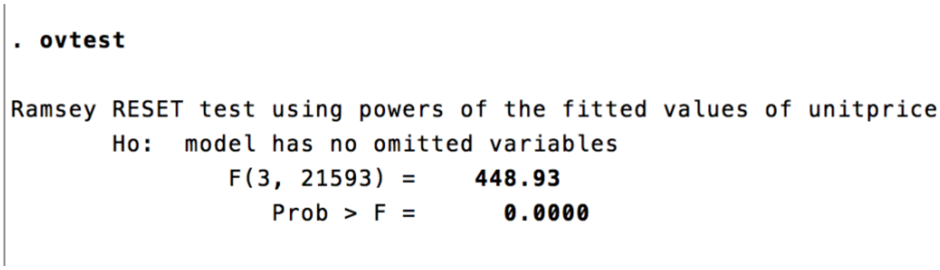


```
      Source |       SS           df       MS      Number of obs   =    21,613
-------------+----------------------------------   F(16, 21596)    =   1426.23
       Model |   185271997        16  11579499.8   Prob > F        =    0.0000
    Residual |   175337215    21,596  8118.96718   R-squared       =    0.5138
-------------+----------------------------------   Adj R-squared   =    0.5134
       Total |   360609213    21,612  16685.6012   Root MSE        =    90.105

---------------------------------------------------------------------------------
      unitprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
   sqft_living15 |  -.0597319   .0024187   -24.70   0.000    -.0644727   -.0549911
              up |   31.45631   5.418644     5.81   0.000     20.83537    42.07726
          middle |  -15.69405    5.22298    -3.00   0.003    -25.93148   -5.456625
  up_sqft_living |   .0132779   .0023342     5.69   0.000     .0087027    .0178532
 mid_sqft_living |   .0175546     .00241     7.28   0.000     .0128308    .0222783
        bedrooms |  -14.25807   .8431843   -16.91   0.000    -15.91077   -12.60536
       bathrooms |   24.93339    1.44661    17.24   0.000     22.09793    27.76885
          floors |   34.05504   1.595126    21.35   0.000     30.92848    37.18161
      waterfront |   231.1219    7.20851    32.06   0.000     216.9927    245.2512
           grade |    59.4379    .937936    63.37   0.000     57.59948    61.27633
      sqft_above |   -.020018   .0018712   -10.70   0.000    -.0236856   -.0163504
      sqft_lot15 |  -.0001894   .0000235    -8.06   0.000    -.0002355   -.0001433
       time_sold |   1.999326   .0280697    71.23   0.000     1.944307    2.054345
  time_renovated |  -.4612589   .1259218    -3.66   0.000    -.7080749   -.2144429
  sqft_renoliving |  -.0776214   .0019236   -40.35   0.000    -.0813918    -.073851
    sqft_renolot |  -.0000114   .0000214    -0.54   0.592    -.0000534    .0000305
           _cons |  -197.5954    6.80434   -29.04   0.000    -210.9324   -184.2584
---------------------------------------------------------------------------------
```

If we regress unitprice on sqft_living15 up middle up_sqft_living mid_sqft_living bedrooms bathrooms floors waterfront grade sqft_above yr_renovated sqft_lot15 time_sold time_renovated sqft_renoliving sqft_renolot, we get the following estimated coefficient with their joint and separate degree of statistical significance:

2.1.2 Omitted variable test

I. Below is the ov test result of our preliminary model:

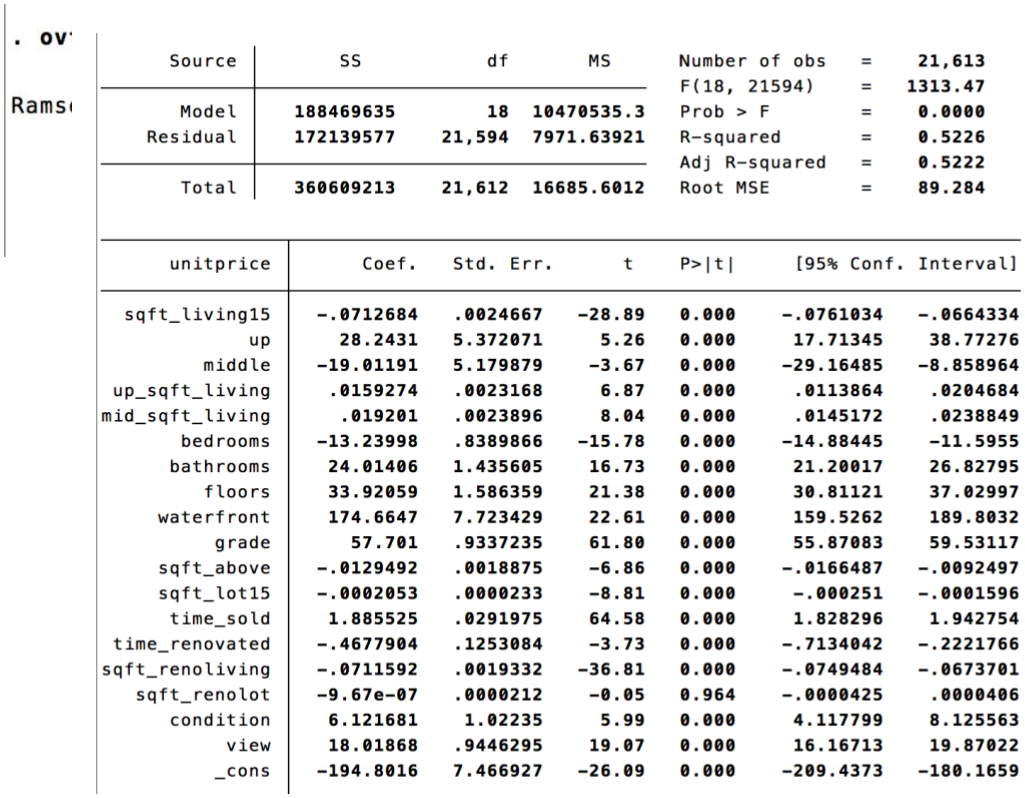


```
. ovtest

Ramsey RESET test using powers of the fitted values of unitprice
       Ho:  model has no omitted variables
                F(3, 21593) =    448.93
                   Prob > F =      0.0000
```

Given the 95% significance level, p-value = 0.0000 < 0.005. Thus, we reject $H_0$and acknowledge that there exists omitted variable problem in the preliminary regression model.

II. To refine the model, our first attempt is to include more independent variables that we dropped in the previous reasoning

Due to multicollinearity and , we can only add 2 more variables *view*, *condition* into the regression model. The regression result is displayed as follows. However, in ov test, we reject H_0as well:

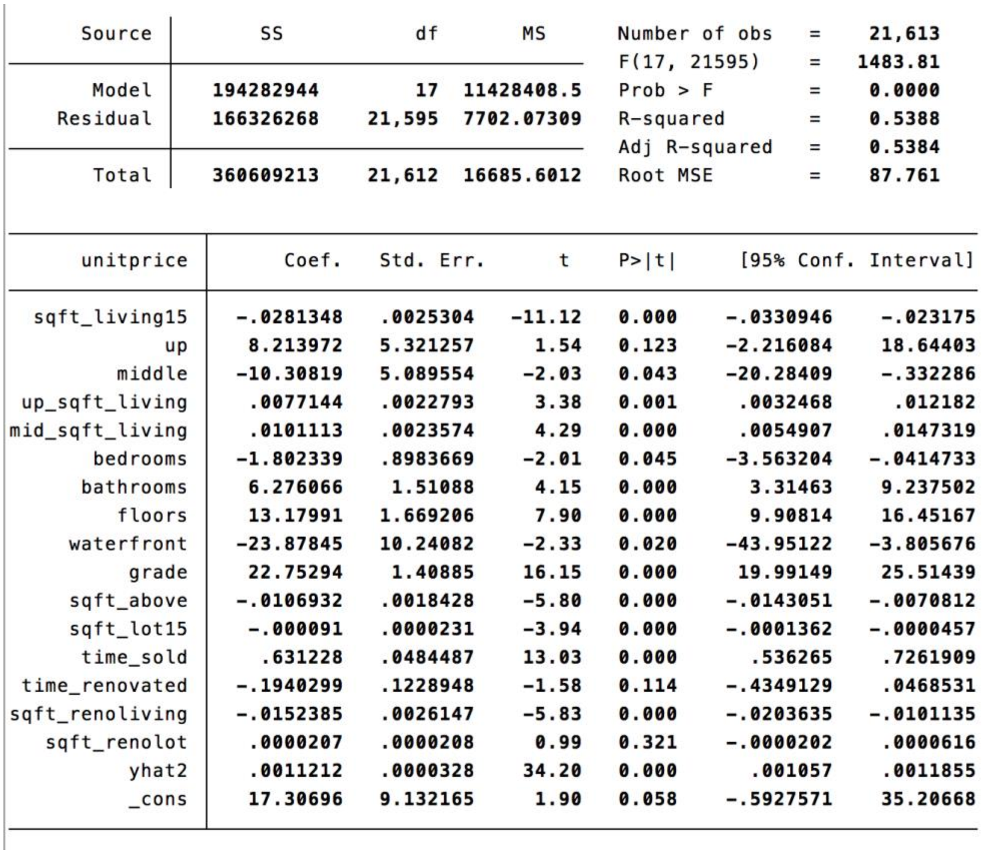


```
. ov

Rams

      Source |       SS           df       MS      Number of obs   =    21,613
-------------+----------------------------------   F(18, 21594)    =   1313.47
       Model |   188469635        18  10470535.3   Prob > F        =    0.0000
    Residual |   172139577    21,594  7971.63921   R-squared       =    0.5226
-------------+----------------------------------   Adj R-squared   =    0.5222
       Total |   360609213    21,612  16685.6012   Root MSE        =    89.284

-------------------------------------------------------------------------------
      unitprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+--------------------------------------------------------------
  sqft_living15 |  -.0712684   .0024667   -28.89   0.000    -.0761034   -.0664334
             up |    28.2431   5.372071     5.26   0.000     17.71345    38.77276
         middle |  -19.01191   5.179879    -3.67   0.000    -29.16485   -8.858964
 up_sqft_living |   .0159274   .0023168     6.87   0.000     .0113864    .0204684
mid_sqft_living |    .019201   .0023896     8.04   0.000     .0145172    .0238849
       bedrooms |  -13.23998   .8389866   -15.78   0.000    -14.88445    -11.5955
      bathrooms |   24.01406   1.435605    16.73   0.000     21.20017    26.82795
         floors |   33.92059   1.586359    21.38   0.000     30.81121    37.02997
     waterfront |   174.6647   7.723429    22.61   0.000     159.5262    189.8032
          grade |     57.701   .9337235    61.80   0.000     55.87083    59.53117
     sqft_above |  -.0129492   .0018875    -6.86   0.000    -.0166487   -.0092497
     sqft_lot15 |  -.0002053   .0000233    -8.81   0.000     -.000251   -.0001596
      time_sold |   1.885525   .0291975    64.58   0.000     1.828296    1.942754
 time_renovated |  -.4677904   .1253084    -3.73   0.000    -.7134042   -.2221766
 sqft_renoliving |  -.0711592   .0019332   -36.81   0.000    -.0749484   -.0673701
   sqft_renolot |  -9.67e-07   .0000212    -0.05   0.964    -.0000425    .0000406
      condition |   6.121681    1.02235     5.99   0.000     4.117799    8.125563
           view |   18.01868   .9446295    19.07   0.000     16.16713    19.87022
          _cons |  -194.8016   7.466927   -26.09   0.000    -209.4373   -180.1659
-------------------------------------------------------------------------------
```

Therefore, adding these two variables does not fix the omitted variable problem.

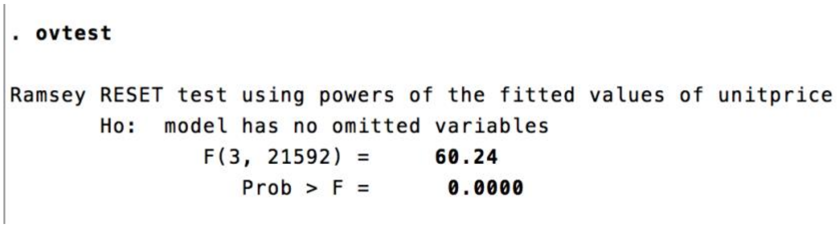
III. Our second attempt is to include the fitted value yhat squared (obtained from regressing the preliminary model) as another regressor in the regression model to test the significance of any quadratic terms or interactions

Following is the regression result:



```
      Source |       SS           df       MS      Number of obs   =    21,613
-------------+----------------------------------   F(17, 21595)    =   1483.81
       Model |  194282944        17  11428408.5   Prob > F        =    0.0000
    Residual |  166326268    21,595  7702.07309   R-squared       =    0.5388
-------------+----------------------------------   Adj R-squared   =    0.5384
       Total |  360609213    21,612  16685.6012   Root MSE        =    87.761

------------------------------------------------------------------------------
   unitprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sqft_living15 |  -.0281348   .0025304   -11.12   0.000    -.0330946    -.023175
          up |   8.213972   5.321257     1.54   0.123    -2.216084    18.64403
      middle |  -10.30819   5.089554    -2.03   0.043    -20.28409    -.332286
up_sqft_living |   .0077144   .0022793     3.38   0.001     .0032468     .012182
mid_sqft_living |   .0101113   .0023574     4.29   0.000     .0054907    .0147319
    bedrooms |  -1.802339   .8983669    -2.01   0.045    -3.563204   -.0414733
   bathrooms |   6.276066    1.51088     4.15   0.000      3.31463    9.237502
      floors |   13.17991   1.669206     7.90   0.000      9.90814    16.45167
  waterfront |  -23.87845   10.24082    -2.33   0.020    -43.95122   -3.805676
       grade |   22.75294    1.40885    16.15   0.000     19.99149    25.51439
  sqft_above |  -.0106932   .0018428    -5.80   0.000    -.0143051   -.0070812
  sqft_lot15 |   -.000091   .0000231    -3.94   0.000    -.0001362   -.0000457
   time_sold |    .631228   .0484487    13.03   0.000      .536265    .7261909
time_renovated |  -.1940299   .1228948    -1.58   0.114    -.4349129    .0468531
sqft_renoliving |  -.0152385   .0026147    -5.83   0.000    -.0203635   -.0101135
 sqft_renolot |   .0000207   .0000208     0.99   0.321    -.0000202    .0000616
       yhat2 |   .0011212   .0000328    34.20   0.000      .001057    .0011855
       _cons |   17.30696   9.132165     1.90   0.058    -.5927571    35.20668
------------------------------------------------------------------------------
```

Nevertheless, we have to reject $H_o$in ov test and conclude there still exists omitted variable problem in the regression model:



```
. ovtest

Ramsey RESET test using powers of the fitted values of unitprice
       Ho:  model has no omitted variables
                F(3, 21592) =     60.24
                   Prob > F =      0.0000
```

IV. Conclusion from omitted variable test

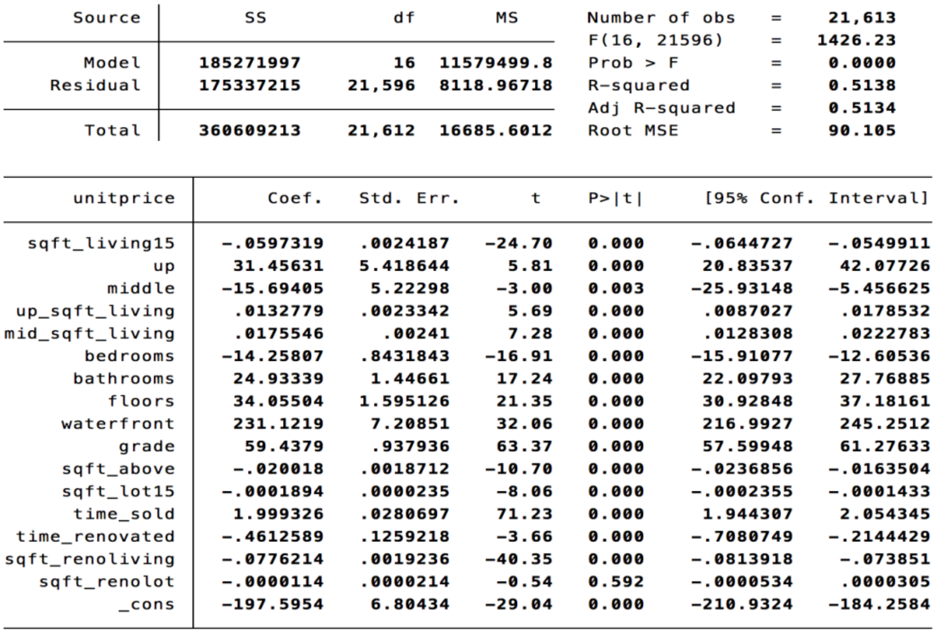
For the sake of brevity, the result from including the fitted value yhat cubic is not displayed here. Still, including this new term does not help to correct the omitted variable problem. The series of experiments conducted above shows that there are some significant variables missing in the original dataset; therefore, we cannot avoid omitted variable problem given the current data.

Noticeably, adding dropped variables or the fitted value yhat squared does not lead to a significant increase in adjusted R-squared. In other words, adding these new terms does not significantly raise the explanatory power of our regression model. Instead, introducing these new terms substantially enhances the problem of multicollinearity, rendering many existing independent variables statistically insignificant.

Based on the preceding reasoning, we therefore adhere to the original model and add no more new terms.

2.1.3 Variable’s individual and joint statistical significance:

When we regress the original model in STATA, the software directly returns the p-value for each individual estimated coefficient. We set the significance level to be 95%. Hence, an estimated coefficient is tested statistically significant if its p-value is smaller than 0.05. The test result shows that: all independent variables, except for sqft_renolot [p-value =0.592 >0.005], are statistically significant.



```
      Source |       SS           df       MS      Number of obs   =    21,613
-------------+----------------------------------   F(16, 21596)    =   1426.23
       Model |  185271997        16  11579499.8   Prob > F        =    0.0000
    Residual |  175337215    21,596  8118.96718   R-squared       =    0.5138
-------------+----------------------------------   Adj R-squared   =    0.5134
       Total |  360609213    21,612  16685.6012   Root MSE        =    90.105

--------------------------------------------------------------------------------
     unitprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
  sqft_living15 |  -.0597319   .0024187   -24.70   0.000    -.0644727   -.0549911
             up |   31.45631   5.418644     5.81   0.000     20.83537    42.07726
         middle |  -15.69405    5.22298    -3.00   0.003    -25.93148   -5.456625
 up_sqft_living |   .0132779   .0023342     5.69   0.000     .0087027    .0178532
mid_sqft_living |   .0175546     .00241     7.28   0.000     .0128308    .0222783
       bedrooms |  -14.25807   .8431843   -16.91   0.000    -15.91077   -12.60536
      bathrooms |   24.93339    1.44661    17.24   0.000     22.09793    27.76885
         floors |   34.05504   1.595126    21.35   0.000     30.92848    37.18161
     waterfront |   231.1219    7.20851    32.06   0.000     216.9927    245.2512
          grade |    59.4379    .937936    63.37   0.000     57.59948    61.27633
     sqft_above |   -.020018   .0018712   -10.70   0.000    -.0236856   -.0163504
     sqft_lot15 |  -.0001894   .0000235    -8.06   0.000    -.0002355   -.0001433
      time_sold |   1.999326   .0280697    71.23   0.000     1.944307    2.054345
 time_renovated |  -.4612589   .1259218    -3.66   0.000    -.7080749   -.2144429
sqft_renoliving |  -.0776214   .0019236   -40.35   0.000    -.0813918    -.073851
   sqft_renolot |  -.0000114   .0000214    -0.54   0.592    -.0000534    .0000305
          _cons |  -197.5954    6.80434   -29.04   0.000    -210.9324   -184.2584
--------------------------------------------------------------------------------
```

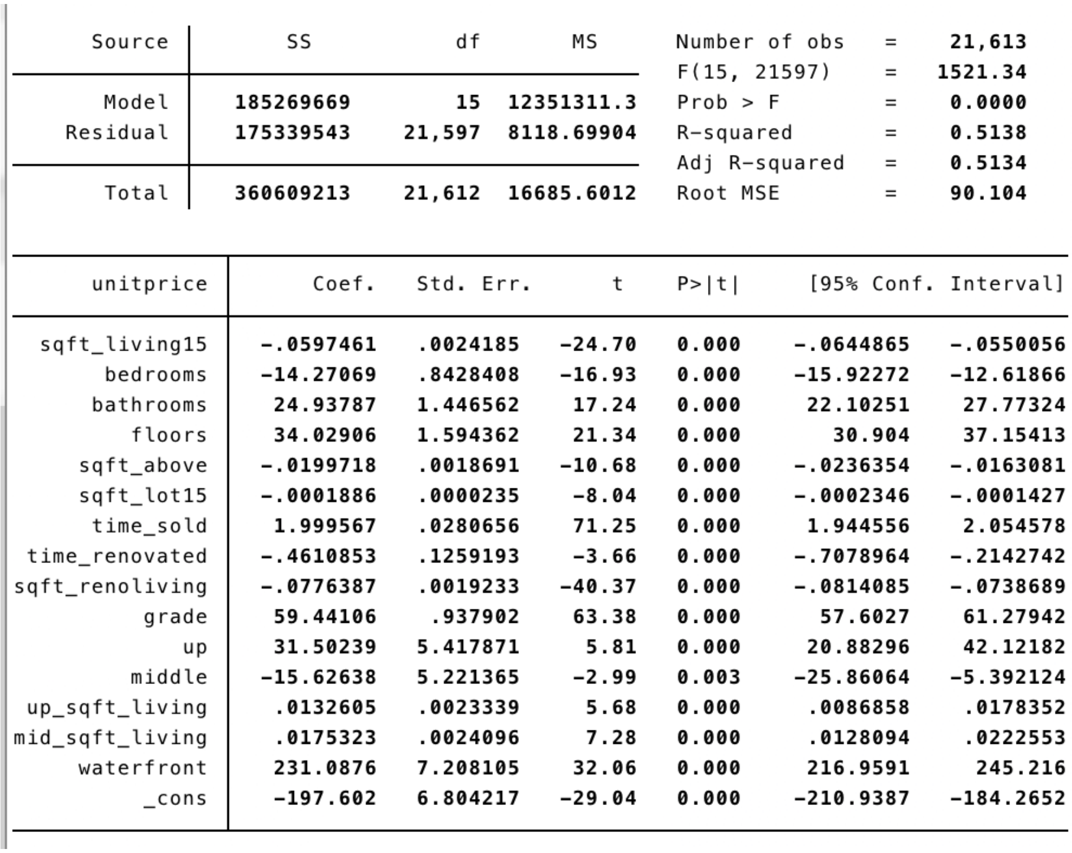
The statistic insignificance of sqft_renolot might be a consequence of its relatively small variance. In empirical analysis, we need big variations of the independent variables to achieve a precise estimation, especially when the dependent variable has large variation. In this sense, we should drop the variable sqft_renolot.

Meanwhile, reckoning this variable in economic sense, we can see the estimated coefficient (-0.0000114) isn’t very significant, which means that the contribution of this variable to the dependent variable is tiny. This indicates slight influence on the model specification even if we drop this variable. The regression model we obtain after dropping this variable is:

$$\begin{aligned} unitprice = {} & \beta_0 + \beta_1 \text{sqft\_living15} + \beta_2 \text{bedrooms} + \beta_3 \text{bathrooms} + \beta_4 \text{floors} \\ & + \beta_5 \text{sqft\_above} + \beta_6 \text{sqft\_lot15} + \beta_7 \text{time\_sold} + \beta_8 \text{time\_renovated} \\ & + \beta_9 \text{sqft\_renoliving} + \beta_{11} \text{grade} + \delta_0 \text{up} + \delta_1 \text{middle} + \delta_2 \text{up\_sqft\_living} \\ & + \delta_3 \text{mid\_sqft\_living} + \delta_4 \text{waterfront} + u \end{aligned}$$

After dropped the variable, we rerun the regression in STATA again, and get all variables statistically significant in a separate and joint perspective at 95% confidence level.

We can see from the specified model, the R-squared, adjusted R-squared don't even change because of dropping one variable. This finding further confirms that the dropped variable has little, if any, explanatory power to the whole model.



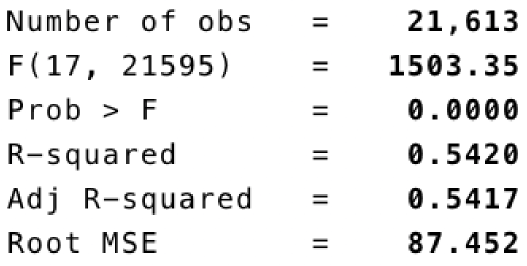
```
      Source |       SS           df       MS      Number of obs   =    21,613
-------------+----------------------------------   F(15, 21597)    =   1521.34
       Model |  185269669        15  12351311.3   Prob > F        =    0.0000
    Residual |  175339543    21,597  8118.69904   R-squared       =    0.5138
-------------+----------------------------------   Adj R-squared   =    0.5134
       Total |  360609213    21,612  16685.6012   Root MSE        =    90.104

------------------------------------------------------------------------------
   unitprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sqft_living15 |  -.0597461   .0024185   -24.70   0.000    -.0644865   -.0550056
    bedrooms |  -14.27069   .8428408   -16.93   0.000    -15.92272   -12.61866
   bathrooms |   24.93787   1.446562    17.24   0.000     22.10251    27.77324
      floors |   34.02906   1.594362    21.34   0.000       30.904    37.15413
  sqft_above |  -.0199718   .0018691   -10.68   0.000    -.0236354   -.0163081
  sqft_lot15 |  -.0001886   .0000235    -8.04   0.000    -.0002346   -.0001427
   time_sold |   1.999567   .0280656    71.25   0.000     1.944556    2.054578
time_renovated |  -.4610853   .1259193    -3.66   0.000    -.7078964   -.2142742
sqft_renoliving |  -.0776387   .0019233   -40.37   0.000    -.0814085   -.0738689
       grade |   59.44106    .937902    63.38   0.000      57.6027    61.27942
          up |   31.50239   5.417871     5.81   0.000     20.88296    42.12182
      middle |  -15.62638   5.221365    -2.99   0.003    -25.86064   -5.392124
up_sqft_living |   .0132605   .0023339     5.68   0.000     .0086858    .0178352
mid_sqft_living |   .0175323   .0024096     7.28   0.000     .0128094    .0222553
  waterfront |   231.0876   7.208105    32.06   0.000     216.9591     245.216
       _cons |   -197.602   6.804217   -29.04   0.000    -210.9387   -184.2652
------------------------------------------------------------------------------
```

2.1.4 Possible Change of functional form in the model:

2.1.4.1 Test the plausibility of the quadratic form

To generally test if any variables' quadratic form should be included in the model, we apply the Ramsey Test. Firstly, we estimate the dependent value from the original regression, denoted as yhat, then we add the square form of yhat and cubic form of yhat into the model, and regress it again. Then we use F test to check the joint significance of the two newly-added-in variables:

When we do the joint test, we can see that these two are jointly insignificant, which means they should be dropped out of the model. In another word, quadratic forms of independent variables are generally not significant in contributing to the model explanation. We can also detect this from the new adjusted R squared, which didn't change by much after the adding in of the quadratic forms [0.5134 vs 0.5417].



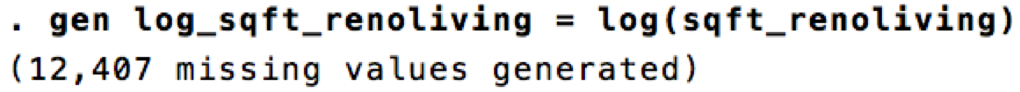
```
Number of obs   =     21,613
F(17, 21595)    =    1503.35
Prob > F        =     0.0000
R-squared       =     0.5420
Adj R-squared   =     0.5417
Root MSE        =     87.452
```

2.1.4.2 Discussion on possible transformation of functional form in both the dependent and independent variables

In our study , we only consider the change of functional form in the sense of logarithm. The objectives of using unit-price as the dependent variable have been discussed before. Since we can already tell the unit change of housing price per square foot, to make the model more explanable in economic sense, we choose to keep the current functional form of the dependent variable. It also doesn't make sense to change the sqft_living and its interaction term (up_sqft_living, mid_sqft_living) into logarithm since they are the key variable of our study given the settled research question.

As for other independent variables, those measuring dummy variables (up, middle, up_sqft_living, mid_sqft_living,waterfront), time-related variables (time_sold, time_renovated), unit numbers (bedrooms,bathrooms, grade, floors) are not suitable to be changed into log form because it doesn't make any economic sense to say that time or number of bedroom increase by one percent. With this being said, the only variables under skeptical for the applied functional form is sqft_above and sqft_renoliving.

Redo the original regression and replace sqft_above and sqft_renoliving by their logarithmic form. However, we can observe from the data that there are negative values for this two variables and if we generate a new variable in their logarithmic form, we will be missing 12,407 observations.



```
. gen log_sqft_renoliving = log(sqft_renoliving)
(12,407 missing values generated)
```

In conclusion, the functional form of our previous model needs no further change in its functional form.

3. Regression diagnostic

3.1 Check on exogeneity of the model and normal distribution of residual [MLR4 & MLR6]

To do this , we first derive the scatter plot and the fitted plot of the residuals according to the fitted value of the model. From this graph we can see the residual is approximately normally distributed around 0 at each point, which means that the conditional zero mean of the model and the normal distribution of the residual fits (MLR4 & MLR6 checked).

From Fig. 5 we can see a scattering funnel-like shape of the residual versus fitted value, which means the variance of the residual increases with the fitted value. This evidence of heteroscedasticity would be discussed in detail in the next part.

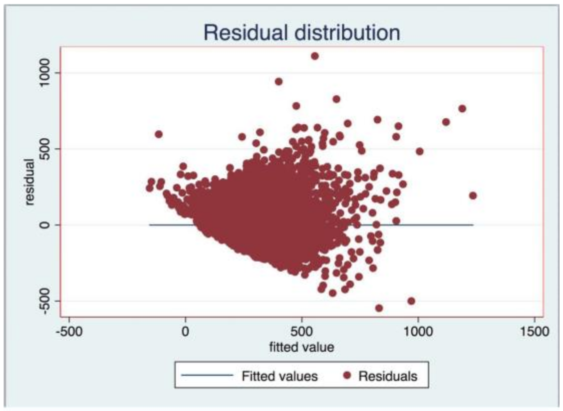
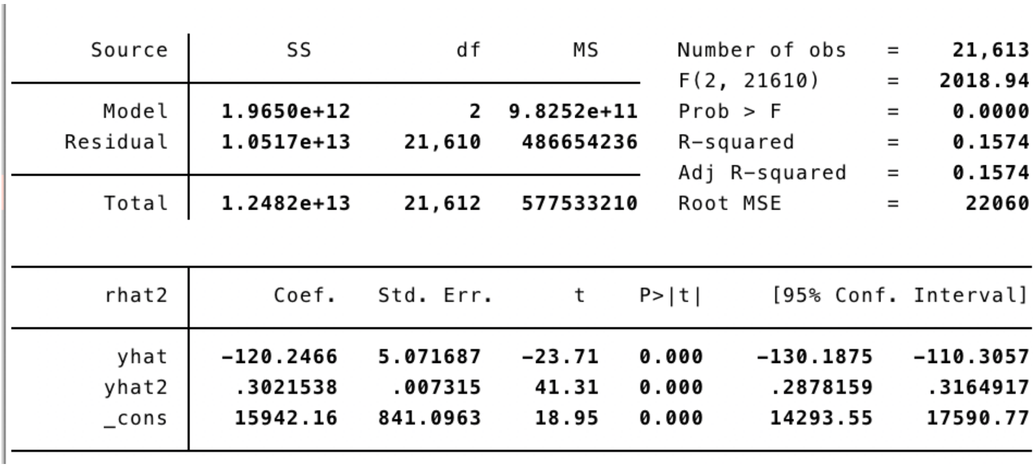


Figure 5

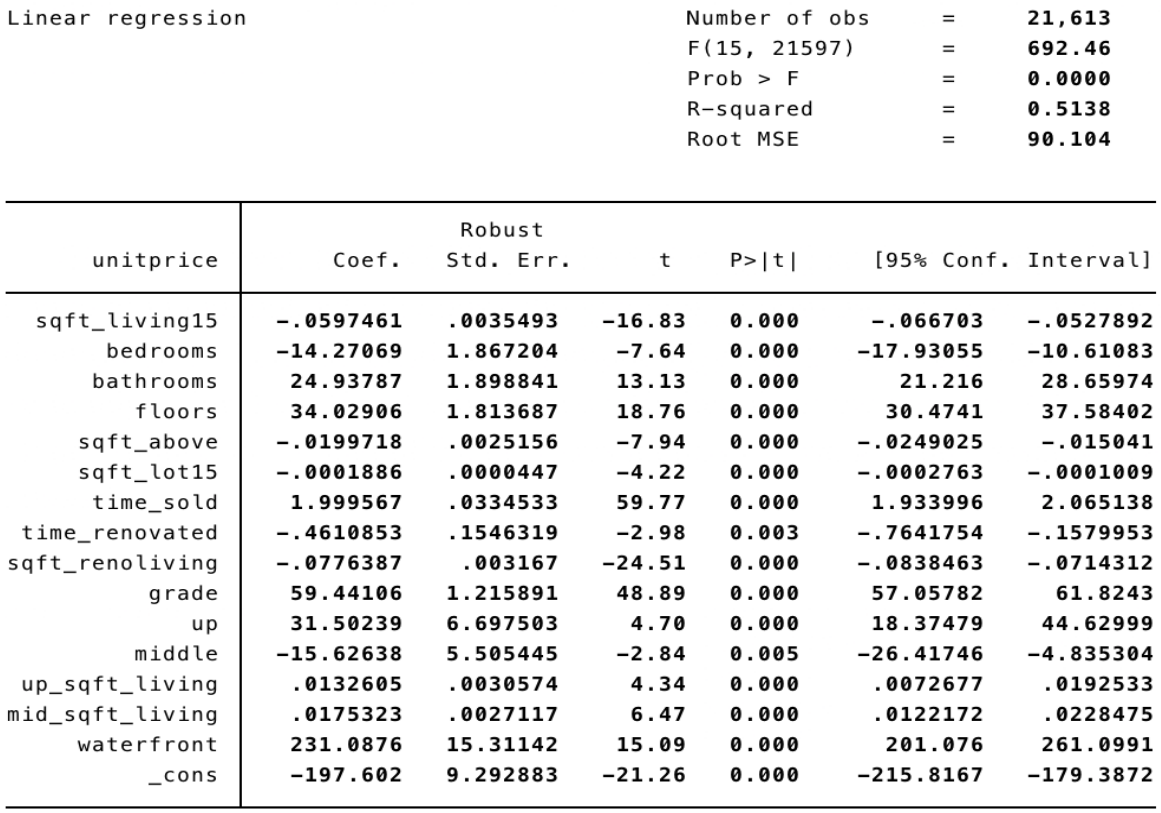
3.2 Check on homoscedasticity [MLR5]

To check if the model satisfies the homoscedasticity [MLR5] assumption for OLS regression, we perform the adjusted-white test. First we we regress the original model, save the residual as rhat, and the fitted value as yhat. Then we regress rhat square on yhat and yhat square to test if the variance of the original model is correlated with the independent variables.



```
      Source |       SS           df       MS      Number of obs   =    21,613
-------------+----------------------------------   F(2, 21610)     =   2018.94
       Model |  1.9650e+12         2  9.8252e+11   Prob > F        =    0.0000
    Residual |  1.0517e+13    21,610   486654236   R-squared       =    0.1574
-------------+----------------------------------   Adj R-squared   =    0.1574
       Total |  1.2482e+13    21,612   577533210   Root MSE        =     22060

------------------------------------------------------------------------------
       rhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yhat |  -120.2466   5.071687   -23.71   0.000    -130.1875   -110.3057
       yhat2 |   .3021538    .007315    41.31   0.000     .2878159    .3164917
       _cons |   15942.16   841.0963    18.95   0.000     14293.55    17590.77
------------------------------------------------------------------------------
```

Given the report from STATA, we find yhat, yhat2 statistically significant in deciding the variance of the model, which means the variance's correlation with the explanatory variance and a violation of the homoscedasticity assumption [MLR5]. So we should use the robust standard error to adjust for the inflated estimated coefficient from the previous OLS regression.
By type "robust" at the end of the previous regression, we get the consistent OLS estimator for the model.



```
Linear regression                               Number of obs     =     21,613
                                                F(15, 21597)      =     692.46
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5138
                                                Root MSE          =     90.104

------------------------------------------------------------------------------
             |               Robust
   unitprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sqft_living15|  -.0597461   .0035493   -16.83   0.000     -.066703   -.0527892
    bedrooms |  -14.27069   1.867204    -7.64   0.000    -17.93055   -10.61083
   bathrooms |   24.93787   1.898841    13.13   0.000       21.216    28.65974
      floors |   34.02906   1.813687    18.76   0.000      30.4741    37.58402
  sqft_above |  -.0199718   .0025156    -7.94   0.000    -.0249025    -.015041
  sqft_lot15 |  -.0001886   .0000447    -4.22   0.000    -.0002763   -.0001009
   time_sold |   1.999567   .0334533    59.77   0.000     1.933996    2.065138
time_renovated| -.4610853   .1546319    -2.98   0.003    -.7641754   -.1579953
sqft_renoliving| -.0776387   .003167   -24.51   0.000    -.0838463   -.0714312
       grade |   59.44106   1.215891    48.89   0.000     57.05782     61.8243
          up |   31.50239   6.697503     4.70   0.000     18.37479    44.62999
      middle |  -15.62638   5.505445    -2.84   0.005    -26.41746   -4.835304
up_sqft_living|  .0132605   .0030574     4.34   0.000     .0072677    .0192533
mid_sqft_living| .0175323   .0027117     6.47   0.000     .0122172    .0228475
  waterfront |   231.0876   15.31142    15.09   0.000      201.076    261.0991
       _cons |   -197.602   9.292883   -21.26   0.000    -215.8167   -179.3872
------------------------------------------------------------------------------
```

4. Regression validation

4.1 Goodness of fit

After testing the residual normal distribution assumption, checking and adjusting the heteroscedasticity, justifying for the exogeneity of our model, we want to further understand how the data we collected and filtered are explained by our model. To do this, we regress the model using robust standard error and check the adjusted R squared reported from STATA.

We get adjusted R squared equals 0.5138, which means that approximately 51.38% of the datas in the dataset we use is explained by the model, which is already pretty good considering the fact that we are studying a heterogeneous commodity (houses) and volatile housing price.

4.2 Further justification for the efficiency of robust OLS estimators

After the model specification and diagnostics, we now test again in a further detailed way for the efficiency of this model.

MLR1: linear in parameter, the model we finalized for our study of housing market is, from which we can see all parameters are in linear relationship to one another, which means that MLR1 linear in parameter is satisfied.

MLR2: Random sampling.From the distribution of the unitprice (dependent variable), we can see the dispersion in its value's distribution, so that we can say that MLR2 random sampling assumption is satisfied.

MLR3: No perfect collinearity, by regressing the model in STATA, no perfect collinearity is detected and dropped, so that MLR3 no perfect collinearity is satisfied.
MLR4: Zero conditional mean of the error term. From the twoway scatter plot and the fitted plot with y-axis the residual and x-axis fitted value, we observe the approximate symmetric distribution of y with respect to x around the fitted value line y=0, this justifies MLR4 endogeneity of the model.
MLR5: Homoscedasticity. This is tested false by the adjusted white test in 4.2. Yet we can adjust the bias caused by heteroscedasticity by using the robust standard error. Because of MLR1 - MLR4 , the OLS estimator is an unbiased and consistent one. Then with the adjustment of heteroscedasticity, we have BLUE (best linear unbiased estimator), which is efficient.
MLR6:Then we further check on MLR6 by deriving the density histogram of the residual distribution. We can see it's approximately normally distributed around zero. In this way, MLR6 is also satisfied, giving stronger assumptions to the model.

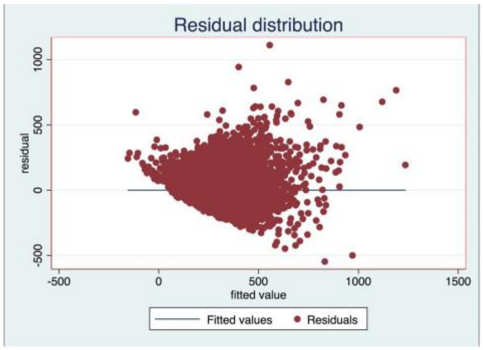


Figure 6

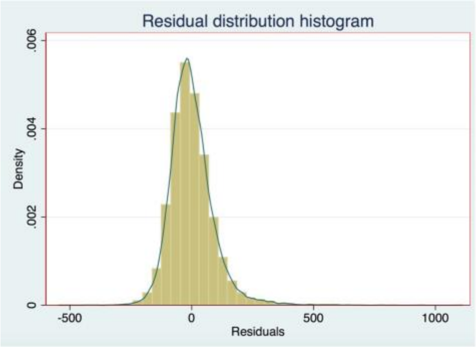


Figure 7

5. Discussion on the regression result

Recall our research topic: we are interested to know how different the unitprice of the house would respond to per unit change in living square across different income groups.

The econometrics model we specified is:

$$\begin{aligned} unitprice = {} & \beta_0 + \beta_1 \text{sqft\_living15} + \beta_2 \text{bedrooms} + \beta_3 \text{bathrooms} + \beta_4 \text{floors} \\ & + \beta_5 \text{sqft\_above} + \beta_6 \text{sqft\_lot15} + \beta_7 \text{time\_sold} + \beta_8 \text{time\_renovated} \\ & + \beta_9 \text{sqft\_renoliving} + \beta_{11} \text{grade} + \delta_0 \text{up} + \delta_1 \text{middle} + \delta_2 \text{up\_sqft\_living} \\ & + \delta_3 \text{mid\_sqft\_living} + \delta_4 \text{waterfront} + u \end{aligned}$$

5.1 direct inference from the regression result about difference in between the lower and upper lower and middle class

1. With one square foot increase in the living square, the housing price per square foot in upper-class increases by 0.0132605 more than that in lower-class on average.
2. With one square foot increase in the living square, the housing price per square foot in middle-class increases by 0.0175323 more than that in lower-class on average.

From the t-test we can see, such differences are statistically significant in between upper-class, middle class and lower class.

5.2 Further discussion on the difference between upper class and middle class

Then we compare whether the difference between the upper and middle class is statistically significant, to do that, we do hypothesis testing on whether the coefficient of up_sqft_living and mid_sqft_living are significantly different in the statistical sense.

*** Hypothesis testing**

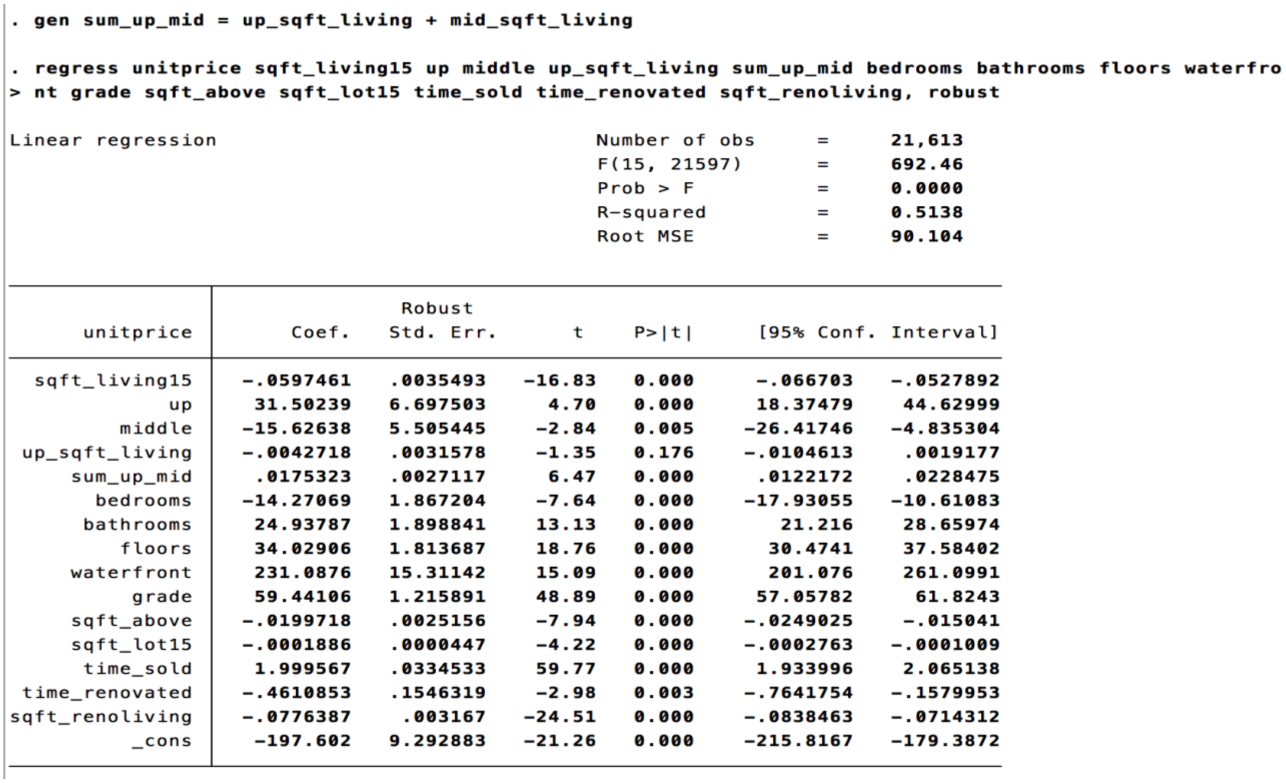
$$H_o: \delta_2 - \delta_3 = 0$$

Let $\gamma = \delta_2 - \delta_3$, the equation can be rewritten into the form:

$$\begin{aligned} unitprice = {} & \beta_0 + \beta_1\text{sqft\_living15} + \beta_2\text{bedrooms} + \beta_3\text{bathrooms} + \beta_4\text{floors} \\ & + \beta_5\text{sqft\_above} + \beta_6\text{sqft\_lot15} + \beta_7\text{time\_sold} + \beta_8\text{time\_renovated} \\ & + \beta_9\text{sqft\_renoliving} + \beta_{11}\text{grade} + \delta_0\text{up} + \delta_1\text{middle} + (\gamma \\ & + \delta_3)\text{up\_sqft\_living} + \delta_3\text{mid\_sqft\_living} + \delta_4\text{waterfront} + u \\ = {} & \beta_0 + \beta_1\text{sqft\_living15} + \beta_2\text{bedrooms} + \beta_3\text{bathrooms} + \beta_4\text{floors} \\ & + \beta_5\text{sqft\_above} + \beta_6\text{sqft\_lot15} + \beta_7\text{time\_sold} + \beta_8\text{time\_renovated} \\ & + \beta_9\text{sqft\_renoliving} + \beta_{11}\text{grade} + \delta_0\text{up} + \delta_1\text{middle} + \gamma\text{up\_sqft\_living} \\ & + \delta_3(\text{up\_sqft\_living} + \text{mid\_sqft\_living}) + \delta_4\text{waterfront} + u \end{aligned}$$

To test if $\gamma$is statistically significant to the new model, we have to derive a new variable sum_up_mid (=up_sqft_living + mid_sqft_living) and run the OLS regression on the new variable using the robust standard error.

The t-test p-value on coefficient of up_sqft_living (γ) is 0.176 > 0.05, given a 95% significance level. Therefore, the difference between these two coefficients is not statistically significant.



```
. gen sum_up_mid = up_sqft_living + mid_sqft_living

. regress unitprice sqft_living15 up middle up_sqft_living sum_up_mid bedrooms bathrooms floors waterfro
> nt grade sqft_above sqft_lot15 time_sold time_renovated sqft_renoliving, robust

Linear regression                               Number of obs     =     21,613
                                                F(15, 21597)      =     692.46
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5138
                                                Root MSE          =     90.104

---------------------------------------------------------------------------------
                |               Robust
      unitprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
  sqft_living15 |  -.0597461   .0035493   -16.83   0.000     -.066703   -.0527892
             up |   31.50239   6.697503     4.70   0.000     18.37479    44.62999
         middle |  -15.62638   5.505445    -2.84   0.005    -26.41746   -4.835304
 up_sqft_living |  -.0042718   .0031578    -1.35   0.176    -.0104613    .0019177
     sum_up_mid |   .0175323   .0027117     6.47   0.000     .0122172    .0228475
       bedrooms |  -14.27069   1.867204    -7.64   0.000    -17.93055   -10.61083
      bathrooms |   24.93787   1.898841    13.13   0.000       21.216    28.65974
         floors |   34.02906   1.813687    18.76   0.000      30.4741    37.58402
     waterfront |   231.0876   15.31142    15.09   0.000      201.076    261.0991
          grade |   59.44106   1.215891    48.89   0.000     57.05782     61.8243
     sqft_above |  -.0199718   .0025156    -7.94   0.000    -.0249025    -.015041
     sqft_lot15 |  -.0001886   .0000447    -4.22   0.000    -.0002763   -.0001009
      time_sold |   1.999567   .0334533    59.77   0.000     1.933996    2.065138
 time_renovated |  -.4610853   .1546319    -2.98   0.003    -.7641754   -.1579953
sqft_renoliving |  -.0776387    .003167   -24.51   0.000    -.0838463   -.0714312
          _cons |   -197.602   9.292883   -21.26   0.000    -215.8167   -179.3872
---------------------------------------------------------------------------------
```

5.3 Interpretation of regression result

5.3.1 Answers to research question

Econometric methodology and careful thoughts have been put into the enquiry of our research question. The above discourse centers around the question of our interest and is structured in an efficient manner to approach the answers.

Based on preceding regression result and subsequent hypothesis testing, we therefore conclude our findings are:

1. Living square of the house has more contributions to the per square foot housing price on average in middle and upper class than in lower class. From the buyers' market, this indicates that there are more demand for larger houses in the middle and upper class. From the constructors' side, this indicates profitability of building bigger houses in the middle and upper class regions, and smaller houses in the lower class regions.
2. However, the increase in square footage of living space does not result in significantly different increase of per square foot housing price in between the middle and upper class. From the buyers' perspective, this indicates more or less similar willingness of middle class and upper class in buying comparatively bigger houses. From the constructors' side, it might be equally profitable to develop big houses in both middle and upper class.
3. General difference (the coefficient of variables up & middle, which measures the difference in per square foot housing price when living square is zero) in housing price per square foot suggests that for upper class, they are more willing to pay higher

(31.50239) for unit price than lower class. However, the middle class are less willing to pay for unit price than lower class by 15.62638.

5.3.2 How we reconcile with unexpected results of 2 & 3?

1. It might be surprising to find that the middle and upper class have similar appetite for big houses, considering about theri income gaps.

This might be explained by their different investment intention for those houses. For the upper class, they buy bigger houses for living, for needs to collaterally showing their wealth and reputation. For the middle class, they may invest in bigger houses for purely investment since most people still believe today that the housing market is the most stable and promising one.

2. In this empirical study, we are surprised to see that the dummy variable middle has a negative coefficient (-15.62638). This finding suggests that, given a house with humble living space, the willingness of middle class to pay for the unit price per square is actually smaller than lower class.

This phenomena might be answered by the desire of middle class for comfortable living experience. Middle class are equipped with enough financial capacity to afford their ideal houses. In this case, houses with small living space become less welcomed by middle class and result in lower unit price than lower class because of the humble demand.

5.3.3 Insights into real estate market - real world implications

While the scope of this empirical study is relatively confined within King County level, we do hope to cast light upon some real world implications from this research and better understand the real estate market mechanism.

Based on our regression results, custructors should follow the rule to build more large houses catered for upper class and middle class while build smaller houses for lower class. This decision will influence the layout of houses and communities along with other factors before construction. When evaluating construction plans, real estate developers should take all these elements into consideration to maximize the profit from a given land.

5.3.4 Further research design

Now that we have learned the divergent willingness of lower class, middle class, and upper class when facing different house options, we suspect the financing methods might be an important variable affecting the utility of purchasing certain houses. For example, buying house with cash or mortgage will certainly make a difference for people about whether buying a big house. Hence, further research can be designed to detect the explanatory power of different financing methods in affecting housing price across different income classes.

6. References

Abdulai, Raymond T., and Owusu-Ansah, Anthony. “House Price Determinants in Liverpool, United Kingdom.” *Current Politics and Economics of Europe*, vol. 22, no. 1, 2011, pp. 1 - 26.

Grum, Bojan, and Darja Kobe Govekar. “Influence of Macroeconomic Factors on Prices of Real Estate in Various Cultural Environments: Case of Slovenia, Greece, France, Poland and Norway.” *Procedia Economics and Finance*, vol. 39, 2016, pp. 597–604., doi:10.1016/s2212-5671(16)30304-5.

Christopher J Mayer. “US Housing price dynamics and behavior finance”.