Are the rich buying more large houses?

An empirical study on King County housing prices

Group III: Zhiyu Chen, Wenxin Feng, Yuxin Zhang

Agenda

1. research question
2. how we filter the variables in the model
3. model specification & diagnostics
4. result and preliminary interpretation
5. conclusions
6. enlightenment of future study

Agenda

1. research question
2. how we filter the variables in the model
3. model specification & diagnostics
4. result and preliminary interpretation
5. conclusions
6. enlightenment of future study

How we approach our research question?

> Data set : What's Available?

> Economic Intuition & Literature Review

Data available

id	date	price	bedrooms	bathrooms	sqft_living	floors
view	condition	grade	sqft_above	sqft_basement	sqft_living15	zipcode
lat	waterfront	Long	sqft_lot15	yr_renovated	sqft_lot	yr_built

Data set

Kept variables						
price	bedrooms	bathrooms	sqft_living	sqft_lot	sqft_living15	sqft_lot15
waterfront	yr_renovate	yr_built	grade	sqft_above	floors	

Dropped variable	id	date	view	condition	sqft_basement
Confusing data	zipcode	lat	long		

How we make sense of the confusing data？

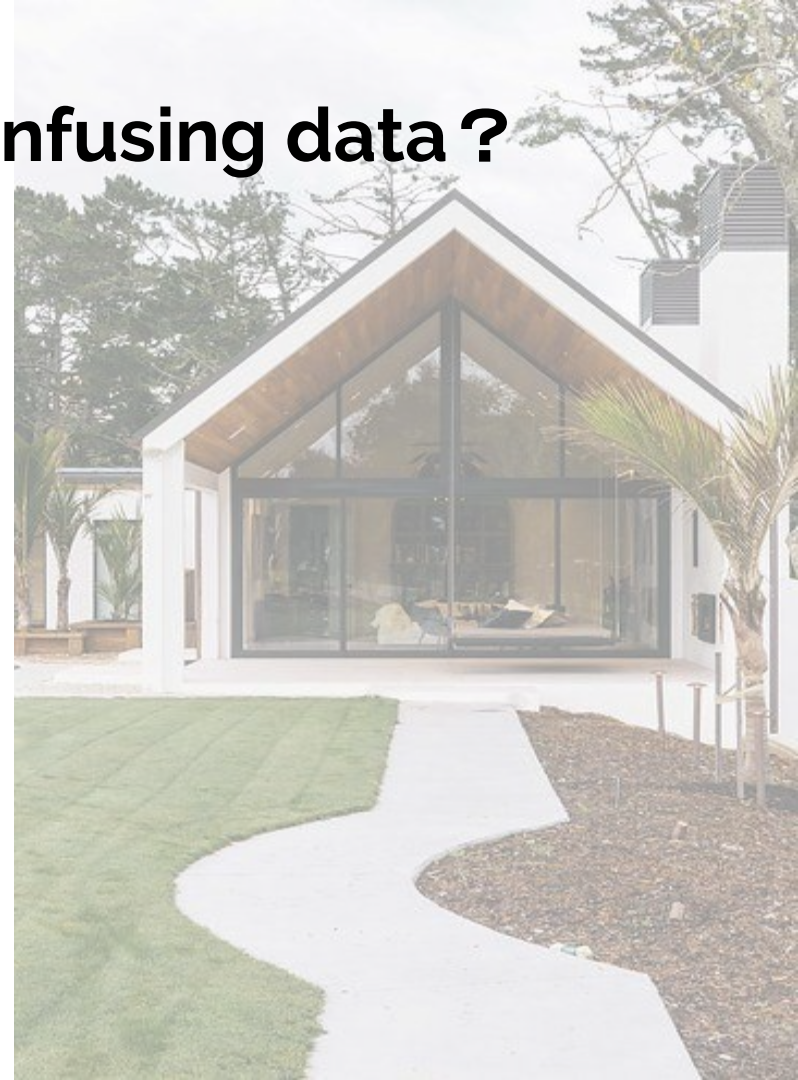
latitude / longitude / zipcode

↓

Location of the house

↓

Different income groups



Inspiration from income class dummy

Literature review and perspective on housing market:

→ Housing price almost always goes up with the living square →

Boring

→ What about its contribution to the unit square housing price?

→ Is it generally positive ?

→ Does it have different contributions across different income levels ?

Intuition

→ the unit price reflects the demand from the buyers [preference]

→ the unit price reflects the profitability of the constructor [supply]

Intuitively

Small houses

Big houses

Upper class

Low income class

Research question

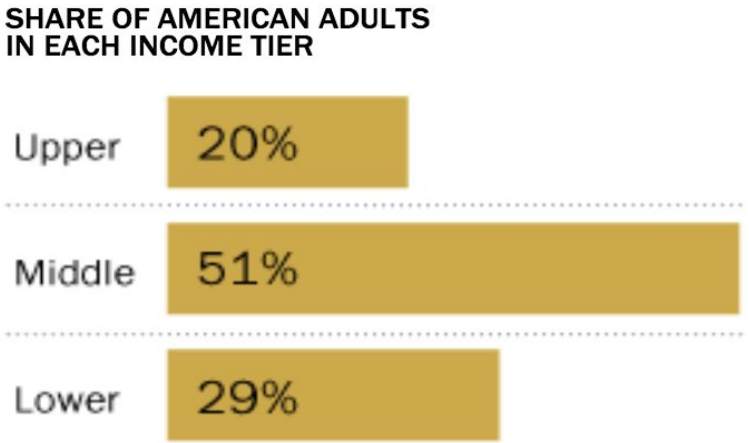
When the square footage of total house living space increases by unit, does the unit price of house per square foot change more in upper class regions than that in middle and lower class regions?

Agenda

1. research question
2. how we filter the variables in the model
3. model specification & diagnostics
4. result and preliminary interpretation
5. conclusions
6. enlightenment of future study

How we obtain the dummy for the income class?

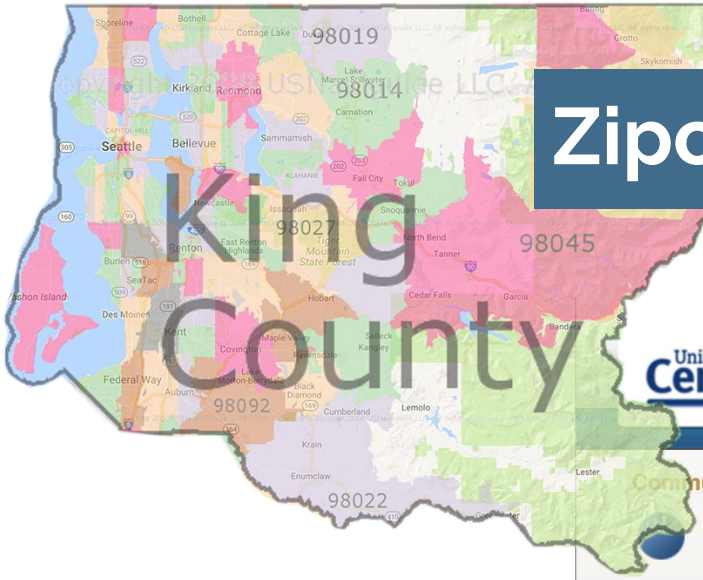
- Median income > 100,000 [20%] → upper income region
- Median income > 30,000 [50%] → middle income region
- Median income < 30,000 [30%] → low income region



Assumption:

The income distribution follows that in the nation wide

How we obtain the dummy for the income class?



Zipcode → **Median income**

United States Census Bureau — AMERICAN FactFinder

MAIN | COMMUNITY FACTS | GUIDED SEARCH | ADVANCED SEARCH | DOWNLOAD CENTER

Community Facts - Find popular facts (population, income, etc.) and frequently requested data about your community.

Enter a state, county, city, town, or zip code: GO

- Population
- Age
- Business and Industry
- Education
- Governments
- Housing
- Income

98001

Median Household Income

73,782 Source: 2012-2016 American Community Survey 5-Year Estimates

Bookmark/Save | Print

Popular tables for this geography:

2016 American Community Survey

- Selected Economic Characteristics (Employment, Commute, Occupation, Income, Health Insurance, Poverty, ...)
- Income in the Past 12 Months (Households, Families, ...)
- Earnings in the Past 12 Months (Sex, Educational Attainment, ...)

Do we have enough data set for each group ?

More frequent housing transactions in low-income area ?

```
. tabulate tag
```

tag	Freq.	Percent
L	13,358	61.81
M	3,821	17.68
U	4,434	20.52

Data Processing

Numerical variable	unitprice	sqft_living	bathrooms	bedrooms	grade
	floor	sqft_above	waterfront	sqft_lot	sqft_lot15
	yr_built	yr_renovated	sqft_living	sqft_living15	
dummy	(L)	middle	up		
Dummy interaction	up_sqft_living	mid_sqft_living			

Processing the data

> Using sqft_lot15 & sqft_living15

> Dropping sqft_lot & sqft _living

> Generate new variable sqft_renolot: sqft_lot15 - sqft_lot

> Generate new variable time_sold: yr_sold - yr_built

> Generate new variable time_renovated: yr_sold - yr_renovated

Preliminary model based on economic intuition

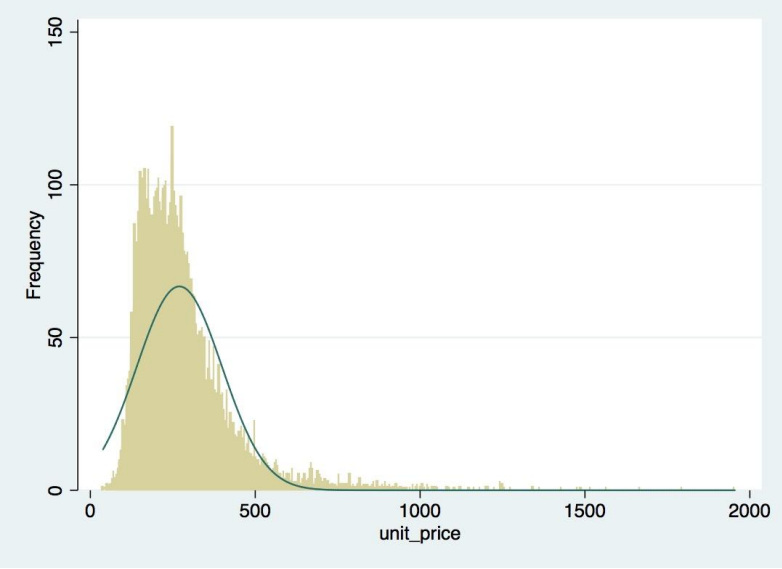
$$
\begin{aligned}
unitprice = {} & \beta_0 + \beta_1 \text{sqft\_living15} + \beta_2 \text{bedrooms} + \beta_3 \text{bathrooms} + \beta_4 \text{floors} \\
& + \beta_5 \text{sqft\_above} + \beta_6 \text{sqft\_lot15} + \beta_7 \text{time\_sold} + \beta_8 \text{time\_renovated} \\
& + \beta_9 \text{sqft\_renoliving}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

Agenda

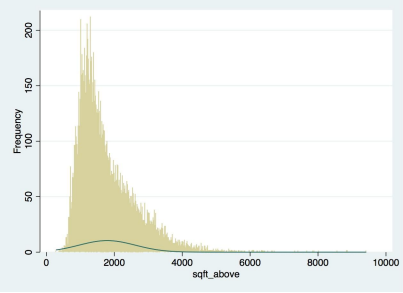
1. research question
2. how we filter the variables in the model
3. model specification & diagnostics
4. result and preliminary interpretation
5. conclusions
6. enlightenment of future study

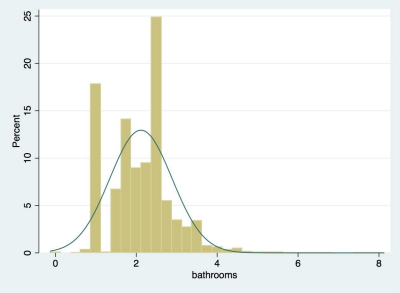
Descriptive Variables

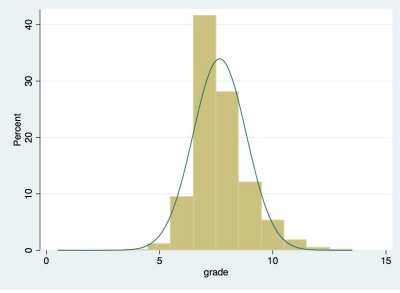
Dependent variable

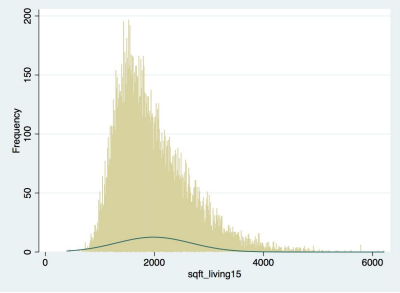


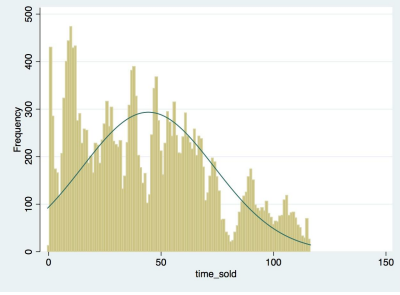
Independent variables

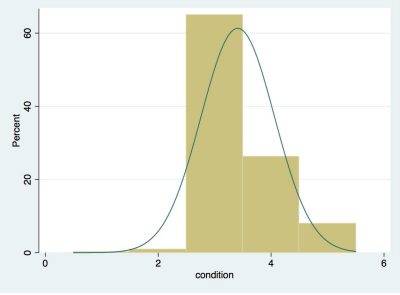












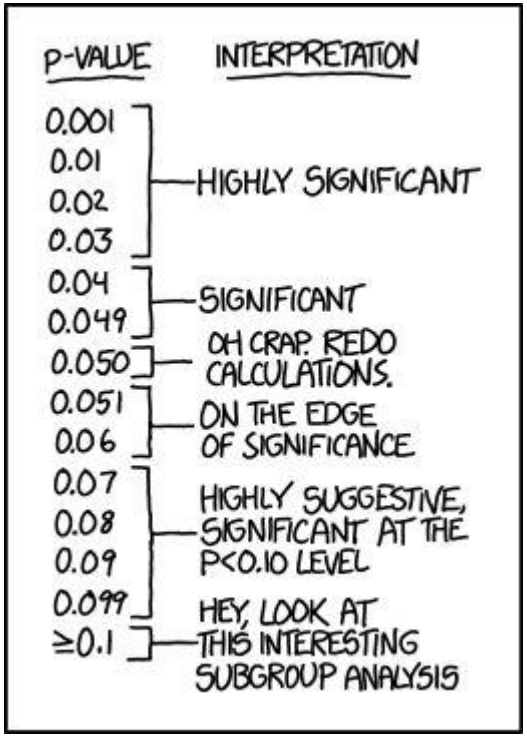
Inferential test

> Omitted variable test

> Quadratic terms

> Other possible transformation

> CLM assumption test



Preliminary regression

Source	SS	df	MS		
Model	185271997	16	11579499.8	Number of obs	= 21,613
Residual	175337215	21,596	8118.96718	F(16, 21596)	= 1426.23
				Prob > F	= 0.0000
				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
sqft_renolot	-.0000114	.0000214	-0.54	0.592	-.0000534	.0000305
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
_cons	-197.5954	6.80434	-29.04	0.000	-210.9324	-184.2584

u

$\text{oms} + \beta_4 \text{floors}$

$- \beta_8 \text{time\_renovated}$

$\text{de} + \delta_0 \text{up} + \delta_1 \text{middle}$

$\text{terfront} + u$

P = 0.592 > > 0.05

Drop sqft_renolot

Rerun the regression after dropping

Source	SS	df	MS
Model	185269669	15	12351311.3
Residual	175339543	21,597	8116.69904
Total	360609213	21,612	16685.6012

Number of obs	21,613
F(15, 21597)	1521.34
Prob > F	0.0000
R-squared	0.5138
Adj R-squared	0.5134
Root MSE	90.104

R-squared = 0.5138
Adj R-squared = 0.5134

No change

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

All statistically significant at 95% confidence level

Possible quadratic form ?

1. No specific economic intuition driven
2. By taking Ramsey test, the new model (with yhat^2, yhat^3) shows insignificance in the joint test

```
. test yhat2 yhat3

 ( 1)  yhat2 = 0
 ( 2)  yhat3 = 0
       Constraint 2 dropped

       F(  1, 21594) =  468.28
            Prob > F =    0.1281
```

Not significant

```
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
```

Previously 0.5134

Functional form change ?

1. Based on research question:

   unitprice, up_sqft_living, mid_sqft_living, up, middle [can't change]

2. Based on economics meaning:

   Dummy variable : waterfront

   Time-measuring variable : time_sold, time_renovated

   Unit measuring variable : bedrooms,bathrooms, grade, floors

3. One possible change : sqft_renoliving

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

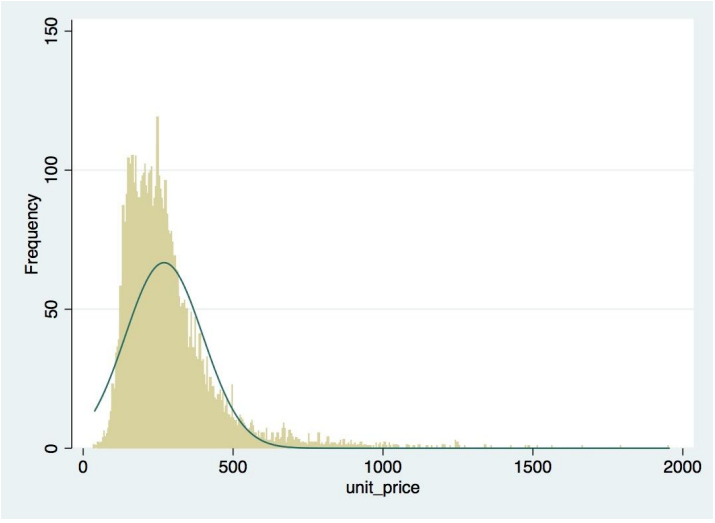
CLM

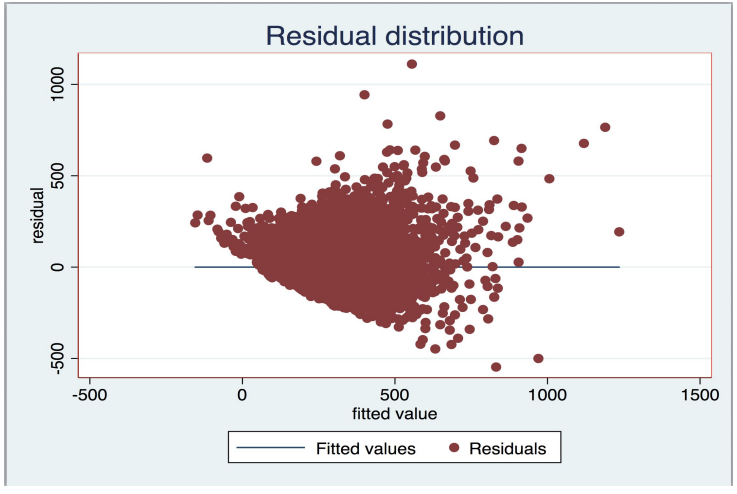
>MLR1: Linear in parameter

>MLR2 : Random sampling

>MLR3: No perfect linearity

>MLR4: Zero conditional mean





MLR5: Homoscedasticity

Adjusted white test :

Regress rhat^2 [predicted residual] on yhat & yhat^2, test joint significance

Source	SS	df	MS		
Model	1.9650e+12	2	9.8252e+11	Prob > F	0.0000
Residual	1.0517e+13	21,610	486654236	R-squared	0.1574
Total	1.2482e+13	21,612	577533210	Root MSE	22060

Number of obs	=	21,613			
F(2, 21610)	=	2018.94			
Prob > F	=	0.0000			
R-squared	=	0.1574			
Adj R-squared	=	0.1574			
Root MSE	=	22060			

→ Heteroscedasticity

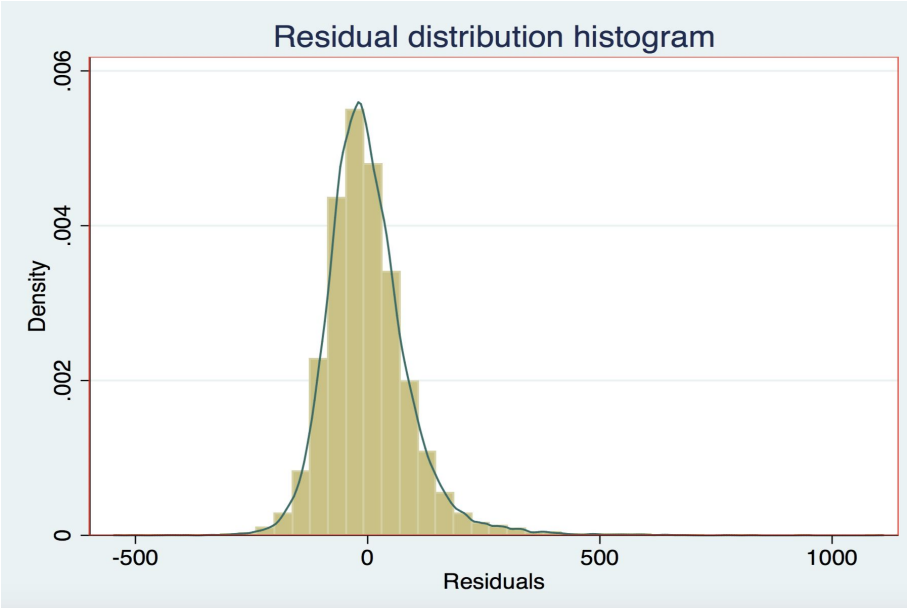
rhat2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
yhat	-120.2466	5.071687	-23.71	0.000	-130.1875	-110.3057
yhat2	.3021538	.007315	41.31	0.000	.2878159	.3164917
_cons	15942.16	841.0963	18.95	0.000	14293.55	17590.77

Classical Linear Model

> MLR1-4 → Efficient estimator

> MLR5 -> Adjusted by robust standard error

> MLR6: random distribution of the error



Agenda

1. research question
2. how we filter the variables in the model
3. model specification & diagnostics
4. result and preliminary interpretation
5. conclusions
6. enlightenment of future study

Specified Econometrics Model

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

What does the model tell us directly?

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

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.

Is there any difference between upper and middle class?

*** Hypothesis testing**

$H_o : \delta_2 - \delta_3 = 0$

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

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

Interest of hypothesis testing

New variable: sum_up_mid

Conclusion:
Upper class and Middle class do not have significantly difference

```
. 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
```

unitprice	Coef.	Robust 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
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

up_sqft_living -.0042718 .0031578 -1.35 0.176

P-value = 0.176 > 0.05

We cannot reject H0

Agenda

1. research question
2. how we filter the variables in the model
3. model specification & diagnostics
4. result and preliminary interpretation
5. Conclusions
6. enlightenment of future study

Conclusion

Increasing profitability of building bigger houses in the middle and upper class regions, and smaller houses in the lower class regions

There are more demand for larger houses in the middle and upper class.

Conclusion

No significant difference between upper class and middle class

It might be equally profitable to develop big houses in both middle and upper class.

It indicates more or less similar willingness of middle class and upper class in buying comparatively bigger houses

Other surprising finding

When facing houses with humble living space:

For upper class, they are more willing to pay higher for unit housing price by 31.50239 than lower class.

For middle class, they want to pay lower for unit housing price by 15.62638 than lower class .

time_sold	1.999567	.0280656	71.25	0.000	
time_renovated	-.4610853	.1259193	-3.66	0.000	
sqft_renoliving	-.0776387	.0019233	-40.37	0.000	
grade	59.44106	.937902	63.38	0.000	
up	31.50239	5.417871	5.81	0.000	
middle	-15.62638	5.221365	-2.99	0.003	
up_sqft_living	.0132605	.0023339	5.68	0.000	
mid_sqft_living	.0175323	.0024096	7.28	0.000	
waterfront	231.0876	7.208105	32.06	0.000	
_cons	-197.602	6.804217	-29.04	0.000	

Conclusion

The willingness of middle class to pay for the unit price is actually smaller than lower class

No significant difference between upper class and middle class

WHY ?

Agenda

1. research question
2. how we filter the variables in the model
3. model specification & diagnostics
4. result and preliminary interpretation
5. conclusion
6. enlightenment of future study

Further Insights

Our Suggestions to Real Estate Market !

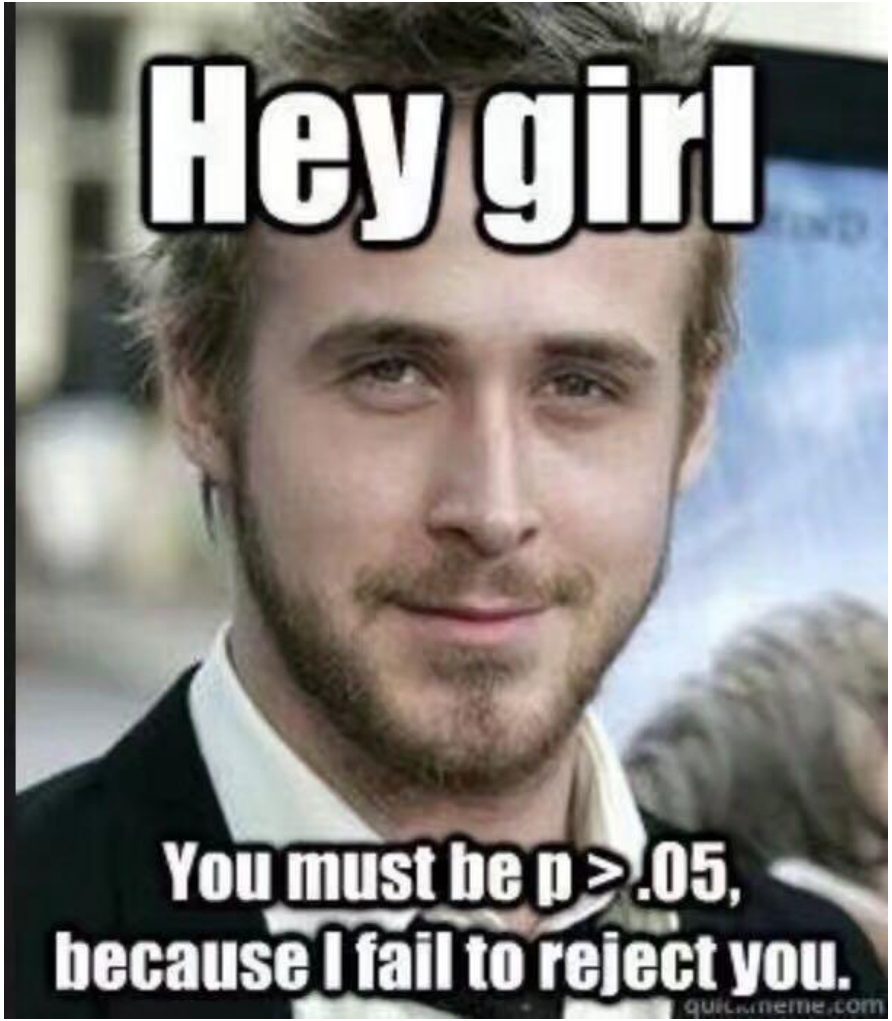
Custructors may follow this empirical finding to build more large houses catered for upper class and middle class while build smaller houses for lower class. This decision will influence the layout of houses and communities along with other factors before construction. When evaluating construction plans,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

Further Insights

What Research Can We Do In the Future ?

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**.


Hey girl
You must be p > .05,
because I fail to reject you.