Yang Huan
A0224968N
e0575602@u.nus.edu

除问答,其余在do文件,和表格中

如有问题,随时联络

All students have to hand in this assignment by 11:59 pm Sunday 20 September 2020. **Please upload both your answers and program code**.

1. Assume that the housing price is determined by the following equation

$$\ln p_i = \beta_0 + \beta_1 \ln s_i + \beta_2 n_i + \epsilon_i \tag{1}$$

where p_i is the price per square foot, s_i is the size of the unit in squared feet, n_i is the number of bedrooms, ϵ_i is an error term, and $E(\epsilon|s,n) = 0$. Suppose a researcher runs the following regression

$$\ln p_i = \alpha_0 + \alpha_1 \ln s_i + e_i \tag{2}$$

(a) Under what assumptions will the OLS estimate of α_1 provide an unbiased estimate of β_1? Are these assumptions realistic?

(a) $\because \alpha_1 = \beta_1 + \beta_2 \dfrac{cov(S_i, n_i)}{Var(S_i)}$ $\therefore$ if $\beta_2 = 0$ or $cov(S_i, n_i) = 0 \to \alpha_1$ unbiased β_1

These two is realistic to sume extent.

(b) If these assumptions are violated, will the OLS estimate of α_1 over- or under-estimate the impact of size on price? Explain your answer.

(d) 1. reduce the absolute value.
2. the data is more stable, weakens the collinearity and heteroscedasticity
3. show the economic implication of elasticity
4. Distribution close to nomal.

(c) Using housing_ass.csv to plot the histogram of resale_price and its log.

(d) Explain why researchers prefer to use log price rather than price in the regression.

(e) Based on the information contained in variable *flat_type* to generate the number of rooms (assume there are 6 rooms in EXECUTIVE apartments and 7 rooms in MULTI-GENERATION apartments) and use the data to justify your answer by reporting the regression results in Table 1 and the correlation between log size and the number of rooms.

(b) if $n_i = C_0 + C_1 \ln S_i + V_i \to (1)\ \ln p_i = (\beta_0 + \beta_2 \cdot C_0) + (\beta_1 + \beta_2 C_1) \ln S_i + \beta_2 V_i + G_i$

if $\beta_1 < 0$ $\beta_2 > 0$ $C_1 > 0$, α_1 over estimates the negative impact of size on price.

$E(\alpha_1) > \beta_1$ 向上偏误.

via (e) solution get

1

(a)

Are these assumptions realistic?

(a) $\because \alpha_1 = \beta_1 + \beta_2 \frac{cov(s_i, n_i)}{var(s_i)}$ $\therefore$ if $\beta_2 = 0$ or $cov(s_i, n_i) = 0 \rightarrow \alpha_1$ unbiased β_1

These two is realistic to sume extent.

(b)

Table 1 and the correlation between log size and the number of rooms.

(b) if $n_i = c_0 + c_1 \ln s_i + v_i \rightarrow (1)\ \ln p_i = (\beta_0 + \beta_2 \cdot c_0) + (\beta_1 + \beta_2 c_1) \ln s_i + \beta_2 v_i + \varepsilon_i$

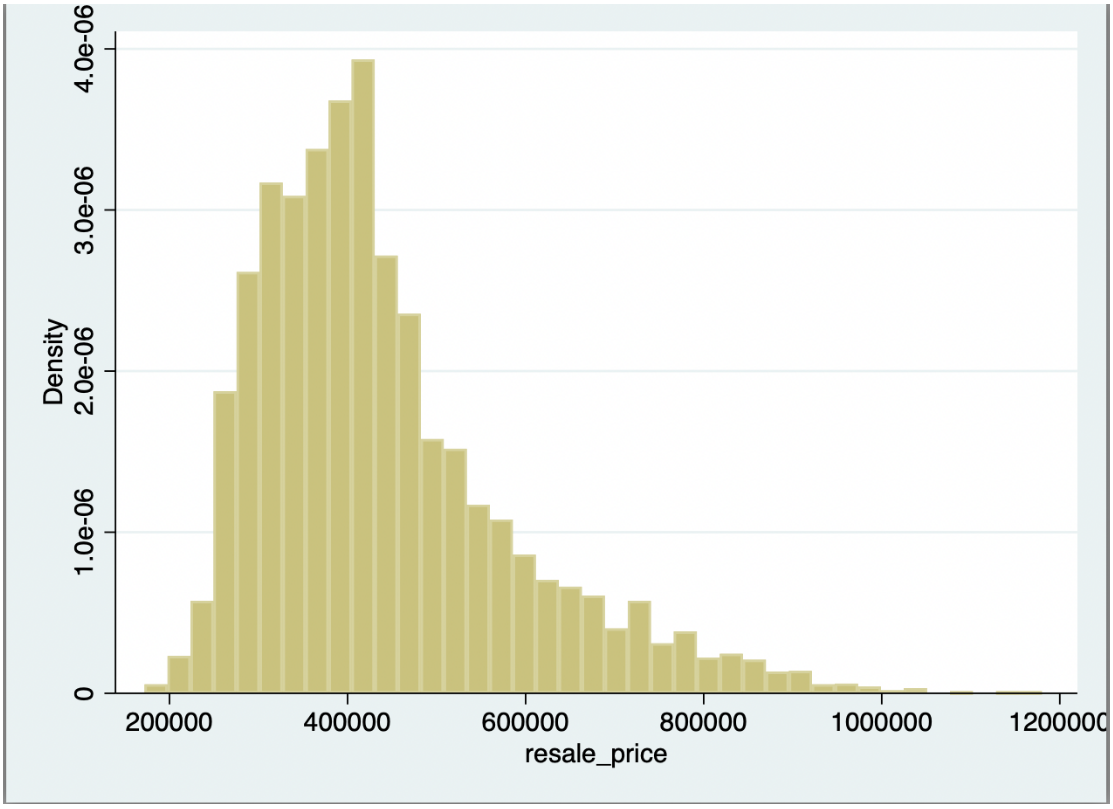
if $\beta_1 < 0$ $\beta_2 > 0$ $c_1 > 0$, α_1 over estimates the negative impact of size on price.

via (e) solution get

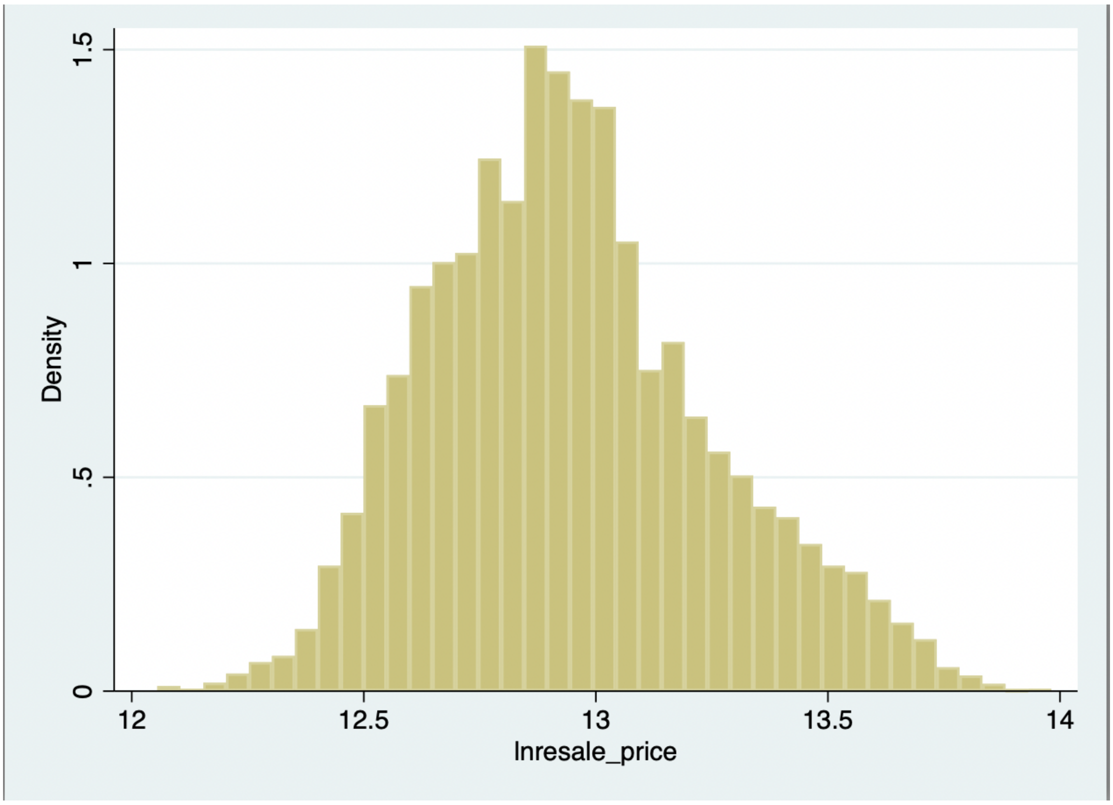
$E(\alpha_1) > \beta_1$ 向上偏误.

1

(c)



And then plot its log



(d)

(d) 1. reduce the absolute value.

2. the data is more stable, weakens the collinearity and heteroscedasticity

esale_price and its log.

3. show the economic implication of elasticity

ather than price in the regression.

flat_type to generate the number of rooms

ients and 7 rooms in MULTI-GENERATION

4. Distribution close to normal.

iswer by reporting the regression results in

(e)

Regression Results

	(1)	(2)
VARIABLES	Model 1	Model 2
lnfloor_area_sqm	-0.548***	-0.157***
	(0.027)	(0.009)
room_number	0.115***	
	(0.007)	
Constant	10.413***	9.106***
	(0.094)	(0.043)
Observations	8,360	8,360
R-squared	0.059	0.032

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

In that time, we adjust total price in original data to price per square feet as Y

Correlation between size and room number

```
. reg lnfloor_area_sqm room_number //变量间 回归

      Source |       SS           df       MS      Number of obs   =     8,360
-------------+----------------------------------   F(1, 8358)      =  62869.25
       Model |  473.186223         1  473.186223   Prob > F        =    0.0000
    Residual |  62.9065984     8,358  .007526513   R-squared       =    0.8827
-------------+----------------------------------   Adj R-squared   =    0.8826
       Total |  536.092822     8,359  .064133607   Root MSE        =    .08676

------------------------------------------------------------------------------
lnfloor_ar~m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 room_number |   .2594174   .0010346   250.74   0.000     .2573893    .2614455
       _cons |   3.478395   .0043836   793.50   0.000     3.469802    3.486988
------------------------------------------------------------------------------
```

```
. pwcorr lnfloor_area_sqm room_number, sig star(0.05) //correlation

             | lnfloo~m room_n~r
-------------+------------------
lnfloor_ar~m |   1.0000 
             |
             |
 room_number |   0.9395*  1.0000 
             |   0.0000
             |
```

	(1)
VARIABLES	Correlation
room_number	0.259***
	(0.001)
Constant	3.478***
	(0.004)
Observations	8,360
R-squared	0.883

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

We justify that $\beta 1<0$, $\beta 2>0$ $\mathrm{Cov}(\beta 1, \beta 2)>0$ so overestimate the negative impact.

2. The data set us_census.xlse is extracted from the US 1970 and 1980 census.

   (a) Pool the two years data together and generate a year variable equals 1970 for the 1970 census and 1980 for the 1980 census. Recode educ into a categorical variable, which takes the value of 1 for educ<12, 2 for educ==12, 3 for $12 < educ \leq 15$ and 4 for educ$\geq$ 16. Test whether the population share of college graduates (group 4) is the same between these two census. (a) 80 group 4 > 70 group 4 not same

   (b) Generate the summary statistics for weekly wage, educ, age and census years and reports the results in Table 2. ✓

   p<0.05 ?

   (c) Variable qob contains birth quarter information. Test whether people born in the first quarter are less educated than people born in other quarters in the 1970 census. (c) Yes, less educated in first quarter 1970.

   (d) Test whether people born in the last quarter are less educated than people born in the first quarter. (d) No, last quarter is better educated → But p>0.05 no significance difference

   (e) Generate potential years of experience using exp = age - edu - 5. ✓

   (f) Regress log wage on education, a quadratic function of potential years of experience. Report the regression results in column 1 of Table 3. ✓

   (g) Explain the coefficient on education? What is the impact of one year of experience in wage for workers with 10 years of experience?

   (h) Generate a census year dummy c80 = 1 for people observed in 1980 and add it as an additional control. Report the regression results in column 2 of Table3.

   (i) Comment on the difference in the coefficients on education between these two columns.

(g) When educ add 1 yrs, wage ↑ 6.6% within 10 yrs so just look exp, one year exp will increase 8.2% wage

2

(a)

```
. tab educ_cate if year==1970
```

educ_cate	Freq.	Percent	Cum.
1	1,999	40.43	40.43
2	1,580	31.96	72.39
3	589	11.91	84.30
4	776	15.70	100.00
Total	4,944	100.00	

```
. tab educ_cate if year==1980 //比较分类变量 比例是否相等
```

educ_cate	Freq.	Percent	Cum.
1	1,400	17.15	17.15
2	2,983	36.54	53.69
3	1,570	19.23	72.92
4	2,211	27.08	100.00
Total	8,164	100.00	

Not same, 1970 15.7%< 1980 27.08%

(b)

```
. summarize lwklywge educ age year //变量描述性统计
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lwklywge	13,108	5.57932	.7507315	-2.341806	9.028119
educ	13,108	12.57324	3.312294	0	20
age	13,108	40.91074	5.777106	30	50
year	13,108	1976.228	4.846976	1970	1980

	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max

Age	13,108	40.91	5.777	30	50
Years of schooling	13,108	12.57	3.312	0	20
Census year	13,108	76.23	4.847	70	80
Weekly wage	13,108	5.579	0.751	-2.342	9.028

(c)

```
. tabstat educ if year==1970&qob==1

    variable |      mean
-------------+----------
        educ |  11.33466
------------------------

. tabstat educ if year==1970&(qob!=1) //观察两条件组均值)

    variable |      mean
-------------+----------
        educ |  11.57793
------------------------

. reg educ dummyqob1 if year==1970 //虚拟变量 观察系数正负和P值，看看有没有显著
> 差异

      Source |       SS           df       MS      Number of obs   =     4,944
-------------+----------------------------------   F(1, 4942)      =      5.05
       Model |  55.4194104         1  55.4194104   Prob > F        =    0.0246
    Residual |  54209.2861     4,942  10.9690988   R-squared       =    0.0010
-------------+----------------------------------   Adj R-squared   =    0.0008
       Total |  54264.7055     4,943  10.9780913   Root MSE        =     3.312

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   dummyqob1 |   -.243273   .1082302    -2.25   0.025    -.4554523   -.0310938
       _cons |   11.57793   .0545295   212.32   0.000     11.47103    11.68484
------------------------------------------------------------------------------
```

P=0.025<0.05, there is a significant difference. When qob=1, edu will minus -0.24, which means less educated in quarter 1.

(d)

```
. tabstat educ if qob==4

    variable |      mean
-------------+----------
        educ |  12.65153
------------------------

. tabstat educ if qob==1 //观察条件组均值

    variable |      mean
-------------+----------
        educ |  12.49907
------------------------

. reg educ dummyqob4 if (qob==1)|(qob==4) //观察系数正负和P值，看看有没有显著差
> 异

      Source |       SS           df       MS      Number of obs   =     6,540
-------------+----------------------------------   F(1, 6538)      =      3.44
       Model |  37.9962626         1  37.9962626   Prob > F        =    0.0638
    Residual |  72274.6244     6,538  11.0545464   R-squared       =    0.0005
-------------+----------------------------------   Adj R-squared   =    0.0004
       Total |  72312.6206     6,539  11.0586666   Root MSE        =    3.3248

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   dummyqob4 |   .1524668   .0822385     1.85   0.064    -.0087476    .3136812
       _cons |   12.49907   .0586472   213.12   0.000      12.3841    12.61403
------------------------------------------------------------------------------
```

P=0.064>0.05. No significant difference. But from mean, quarter 4 is well educated than quarter 1. When qob=4, edu increases 0.15.

(e)

gen exp=age-educ-5 //声明新变量

(f)

	(1)
VARIABLES	Model 1
educ	0.066***
	(0.002)
exp	0.032***
	(0.005)

exp2	-0.001***
	(0.000)
Constant	4.616***
	(0.063)
Observations	13,108
R-squared	0.164

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

(g) When education add 1 years, wage increases 6.6%. We can conclude that edu has positive relationship to wage from (f).

```
. display "the impact of one year exp to wage with 10 years exp^2=" _b[exp]+2*_b
> [exp2]*10
the impact of one year exp to wage with 10 years exp^2=.01147372
```

With 10 years' experience, we should look the derivative of model 1 by exp at 10. So one year exp will increase 1.1% wages.

(h)

	(1)	(2)
VARIABLES	Model 1	Model 2
educ	0.066***	0.079***
	(0.002)	(0.002)
exp	0.032***	0.056***
	(0.005)	(0.004)
exp2	-0.001***	-0.001***
	(0.000)	(0.000)
yrdummy		0.654***
		(0.014)
Constant	4.616***	3.435***
	(0.063)	(0.063)

Observations	13,108	13,108
R-squared	0.164	0.290

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

(i)

As you can see, when add dummy variable, the coefficient of education is increasing from 0.066 to 0.079, and the R-squared increases, showing that the model interpretation increases.

Also, the coefficient of education is increasing. It may be when control time-year effect, model 2 alleviated the endogenous problem of mutual cause and effect that caused the negative impact on education in the 1980s due to low wages in the 1970s. 1980 wages is higher than 1970 possibly due to inflations, so we should distinguish two years to make other variable more robust.

It also shows that 80s achieve better education linked to the wage, and positive shock on wage happened in 1980, implying some event happened and 1980 matters education more on wage.

Add yrdummy80 may control for the potential erogeneity of education as year growing to 1980 from 1970, people are also more likely to received longer education than before