Applied Statistics | NUEZ Agency Case – TASK 2 | Shayan Abbasi & Ivan Gutiérrez

Q1)

Command:

ttest age == 8

Result: The null hypothesis is that the population's age variable mean is equal to 8 and considering 95% confidence, we reject the null hypothesis. We can also draw the same conclusion by looking at the confidence interval. Confidence interval mentions that the population's mean is between the two below values given 95% confidence.

```
One-sample t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
age	518	6.175676	.0915994	2.084767	5.995723	6.355629

```
    mean = mean(age)                                      t = -19.9163
Ho: mean = 8                                 degrees of freedom =      517

    Ha: mean < 8                 Ha: mean != 8                 Ha: mean > 8
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Q2)

Command:

prtest agency == 0.5

Result: Null hypothesis asserts that the proportion of the house sold by Nuez is 50% and we reject the null hypothesis.

```
One-sample test of proportion                 Number of obs      =       518
```

Variable	Mean	Std. Err.	[95% Conf.	Interval]
agency	.0945946	.0128585	.0693924	.1197968

```
    p = proportion(agency)                                z = -18.4537
Ho: p = 0.5

    Ha: p < 0.5                  Ha: p != 0.5                  Ha: p > 0.5
 Pr(Z < z) = 0.0000         Pr(|Z| > |z|) = 0.0000          Pr(Z > z) = 1.0000
```

Q3)

Command:

tabulate area month, cell chi2 column expected row

Result: The first row of each cell shows the frequency of houses sold in a specific area and a specific month **(first part of the question)**. We can conduct a χ^2 test using the observed frequencies and expected frequencies which test the null hypothesis (The population means are not significantly different). Finally, **we couldn't reject the null hypothesis**. **(Second part of the question)**

Key
frequency *expected frequency* *row percentage* *column percentage* *cell percentage*

Area where the house is	Month of the sale				
	March	April	May	June	Total
Granollers	73	60	52	58	243
	65.7	61.5	59.6	56.3	243.0
	30.04	24.69	21.40	23.87	100.00
	52.14	45.80	40.94	48.33	46.91
	14.09	11.58	10.04	11.20	46.91
Manresa	42	43	41	40	166
	44.9	42.0	40.7	38.5	166.0
	25.30	25.90	24.70	24.10	100.00
	30.00	32.82	32.28	33.33	32.05
	8.11	8.30	7.92	7.72	32.05
Other	25	28	34	22	109
	29.5	27.6	26.7	25.3	109.0
	22.94	25.69	31.19	20.18	100.00
	17.86	21.37	26.77	18.33	21.04
	4.83	5.41	6.56	4.25	21.04
Total	140	131	127	120	518
	140.0	131.0	127.0	120.0	518.0
	27.03	25.29	24.52	23.17	100.00
	100.00	100.00	100.00	100.00	100.00
	27.03	25.29	24.52	23.17	100.00

Pearson chi2(6) = 5.2201 Pr = 0.516

Q4)

Command:

tabulate agency area, cell chi2 column expected row

Result: As we expected conventional real state agencies are local and therefore, the population mean (which here means proportion) could be different between different categories. Naturally, each agency focuses on some area more than others. Here, we are testing whether NUEZ have different proportion of houses sold in each area than the general "Other" agencies. The **p-value is significant**. We reject the null hypothesis, therefore, **the portion of houses Nuez sold in different areas are different with Other's.**

Key
frequency *expected frequency* *row percentage* *column percentage* *cell percentage*

Agency that sold the house	Area where the house is			
	Granoller	Manresa	Other	Total
Other	236	139	94	469
	220.0	150.3	98.7	469.0
	50.32	29.64	20.04	100.00
	97.12	83.73	86.24	90.54
	45.56	26.83	18.15	90.54
Nuez	7	27	15	49
	23.0	15.7	10.3	49.0
	14.29	55.10	30.61	100.00
	2.88	16.27	13.76	9.46
	1.35	5.21	2.90	9.46
Total	243	166	109	518
	243.0	166.0	109.0	518.0
	46.91	32.05	21.04	100.00
	100.00	100.00	100.00	100.00
	46.91	32.05	21.04	100.00

Pearson chi2(2) = 23.6122 Pr = 0.000

Q5)

Command:

oneway price agency

Result: First, based on Bartlett's test, we cannot reject that the variance of two price sets related for each agency (other or Nuez) is the same. Thus, we can carry on the ANOVA. The null hypothesis asserts that there is no different population means of prices among agencies. The p-value in ANOVA shows that **we can reject the null hypothesis**, meaning that **population average prices are different between Nuez and others**.

```
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      2.1773e+13      1   2.1773e+13      6.82     0.0093
 Within groups      1.6461e+15    516   3.1902e+12
------------------------------------------------------------------------
    Total           1.6679e+15    517   3.2261e+12

Bartlett's test for equal variances:  chi2(1) =   3.2043  Prob>chi2 = 0.073
```

Moreover, we can conduct the same analysis with t-test and have a more detailed result on the difference of means within each category as well.

Q6)

Command: oneway price month

Result: First, based on Bartlett's test, we cannot reject that the variance of price sets related for each month is the same. Thus, we can carry on the ANOVA. The null hypothesis asserts that there is no different in population's mean prices and the month the houses were sold in. The p-value in ANOVA shows that **we can reject the null hypothesis**, meaning **that there is a different in average prices of sold houses among months**.

```
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1.0689e+14      3   3.5632e+13     11.73     0.0000
 Within groups      1.5610e+15    514   3.0370e+12
------------------------------------------------------------------------
    Total           1.6679e+15    517   3.2261e+12

Bartlett's test for equal variances:  chi2(3) =   1.5139  Prob>chi2 = 0.679
```

In the next step, we can see preform a post-hoc test, showing the significant difference between groups:

```
. pwmean price, over(month) mcompare(tukey) effects

Pairwise comparisons of means with equal variances

over          : month
```

	Number of Comparisons
month	6

price	Contrast	Std. Err.	Tukey t	Tukey P>|t|	Tukey [95% Conf.	Interval]
month						
April vs March	-344257.4	211838.8	-1.63	0.365	-890262.4	201747.7
May vs March	-811898.8	213555.3	-3.80	0.001	-1362328	-261469.6
June vs March	-1197976	216796.7	-5.53	0.000	-1756760	-639192.5
May vs April	-467641.4	217016.6	-2.15	0.137	-1026992	91709.06
June vs April	-853718.8	220207	-3.88	0.001	-1421293	-286145.1
June vs May	-386077.4	221858.8	-1.74	0.304	-957908.4	185753.5

Based on the pairwise mean comparison, we can see that pairs May-March, June_March, and June_April have significant differences in mean prices for the houses sold.

Q7)

Command: oneway room area

Result: First, based on Bartlett's test, we cannot reject that the variance of price sets related for each month is the same. The null hypothesis asserts that there is not a difference between the population's average number of rooms and the area that the sold houses were located in. The **f-test rejects the null hypothesis with 95% confidence interval**, however, **the question asks for 99% confidence** which requires a p-value below 1% for null hypothesis rejection. Since the null hypothesis is not rejected, **we cannot say that there is a difference in average number of rooms among zones.**

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	3.45878889	2	1.72939445	4.42	0.0125
Within groups	201.631945	515	.391518339		
Total	205.090734	517	.396693875		

```
Bartlett's test for equal variances:  chi2(2) =   2.7255  Prob>chi2 = 0.256
```

```
. pwmean room, over(area) mcompare(tukey) effects

Pairwise comparisons of means with equal variances

over         : area
```

	Number of Comparisons
area	3

rooms	Contrast	Std. Err.	Tukey t	Tukey P>|t|	Tukey [95% Conf.	Interval]
area						
Manresa vs Granollers	.0536219	.0630058	0.85	0.671	-.0944737	.2017175
Other vs Granollers	.2138408	.0721325	2.96	0.009	.0442928	.3833887
Other vs Manresa	.1602189	.0771392	2.08	0.096	-.0210974	.3415351

Although, the mean differences where not significant at 99% confidence, preforming pairwise mean comparison can tell us under a milder confidence, which pair had significantly different means (Other-Granollers).

Q8) It's better to llok at Q8-11 as a single thred.

Command: pwcorr room price, obs sig star(5)

Result: First, we calculate correlation coefficient, then we perform a simple OLS regression and test R^2 and the β. There is 0.4331 positive correlation between the number of rooms and sale's price. Moreover, we reject the null hypothesis (asserting that the correlation is zero).

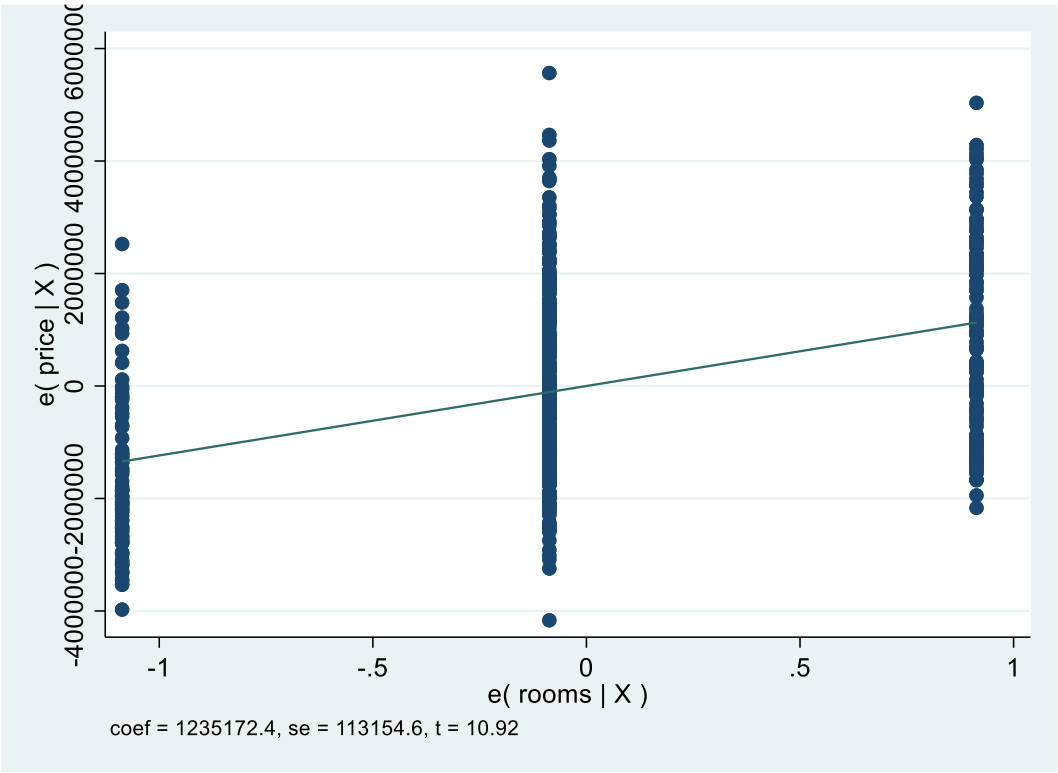
	rooms	price
rooms	1.0000	
	518	
price	0.4331*	1.0000
	0.0000	
	518	518

As we all know, correlations do not necessarily induce causation, and by relationship, the question may concern correlation or explanation(causation). Therefore, we look at the OLS regression. Room's beta show that there is a positive relation between price and number of rooms. Room's beta (or even the intercept) is significant and F-test shows that R^2 is significant too. For simple linear regression, R^2 is the square of the sample correlation r_{xy}, showing that room variance explains 18.7% of variability in price variable. The f-test shows if the independent variables used in the model are jointly significant.

```
. regress price rooms
```

Source	SS	df	MS	Number of obs	=	518
				F(1, 516)	=	119.15
Model	3.1290e+14	1	3.1290e+14	Prob > F	=	0.0000
Residual	1.3550e+15	516	2.6260e+12	R-squared	=	0.1876
				Adj R-squared	=	0.1860
Total	1.6679e+15	517	3.2261e+12	Root MSE	=	1.6e+06

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rooms	1235172	113154.6	10.92	0.000	1012872	1457473
_cons	5532682	356476.7	15.52	0.000	4832358	6233006



But how we can say that we have the right model specification? How can we say that the relationship is linear? Looking at the regression line and scattered data points we see that the models does not fit the data at all and only shows a positive relationship.

Q9)

First, we can look at the correlation between price and size:

	price	size
price	1.0000	
	518	
size	0.7121*	1.0000
	0.0000	
	518	518

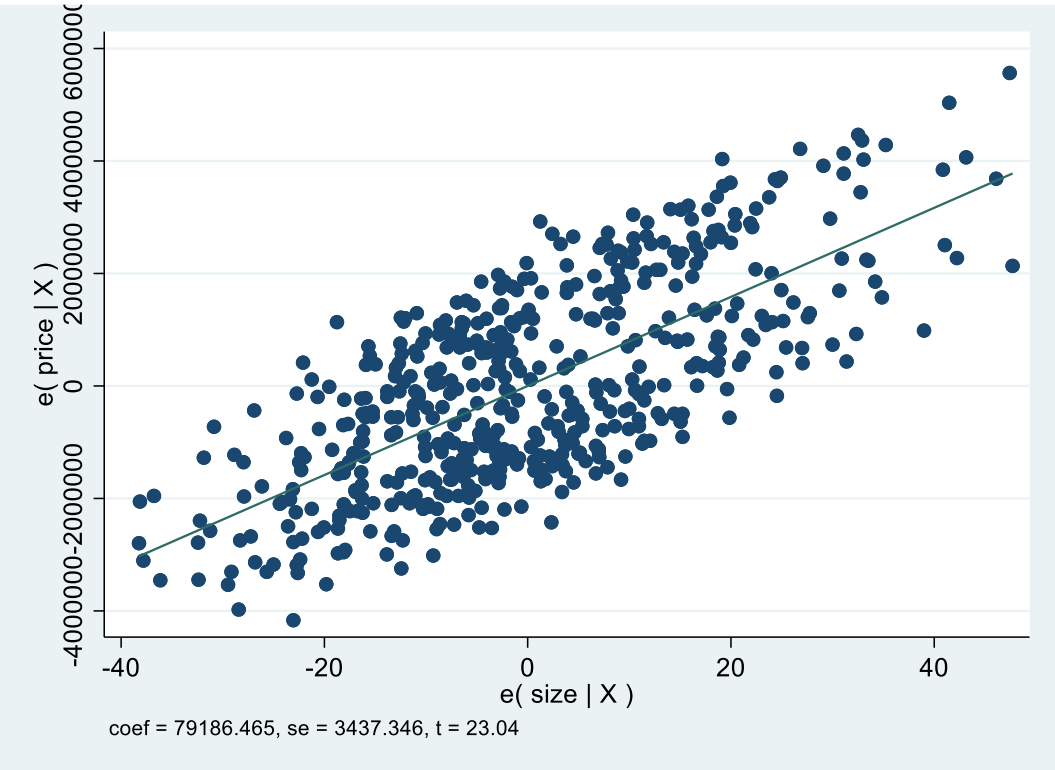
We can have the similar conclusion for the correlation results at Q8.

Moreover, we can do similar analysis for the causation relationship between price and the size of the house. First, we can have a simple regression:

```
. regress price size
```

Source	SS	df	MS	Number of obs	=	518
				F(1, 516)	=	530.71
Model	8.4567e+14	1	8.4567e+14	Prob > F	=	0.0000
Residual	8.2223e+14	516	1.5935e+12	R-squared	=	0.5070
				Adj R-squared	=	0.5061
Total	1.6679e+15	517	3.2261e+12	Root MSE	=	1.3e+06

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
size	79186.46	3437.346	23.04	0.000	72433.55	85939.38
_cons	2293874	311083.1	7.37	0.000	1682729	2905019



The plot shows approximately linear specification, thus, we don't go through log-log specification or any other specification.

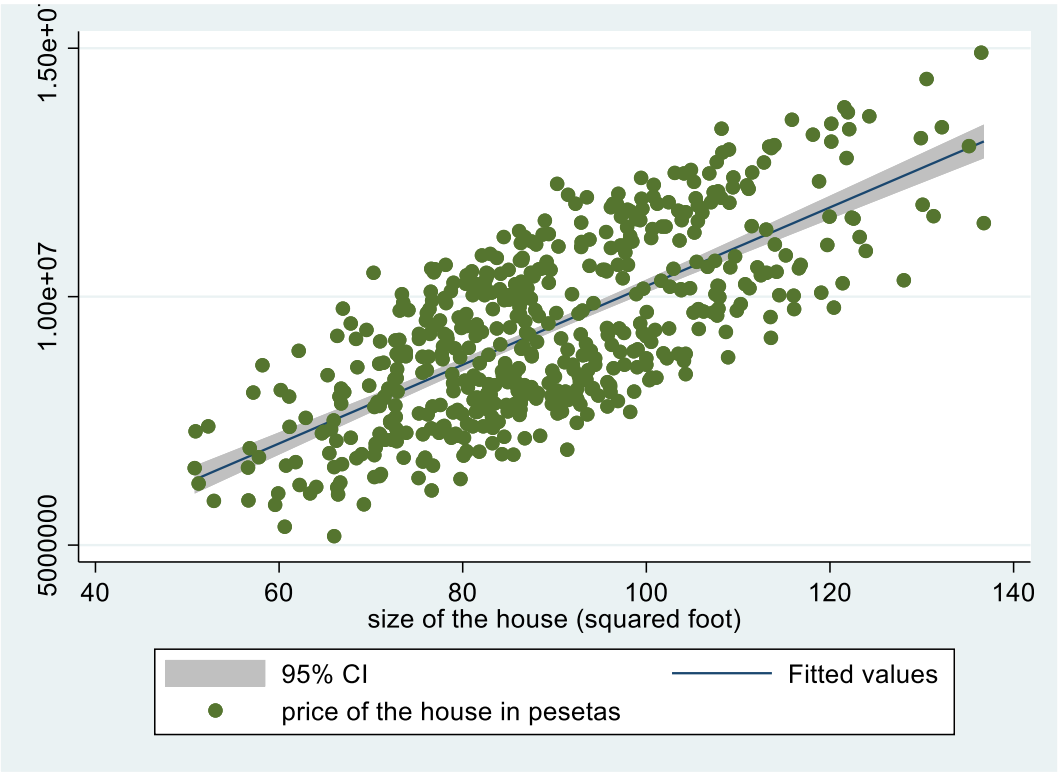
Q10) Therefore, we can say that 50% of variability in price is explained by size based on the naive model that we made.

However, what if other variables are causing change in price that are causing size too? In other words, is there an omitted variable bias? Isn't number of rooms highly correlated with house size? Don't houses with high number of rooms tend to have bigger sizes?

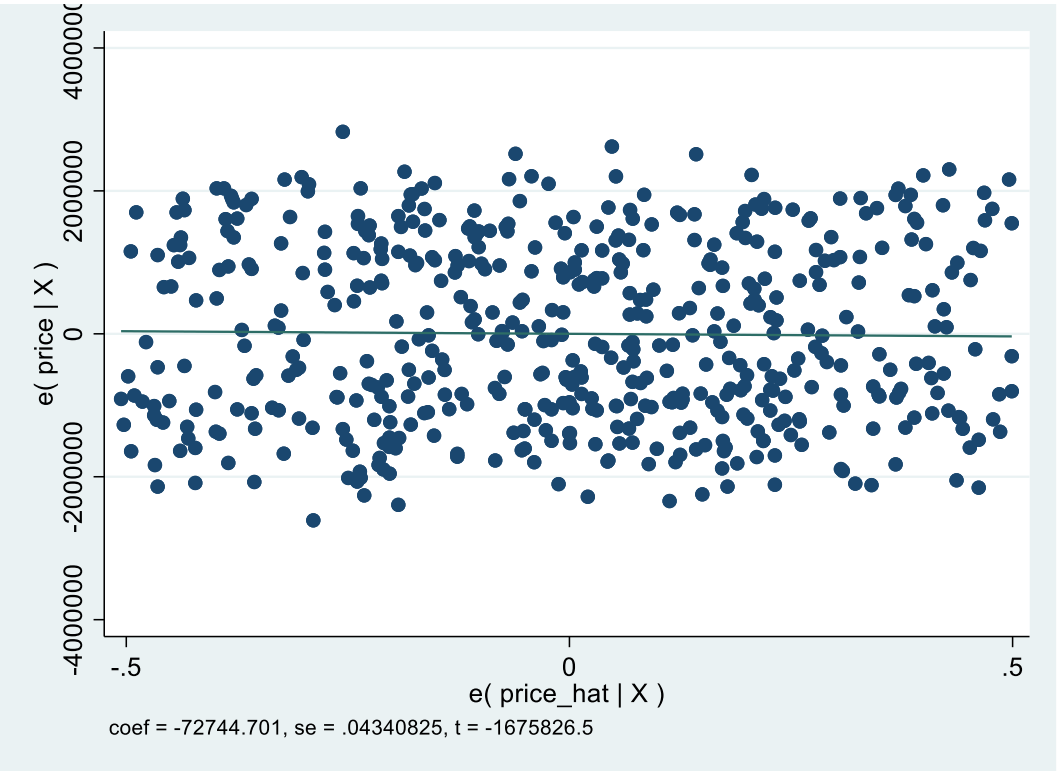
```
. pwcorr room size, obs sig star(5)

             |    rooms     size
-------------+--------------------
       rooms |   1.0000
             |      518
             |
        size |   0.6439*  1.0000
             |   0.0000
             |      518      518
```

As we can see, rooms and size are considerably correlated. Therefore, we shouldn't add rooms as control variable to our model. However, we can add other control variables, like neighborhood.

Q11) Based on the regressions, we can predict price and considering the root MSE as a sort of accuracy measure. However, the accuracy is measured in-sample. The important thing in prediction, is that we can train our model with in-sample data and test it with out of sample data, trying to avoid over-fitting/under-fitting and remain at a high level of accuracy. The simple regression of price ~ size has the following illustration:



We can also show the error (y-y_pred) for the trained model on all the data as the following:



There is no wrong prediction here, however, the important thing is how accurate we are. Here we have up to approximately 200,000 error for some data points which is not perfect at all. Moreover, the MSE in this regression does not satisfy me at all. I believe with a multivariate regression we can achieve a better MSE and a robust prediction that does not under-fit like the simple regression.

Moreover, we can train the model for have of the randomly selected data and predict the data for the other half (using it as test data for the purpose of checking for under fitting or over fitting or any extrapolation problem).

Commands:

set seed 1234

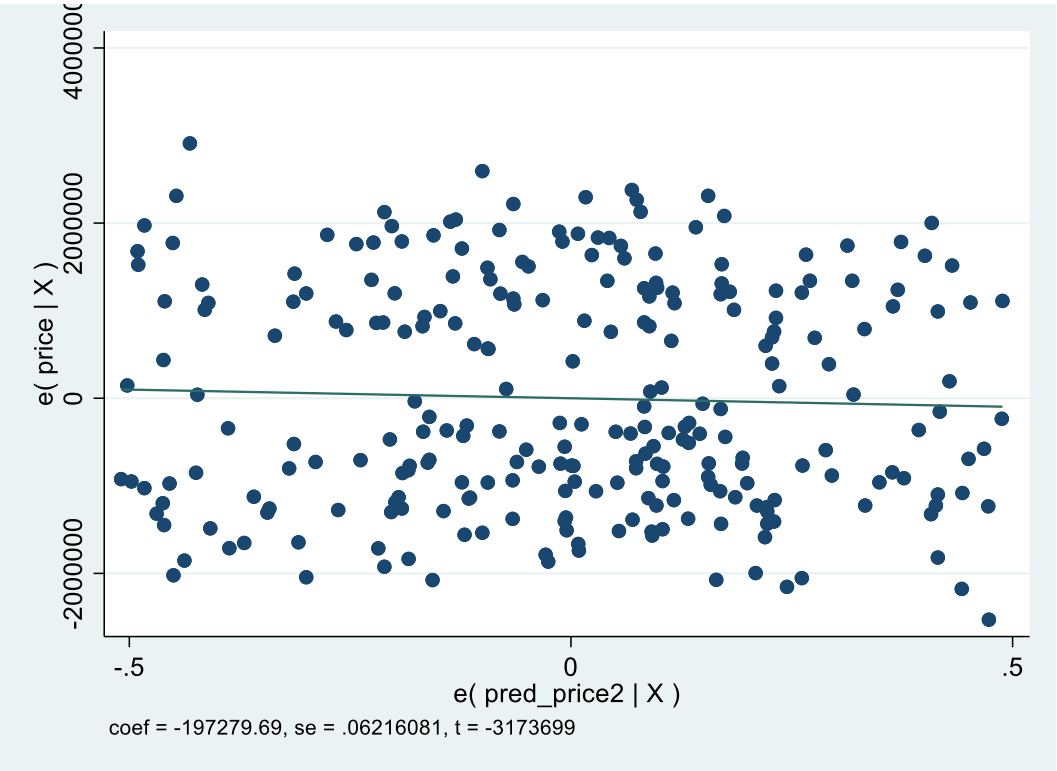
g train = (runiform() > 0.5)

regress price size if train

predict pred_price

predict resid_price, residuals

avplot pred_price



However, the error term scatters does not seem so much different, we can always calculate MSE or other similar measures for in-sample and out-of-sample data to have a better understanding of the model prediction power and possible over-fits.

Notice that in the previous regressions, for the purpose of simplicity, we didn't check for violations in other OLS assumptions, and we did not add a lot of control variables or check for model's robustness. Adding the agency dummy to the model can be beneficial to resolving the Nuez dispute with costumers, however, I guess that we are going to do that in the third assignment.