Application Task: Prof. Qian, Spring 2022 the U.S. housing market: Patterns in Trading Records from Homesnap.com

Jiawei Yang

May 25, 2022

Task 1: Event-level aggregated summary statistics

- The raw data set contains 556 observations each specific to one action in one bargaining event: the overall goal in this section is to aggregate the observations so that each entry represents one event. With the aggregated event-level data, we then obtain the summary statistics for the following features: property age (in years), number of buyers, number of price revisions, duration until Off-Market (in days), sales prices (inflation corrected).

- The wanted features are not given directly in the raw data set, expect for the deflated sales price. The following are specific procedures used to derive them:

  Property Age: calculated by taking the difference between 2022 and year of the property built

  Number of buyers: by counting number of unique buyer ids per event

  Price revision count: by counting number of unique buyer ids per event

  Duration: by taking the difference between listing date and sales date. Rounded down to the nearest whole day.

- abnormal data point: one property features same-day listing and sales in the data set, the listing duration is manually corrected to be 0 day.

- the summary statistics are as the following:

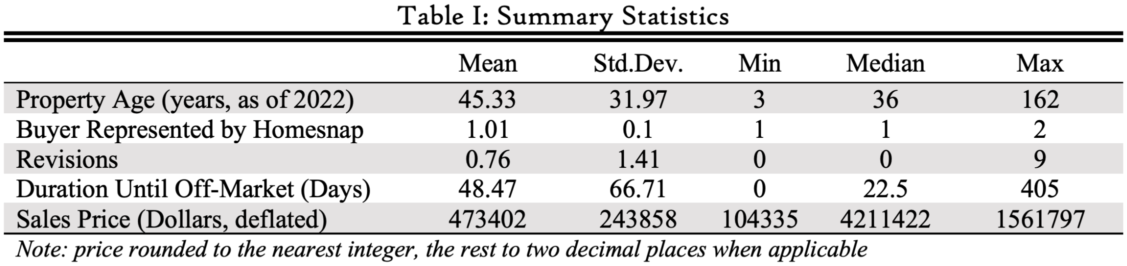


Table I: Summary Statistics

	Mean	Std.Dev.	Min	Median	Max
Property Age (years, as of 2022)	45.33	31.97	3	36	162
Buyer Represented by Homesnap	1.01	0.1	1	1	2
Revisions	0.76	1.41	0	0	9
Duration Until Off-Market (Days)	48.47	66.71	0	22.5	405
Sales Price (Dollars, deflated)	473402	243858	104335	4211422	1561797

Note: price rounded to the nearest integer, the rest to two decimal places when applicable

Task 2: Geographical distribution of the bargaining events

- This section plots the geographical distribution of the bargaining events in the data set.
- The location information is extracted from the ***censustract*** column in the data set, where the first two digits represents the FIPS code of the state where the bargaining event/action took place.
- The FIPS code contains higher level of specification: the 3th-5th digits contains county code. However, given the low granularity of the data set that only around 500 actions are recorded out of 200 events taking place in less than 30 states, the map in this section are specified only to a state-level.
- The choropleth:

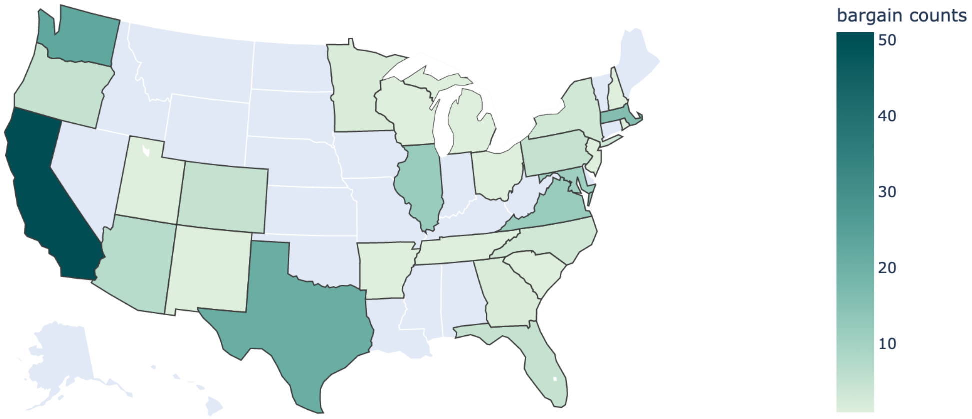


Figure 1: Event count by U.S. State

Task 3/4/5: Regression Analysis on Modeling Property Sales Price

- This section explores the possible factors that influence the final deflated sales price of a property with regression models. To begin with, I categorize the variables recorded in the data set into two types.

  Market Reaction Indicators (M.R.I): variables that describe the market interactions regarding the selling of this property (Such as initial listing price, length of listing period, number of buyers attracted)

  Property Feature Indicators (P.F.I): variables that records intrinsic attributes related to the property. (Such as property age, area, number of rooms)

- While it is reasonable to assume that both types of indicators pertain to the final sales price, I choose not to include them simultaneously in regression analysis to begin with.

  The reason is the concern of bad control bias. My claim is the causal relationship between **M.R.I** and **P.F.I**: specifically **P.F.I** cause **M.R.I** to a great extent while both **P.F.I** and **M.R.I** cause final sales price. An interesting fact to note is that the uni-variate regression of final sales price on initial listing price yields a r-squared value over 0.9, which effectively proves any correlation with the final sales price can be effectively mirrored by that with the initial listing price. Formally, **P.F.I** feature the most endogenous characteristic of one property which factor to different extent into the decision making into the trading activities related to the property, and the trading activities are represented by **M.R.I**.

  A detailed example would be to consider the interaction between area of the property (as an instance of P.F.I) and initial listing price (as an instance of M.R.I). It makes perfect sense for the sellers, they place higher initial listing price on larger properties; on the other hand, the final sales price is higher for larger properties as well. In this case, we may consider the final sales price as the collider variable. The related inference and prediction would be rendered invalid upon considering two inter-correlated causal factors of the collider.

- As a result, I assume there are two different data generating processes (or casual inference channels)to be estimated independent of each other. Simply put

  Market Determination: sales price $= f(M.R.I)$

  Feature Determination: sales price $= f(P.F.I)$

- The above model requires strong assumption which will be loosened in Task 6.

- General variable-selection/model-optimization strategy: iteratively including variable with highest marginal R-square. Regardless of the model, I adapt this greedy algorithm: given $(x_1, x_2, ...x_k)$ selected, regress the outcome on $(x_1, x_2, ...x_k)$ plus one additional variable from the remaining $(x_{k+1}, ..., x_n)$ select x_r from the remaining such that the R-square of $(x_1, x_2, ...x_k, x_r)$ model beats the other candidates. The model starts as an uni-variate model with the single most predictive independent variable, then grows to bi-variate, tri-variate and eventually n-variate. This strategy gives an unique ordering of including of the regressors. The last step is to cut the model by excluding the regressors whose inclusion contributes little to the r-square value.

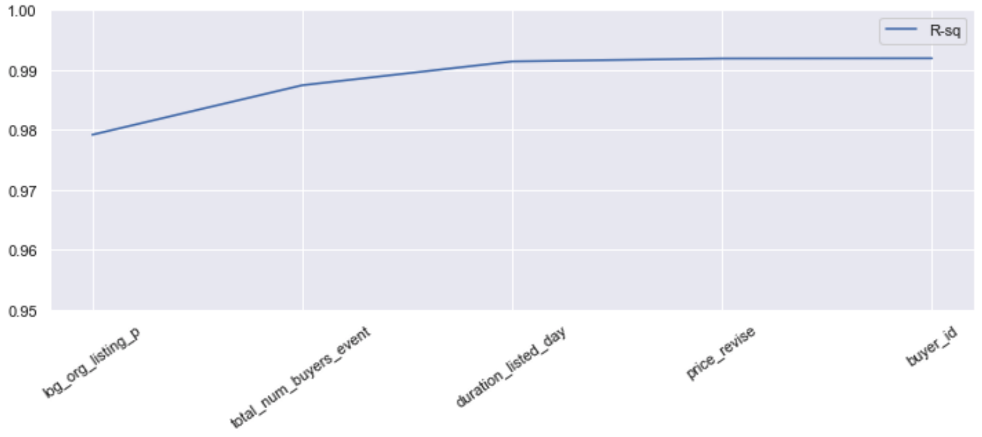
Task 3/4/5: Procedure, Results and Analysis

- Additional variables have been included by feature engineering. *log_de_p*, *log_org_listing_p* and *log_sq_fr* are respectively the log transformed deflated final sales price, initial listing price and square foot area of the property due to the high skewness in the original distribution. A side-effect to the log transformation is the improvement in the fit since sparsity is reduced. *score_trspt* is the unweighted summation of bike, walk, and transit score as they are highly inter-correlated and could be reasonably represented by this overall index of transportation accessibility.

- Pair-wise variable pre-screening has been done to observe the overall correlation pattern and filter out variables that are evidently uncorrelated with final sales price. (See pair maps in appendix I) Homoskedasticity is assumed as a result for later analysis.

- **the pure Market Determination Model**

  - Marginal Variable-Rsq Optimization:



  This is less of an interesting case where we see the selection process immediately captures log listing price as the first predictor while explaining around 98% change in the log final sales price. The discussion of cutting extra predictors is trivial in that any marginal improvement is almost not meaningful given the over powerful first predictor. However, I still include total number of buyers and duration listed day.

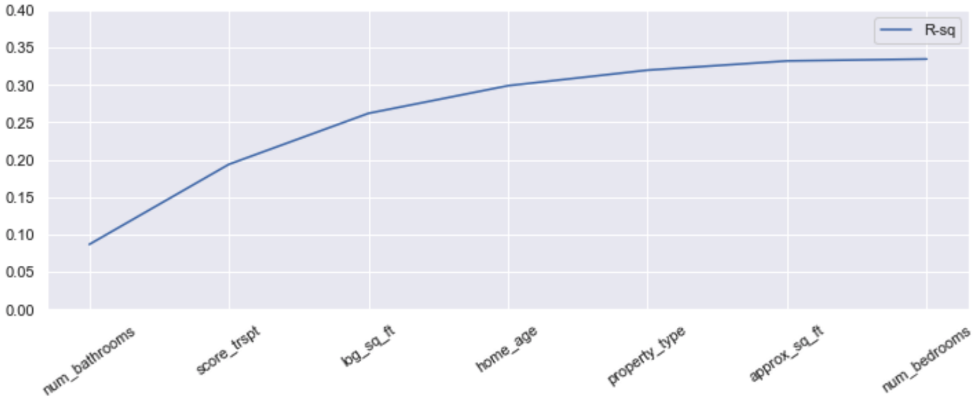
  - Model and Interpretation:

$$\hat{log\_de\_price} = \hat{\beta}_0 + \hat{\beta}_1 log\_listing\_price + \hat{\beta}_2 listing\_duration + \hat{\beta}_3 num\_buyers\_event \qquad (1)$$

  Check Appendix II for the whole regression result table. I find the following significant, at 95% level, predictors. Every 1 percent change in the deflated initial listing price is associated with a 1.02 percent approximately change in the deflated final sales price, ceteris paribus. One more buyer present in the event (buyers in general, not limited those who use Homesnap.com) is associated with a 0.8% increase in the deflated final price, ceteris paribus. We therefore discover the fact that the listing price dominantly influences predicts the final sales price, and the presence of a biding effect to push up the price in the presence of more buyers competing per event.

- **the pure Feature Determination Model**

  - Marginal Variable-Rsq Optimization:



We see that the predicting power of the model peaks out at around 30% R-squared up to including property type. Before the peak, each variable contributes similar marginal predictive power. I include variables up to (and including) property type.

  - Model and Interpretation:

$$log\_de\hat{\_}price = \hat{\beta}_0 + \hat{\beta}_1 num\_bathrooms + \hat{\beta}_2 score\_trspt + \hat{\beta}_3 log\_sq\_ft + \hat{\beta}_4 home\_age + \hat{\lambda_{pt}} \quad (2)$$

Note the $\hat{\lambda_{pt}}$ denotes the intercept change by different property types, analogous to fixed effects. Check appendix for the whole regression result table.
The overall fit or predicting power of the pure feature determination model is much weaker: the model only explains roughly 30% of the variation in the deflated final sales price. There are two significant predictors identified. For every one percent increase in the living area of the property, we expect to see an 0.61 percent increase in the final sales price deflated, ceteris paribus. For every 1 unit increase in the overall transportation score (in the unit it is collected), we expect the final sales price deflated to increase by 0.25 percent. Note that the estimator for overall transportation score is barely significant at 95%level. This regression model gives evidence for the common sense that the property price is proportional to its area, which is significant in effect size.

Task 6: Discussion, Limitations, Improvement

- As discussed in Task 3/4/5 methodology part, the binary partition of variables in the data set is a very strong assumption whose only validity is still to be tested. On the other hand, one would still consider the potential deficiency of models restricted by such set of assumptions. One source of concern is omitted variable bias. It would make sense for now to loosen the assumption when a good balance between the omitted variable bias and bad control bias can be reached. Ideally, we want to include as many regressors as possible while not biasing the model. I therefore propose the mixed model that both P.F.I and M.R.I factor into but seeks to minimize collision by excluding the initial listing price. The specifics of this mixed model are presented in Appendix III: it follows nevertheless the variable selection and model optimization strategy just with a new set of mixed variables. The particular specification of mixed model I suggested improved R-sq to around 40%.

- The data set itself contains limited observation with non-trivial missingness: as the number of predictors increases the effective sample size in regression decreases considerably

- More complicated functional form of regression could be explored. In this demo, only log transformation is presented. Additional feature engineering could also be an option to test the model robustness in response to different indicators when they practically quantify the same underlying feature.

Appendix I: Pair Plots

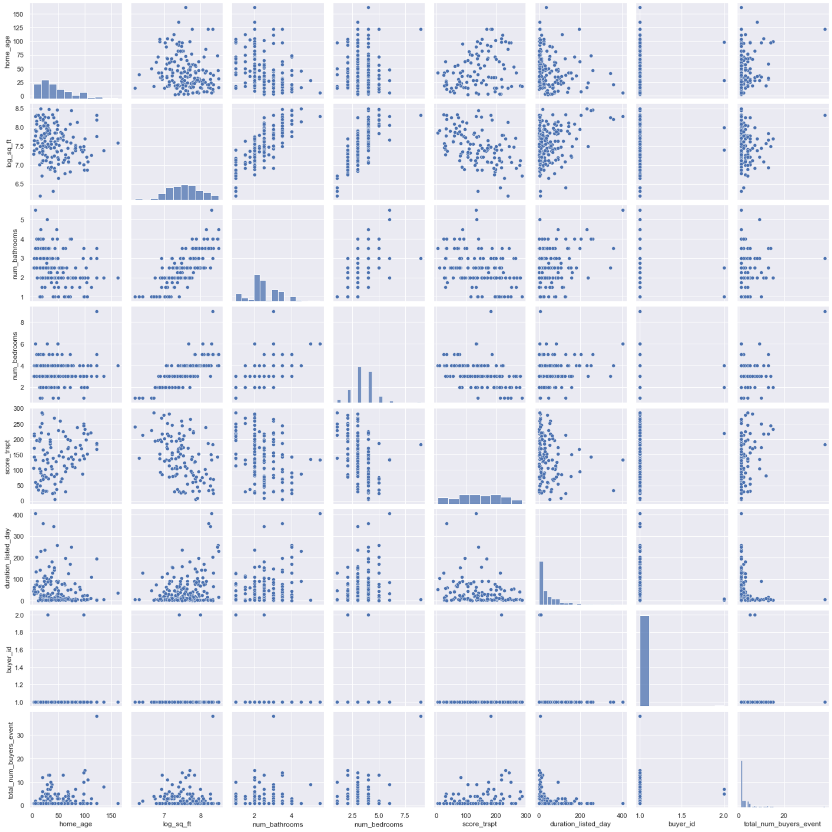


Figure 2: Pair Scatter Plot

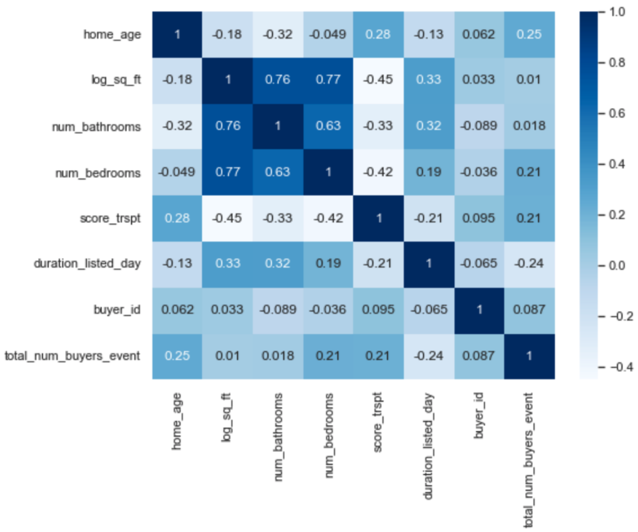
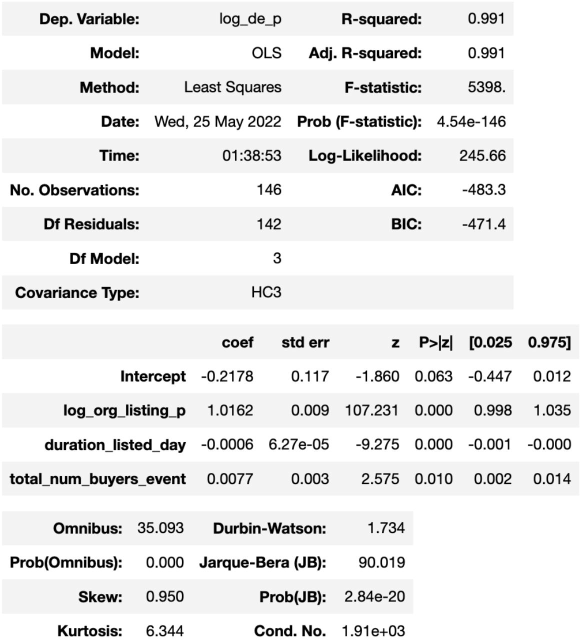


Figure 3: Correlation Heat Map

Appendix II: Regression Tables

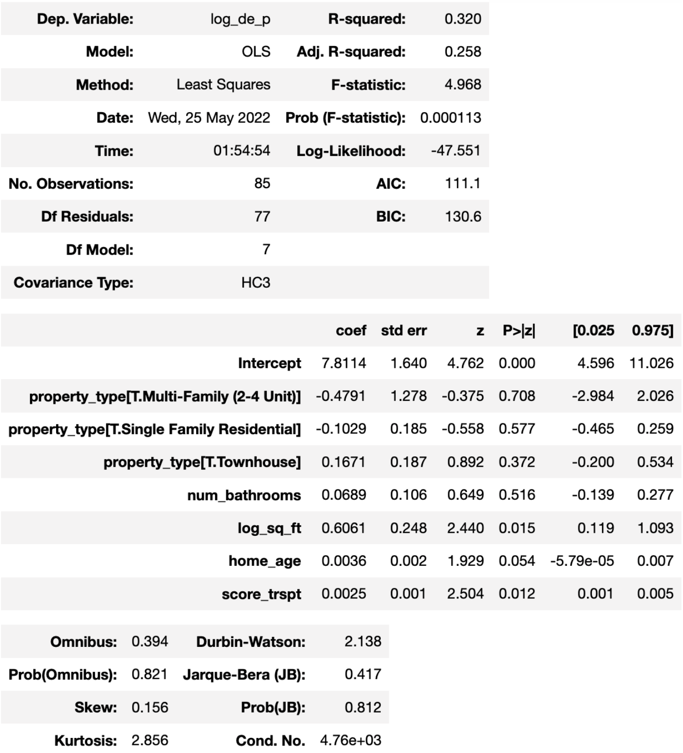


Dep. Variable:	log_de_p	**R-squared:**	0.991
Model:	OLS	**Adj. R-squared:**	0.991
Method:	Least Squares	**F-statistic:**	5398.
Date:	Wed, 25 May 2022	**Prob (F-statistic):**	4.54e-146
Time:	01:38:53	**Log-Likelihood:**	245.66
No. Observations:	146	**AIC:**	-483.3
Df Residuals:	142	**BIC:**	-471.4
Df Model:	3		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-0.2178	0.117	-1.860	0.063	-0.447	0.012
log_org_listing_p	1.0162	0.009	107.231	0.000	0.998	1.035
duration_listed_day	-0.0006	6.27e-05	-9.275	0.000	-0.001	-0.000
total_num_buyers_event	0.0077	0.003	2.575	0.010	0.002	0.014

Omnibus:	35.093	**Durbin-Watson:**	1.734
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	90.019
Skew:	0.950	**Prob(JB):**	2.84e-20
Kurtosis:	6.344	**Cond. No.**	1.91e+03

Figure 4: Regression Results for Model (1)



Dep. Variable:	log_de_p	**R-squared:**	0.320
Model:	OLS	**Adj. R-squared:**	0.258
Method:	Least Squares	**F-statistic:**	4.968
Date:	Wed, 25 May 2022	**Prob (F-statistic):**	0.000113
Time:	01:54:54	**Log-Likelihood:**	-47.551
No. Observations:	85	**AIC:**	111.1
Df Residuals:	77	**BIC:**	130.6
Df Model:	7		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	7.8114	1.640	4.762	0.000	4.596	11.026
property_type[T.Multi-Family (2-4 Unit)]	-0.4791	1.278	-0.375	0.708	-2.984	2.026
property_type[T.Single Family Residential]	-0.1029	0.185	-0.558	0.577	-0.465	0.259
property_type[T.Townhouse]	0.1671	0.187	0.892	0.372	-0.200	0.534
num_bathrooms	0.0689	0.106	0.649	0.516	-0.139	0.277
log_sq_ft	0.6061	0.248	2.440	0.015	0.119	1.093
home_age	0.0036	0.002	1.929	0.054	-5.79e-05	0.007
score_trspt	0.0025	0.001	2.504	0.012	0.001	0.005

Omnibus:	0.394	**Durbin-Watson:**	2.138
Prob(Omnibus):	0.821	**Jarque-Bera (JB):**	0.417
Skew:	0.156	**Prob(JB):**	0.812
Kurtosis:	2.856	**Cond. No.**	4.76e+03

Figure 5: Regression Results for Model (2)

Appendix III: Regression Tables

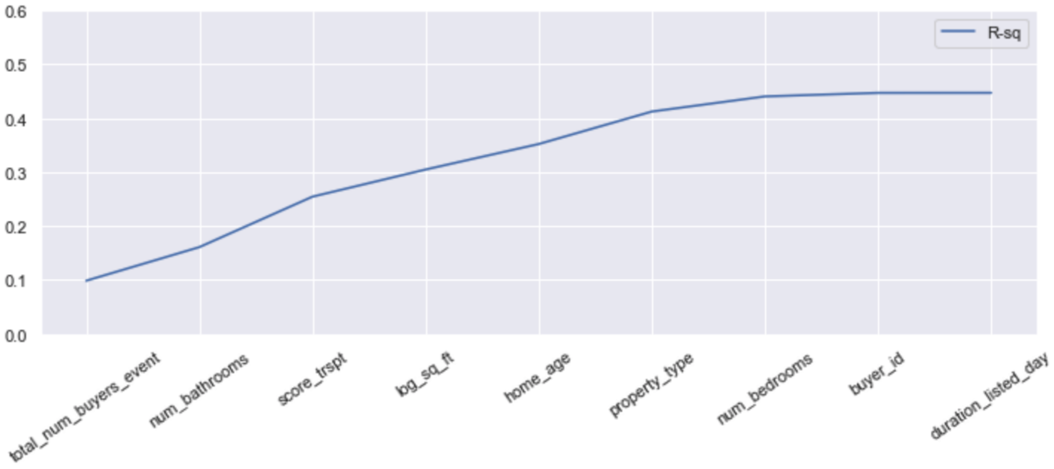
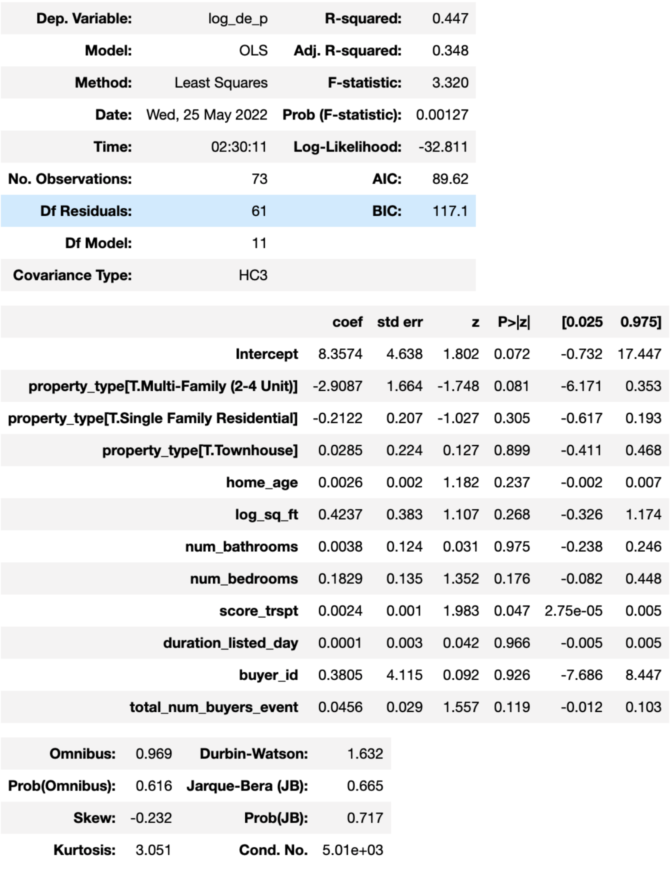


Figure 6: Variable-Rsq plot: mixed model



Dep. Variable:	log_de_p	**R-squared:**	0.447
Model:	OLS	**Adj. R-squared:**	0.348
Method:	Least Squares	**F-statistic:**	3.320
Date:	Wed, 25 May 2022	**Prob (F-statistic):**	0.00127
Time:	02:30:11	**Log-Likelihood:**	-32.811
No. Observations:	73	**AIC:**	89.62
Df Residuals:	61	**BIC:**	117.1
Df Model:	11		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	8.3574	4.638	1.802	0.072	-0.732	17.447
property_type[T.Multi-Family (2-4 Unit)]	-2.9087	1.664	-1.748	0.081	-6.171	0.353
property_type[T.Single Family Residential]	-0.2122	0.207	-1.027	0.305	-0.617	0.193
property_type[T.Townhouse]	0.0285	0.224	0.127	0.899	-0.411	0.468
home_age	0.0026	0.002	1.182	0.237	-0.002	0.007
log_sq_ft	0.4237	0.383	1.107	0.268	-0.326	1.174
num_bathrooms	0.0038	0.124	0.031	0.975	-0.238	0.246
num_bedrooms	0.1829	0.135	1.352	0.176	-0.082	0.448
score_trspt	0.0024	0.001	1.983	0.047	2.75e-05	0.005
duration_listed_day	0.0001	0.003	0.042	0.966	-0.005	0.005
buyer_id	0.3805	4.115	0.092	0.926	-7.686	8.447
total_num_buyers_event	0.0456	0.029	1.557	0.119	-0.012	0.103

Omnibus:	0.969	**Durbin-Watson:**	1.632
Prob(Omnibus):	0.616	**Jarque-Bera (JB):**	0.665
Skew:	-0.232	**Prob(JB):**	0.717
Kurtosis:	3.051	**Cond. No.**	5.01e+03

Figure 7: Regression Results: mixed model