

Question

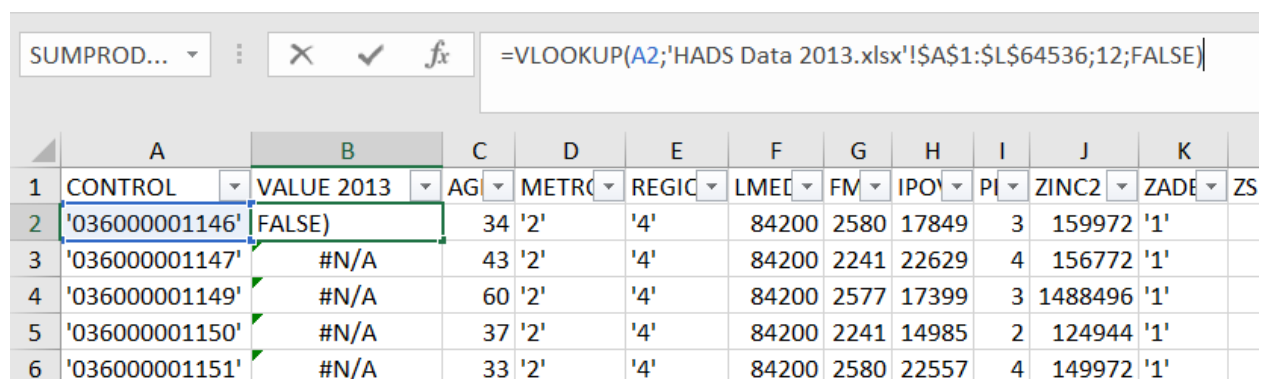
Can we predict the market value of housing in 2013 based on data from 2011?

Analysis Plan

1. Merge the VALUE variable from 2013 data into the 2011 data.
2. Data cleaning
 - a. Delete all rows for housing units that are not common across the two years.
 - b. Consider only 'Single Family Housing' (TYPE = 1 and STRUCTURETYPE = 1)
 - c. Delete all Housing units which have a market value of less than \$1000, that is, delete units with $VALUE < \$1000$
3. Conduct Holdout Analysis
 - a. Hold out some data and not include it in the regression. (1 000 random units)
 - b. Estimate a regression model using the 2013 VALUE variable and the X variables from the 2011 data.
 - c. Make predictions for VALUE 2013 using the holdout data
4. Compare these thousand predictions with the actual market value of those housing units.
 - a. Calculate Data Descriptors
 - b. Create Graphical Representation
5. Prepare a summary report

Course of Analysis

1. Data merging
 - Copied all the data from 2011 onto the 'Regression' sheet
 - Used VLOOKUP to insert VALUE figures from the 2013 data (as values)



The screenshot shows an Excel spreadsheet with a formula bar at the top containing the formula: `=VLOOKUP(A2;'HADS Data 2013.xlsx'!A1:L64536;12;FALSE)`. Below the formula bar is a table with 13 columns (A-M) and 6 rows (1-6). The table contains data for housing units, including a control variable, value for 2013, and various other attributes.

	A	B	C	D	E	F	G	H	I	J	K	L
1	CONTROL	VALUE 2013	AG	METR	REGIC	LMET	FM	IPO	PI	ZINC2	ZADE	ZS
2	'036000001146'	FALSE)	34	'2'	'4'	84200	2580	17849	3	159972	'1'	
3	'036000001147'	#N/A	43	'2'	'4'	84200	2241	22629	4	156772	'1'	
4	'036000001149'	#N/A	60	'2'	'4'	84200	2577	17399	3	1488496	'1'	
5	'036000001150'	#N/A	37	'2'	'4'	84200	2241	14985	2	124944	'1'	
6	'036000001151'	#N/A	33	'2'	'4'	84200	2580	22557	4	149972	'1'	

- Deleted all entries where 2013 values were not found
2. Data cleaning
 1. Deleted all entries that are not interpreted as 'Single Family Housing' (TYPE = 1 and STRUCTURETYPE = 1)
 2. Deleted all entries where the VALUE (2011 and 2013) was below \$1,000

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	CONTROL	VALUE 2013	AGE	METR	REGIO	LMEE	FM	IPOV	PER	ZINC2	ZADE	ZSMH	STATV	BEDRN	BUI	TY	VALUE 2011
6148	'100007130148'	-6	-9	'1'	'3'	62084	711	-9	-6	-6	'1'	-6	'3'	2	1980	1	-6
6149	'100007390148'	-6	54	'2'	'3'	63499	673	11536	1	9700	'1'	467	'1'	1	1985	1	-6
6151	'100008700141'	-6	34	'4'	'4'	54064	796	14849	2	56982	'1'	974	'1'	2	1980	1	-6
6153	'100009170148'	-6	59	'4'	'2'	61959	531	15026	2	7482	'1'	731	'1'	1	1985	1	-6
6155	'100013330103'	-6	-9	'1'	'2'	64028	686	-9	-6	-6	'1'	-6	'3'	2	2007	1	-6

3. Preparation for Holdout Analysis

- Identified only variables useful for prediction, discarding all variables that are either difficult to use in forecasting (like mortgage payments) or do not have a clear impact on market price (like monthly expenditures of residents). Deleted columns ZSMHC, STATUS, TYPE, STRUCTURETYPE, COST8, OWNRENT, COST08, COST09, COST12, COSTMED, ASSISTED, NUNITS
- Deleted those entries that have incorrect values for the remaining variables (such as negative age or monthly rent)

Remaining Values:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	CONTROL	VALUE 2013	AGE1	METRO3	REGION	LMED	FMR	IPOV	PER	ZINC2	ZADEQ	BEDRMS	BUILT	VALUE 2011	ROOMS	UTILITY	OTHERCOST	
2	'100006110249'	130000	40	'5'	'3'	55770	1003	11572	1	44982	'1'	4	1980	125000	8	220.5	41.6666667	
3	'100006520140'	200000	65	'5'	'3'	55770	895	13403	2	36781	'1'	3	1985	250000	5	230	40.3333333	
4	'100007540148'	260000	48	'1'	'3'	62084	935	17849	3	57446	'1'	3	1985	169000	6	236.333333	108.333333	

- Logarithmically transformed continuous values to reduce the impact of outliers and improve model accuracy. Transformed columns: VALUE 2013, LMED, Zinc2, VALUE 2011, UTILITY, OTHERCOST
- Created dummy variables for categorical variables, namely: Central City (Metro3 = '1'), RegNorth (REG = '1'), RegMidW (REG = '2'), RegSouth (REG = '3')

CONTROL	VALUE 2013	Ln(VALUE 2013)	AGE1	Central/City	RegNorth	RegMidW	RegSouth	Ln(LMED)	Ln(FMR)	PER	Ln(ZINC2)	IsAdequate	BEDRMS	BUILT	Ln(VALUE 2011)	ROOMS	Ln(UTILITY)	Ln(OTHERCOST)	
'100006110249'	130000	11.77528973	40	0	0	0	0	1	10.9289914	6.91075079	1	10.7140177	1	4	1980	11.73606902	8	5.39589769	3.729701449
'100006520140'	200000	12.20607265	65	0	0	0	0	1	10.9289914	6.79682372	2	10.5127367	1	3	1985	12.4292162	5	5.43807931	3.697178257
'100007540148'	260000	12.46843691	48	1	0	0	0	1	11.0362436	6.84054653	3	10.9586007	1	3	1985	12.03765399	6	5.46524324	4.685212894
'100008960141'	170000	12.04355372	58	0	0	0	0	0	10.8966467	7.10987946	2	11.7750128	1	3	1985	11.8493977	7	5.38296484	4.007333185
'100010190145'	230000	12.34583459	57	0	0	0	0	1	10.9289914	6.79682372	2	12.4632852	1	3	1985	12.32385568	5	5.23110862	3.825011628

4. Determining Holdout Data (1,000 Records)

- Identified random values using RAND(), creating a random number within the range of the number of records, which will be the record number after which the next 1000 values will be transferred.
- Cut these 1,000 records and moved them to the 'Predictive Test' sheet.

44	Holdout Data For Predictive Testing (1000 Housing units)																	
45																		
46	CONTROL	VALUE 2013	AGE1	CentralCity	RegNorth	RegMidW	RegSouth	Ln(LMED)	Ln(FMR)	PER	Ln(ZINC2)	IsAdequate	BEDRMS	BUILT	Ln(VALUE 2011)	ROOMS	Ln(UTILITY)	Ln(OTHERCOST)
47	'183932700148'	160000	59	0	0	1	0	11.0588897	6.94889722	3	11.9181572	1	5	1919	11.73606902	13	6.16506765	3.761200116
48	'18394148014C'	400000	83	1	0	0	0	11.2239092	7.60040233	2	10.7223859	1	3	1950	12.8346813	5	5.53338949	3.678408154
49	'18394189014C'	300000	48	1	0	0	0	11.2239092	7.60040233	2	10.6454249	1	3	1970	10.30895266	7	4.82028157	3.912023005
50	'184005390141'	200000	70	1	0	0	1	10.8570741	7.07665382	1	9.35876038	1	2	1970	11.91839057	6	5.28489435	5.675611598
51	'184029990143'	280000	78	1	0	0	1	11.131665	7.1172055	2	12.3542857	1	5	1970	14.22628355	9	6.69311712	5.339139361
52	'18403716014E'	2520000	51	0	0	0	0	11.0666384	7.58426482	6	11.0013662	1	3	1940	12.61153775	6	5.2505267	4.317488114

5. Holdout Analysis: Building the Regression Model

«Linear Regression» sheet

- Formulated an equation for the regression model based on previously identified variables.

$$\text{Ln}(\text{VALUE } 2013) = \beta_0 + \beta_1 \text{AGE1} + \beta_2 \text{CentralCity} + \beta_3 \text{RegNorth} + \beta_4 \text{RegMidW} + \beta_5 \text{RegSouth} + \beta_6 \text{Ln}(\text{LMED}) + \beta_7 \text{Ln}(\text{FMR}) + \beta_8 \text{PER} + \beta_9 \text{Ln}(\text{ZINC2}) + \beta_{10} \text{Adequate} + \beta_{11} \text{BEDRMS} + \beta_{12} \text{BUILT} + \beta_{13} \text{Ln}(\text{Value } 2011) + \beta_{14} \text{ROOMS} + \beta_{15} \text{Ln}(\text{UTILITY}) + \beta_{16} \text{Ln}(\text{OTHERCOST})$$

- Used built-in data analysis tools to build the regression model.

Regression Statistics									
Multiple R		0.774380042							
R Square		0.599664449							
Adjusted R Square		0.599340995							
Standard Error		0.491351158							
Observations		19820							
ANOVA									
		df	SS	MS	F	Significance F			
Regression		16	7161.419242	447.5887026	1853.9378	0			
Residual		19803	4780.958286	0.24142596					
Total		19819	11942.37753						
		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept		-4.130417548	0.419340978	-9.849782795	7.752E-23	-4.952361	-3.30847	-4.95236	-3.30847
AGE1		0.000795673	0.00026116	3.04668195	0.0023169	0.0002838	0.00131	0.00028	0.00131
CentralCity		-0.041550847	0.008815226	-4.713531517	2.451E-06	-0.058829	-0.02427	-0.05883	-0.02427
RegNorth		-0.076984344	0.013378086	-5.754511219	8.818E-09	-0.103207	-0.05076	-0.10321	-0.05076
RegMidW		-0.097239824	0.013988274	-6.951524219	3.726E-12	-0.124658	-0.06982	-0.12466	-0.06982
RegSouth		-0.109872807	0.011391554	-9.645111095	5.762E-22	-0.132201	-0.08754	-0.1322	-0.08754
Ln(LMED)		0.162000987	0.034848619	4.648706062	3.362E-06	0.0936948	0.23031	0.09369	0.23031
Ln(FMR)		0.362626159	0.024502707	14.79943272	2.71E-49	0.3145988	0.41065	0.3146	0.41065
PER		-0.00839105	0.00300017	-2.796857784	0.0051652	-0.014272	-0.00251	-0.01427	-0.00251
Ln(ZINC2)		0.051525405	0.004263759	12.08450101	1.673E-33	0.0431681	0.05988	0.04317	0.05988
IsAdequate		0.022443796	0.023548578	0.953084984	0.3405586	-0.023713	0.0686	-0.02371	0.0686
BEDRMS		-0.040844294	0.007515622	-5.434585717	5.557E-08	-0.055576	-0.02611	-0.05558	-0.02611
BUILT		0.001915657	0.000146381	13.0867569	5.689E-39	0.0016287	0.0022	0.00163	0.0022
Ln(VALUE 2011)		0.582281474	0.006588649	88.37645477	0	0.5693672	0.5952	0.56937	0.5952
ROOMS		0.048288268	0.003487659	13.84546888	2.159E-43	0.0414522	0.05512	0.04145	0.05512
Ln(UTILITY)		0.059288826	0.009074816	6.53336368	6.589E-11	0.0415014	0.07708	0.0415	0.07708
Ln(OTHERCOST)		0.009918024	0.00519141	1.910468169	0.0560874	-0.000258	0.02009	-0.00026	0.02009

R Square is approximately 0.6, meaning the **model can explain about 60% of the variance**.

Adjusted R Square is almost identical to R Square, suggesting that the model does not have 'superfluous' variables.

After analyzing the t-statistics for each value, it can be concluded that **each variable is statistically significant in the model**.

6. Holdout Analysis: Predictive test

Spreadsheet 'Predictive Test'

- Performed prediction of values for Ln(VALUE 2013) using a previously developed equation and the results obtained from the linear regression model.

Note: The logarithmically transformed value of VALUE 2013 was used in the model.

- In the column Predicted VALUE 2013, transformed the Ln(VALUE 2013) value back to standard form using the EXP() formula.
- In the 'Absolute Difference' column, calculated the difference between the actual and predicted values for each observation.

SUMPROD...		=D\$21+D\$22*G47 + D\$23*H47+D\$24*I47+D\$25*J47+D\$26*K47+D\$27*L47+D\$28*M47+D\$29*N47+D\$30*O47+D\$31*P47+D\$32*Q47+D\$33*R47+D\$34*S47														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
46	CONTROL	VALUE 2013	Predicted Ln(VALUE 2013)	Predicted VALUE 2013	Absolute Difference		AGE1	CentralCity	RegNorth	RegMidW	RegSouth	Ln(LMED)	Ln(FMR)	PER	Ln(ZINC2)	IsAdequ
47	'183932700148'	160000	=D\$21+D\$22*G47 +	176000.0894	16000.08944		59	0	0	1	0	11.0588897	6.94889722	3	11.9181572	
48	'183941480140'	400000	12.71898628	334030.0653	65969.9347		83	1	0	0	0	11.2239092	7.60040233	2	10.7223859	
49	'183941890140'	300000	11.31141464	81749.47283	218250.5272		48	1	0	0	0	11.2239092	7.60040233	2	10.6454249	
50	'184005390141'	200000	11.88652878	145296.0662	54703.93378		70	1	0	0	1	10.8570741	7.07665382	1	9.35876038	

7. Comparison of Predicted and Actual Values

After calculating the average from the 'Absolute Difference' column ("Predictive Test" sheet), determined that **the prediction deviates from the actual value by an average of ~70,000.**

On the 'Data Descriptors' sheet, calculated the central tendency and data dispersion for VALUE 2011, VALUE 2013, and Predicted VALUE 2013.

10				
11		VALUE 2011	VALUE 2013	Predicted VALUE 2013
12	Mean	247582.789	230169.83	212802.2772
13	Median	180000	150000	164377.2225
14	STDV	258729.621	254811.442	158361.7063
15	Min	3000	10000	8465.232599
16	Max	4414135	2520000	1222195.373
17	Range	4411135	2510000	1213730.14
18	25-percentile	113000	100000	109289.6302
19	75-percentile	300000	280000	265690.626
20				

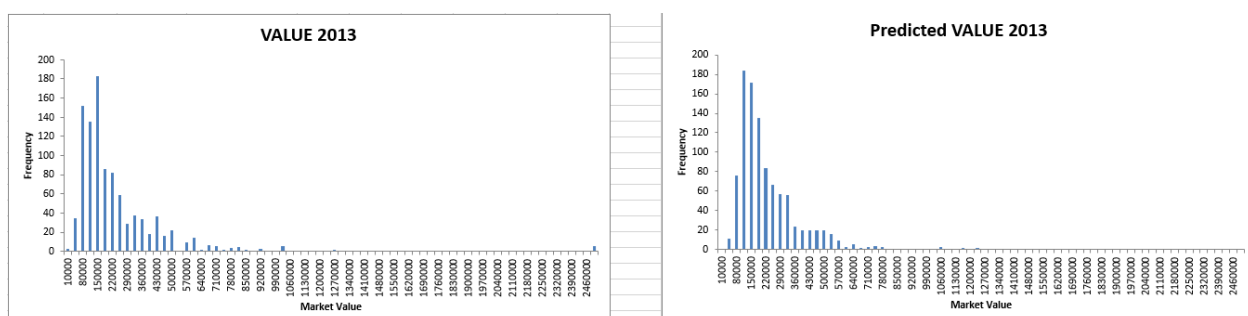
Note: For Value 2013 and Predicted VALUE 2013, values from the retained data (on the 'Predictive Test' sheet) were used.

It can be observed that the descriptive statistics for the actual and predicted values of VALUE 2013 are largely similar, except for the **maximum values and standard deviation, where the difference is approximately twice as large.**

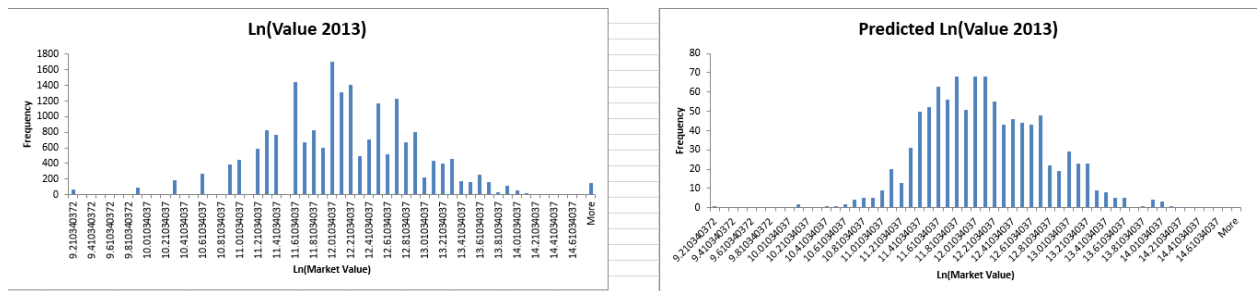
8. Comparison of Predicted and Actual Values: Graphical Representation

On the 'Charts and Graphs' sheet, created histograms for VALUE 2013 for both actual observations and predictions.

Both histograms have identical bins.

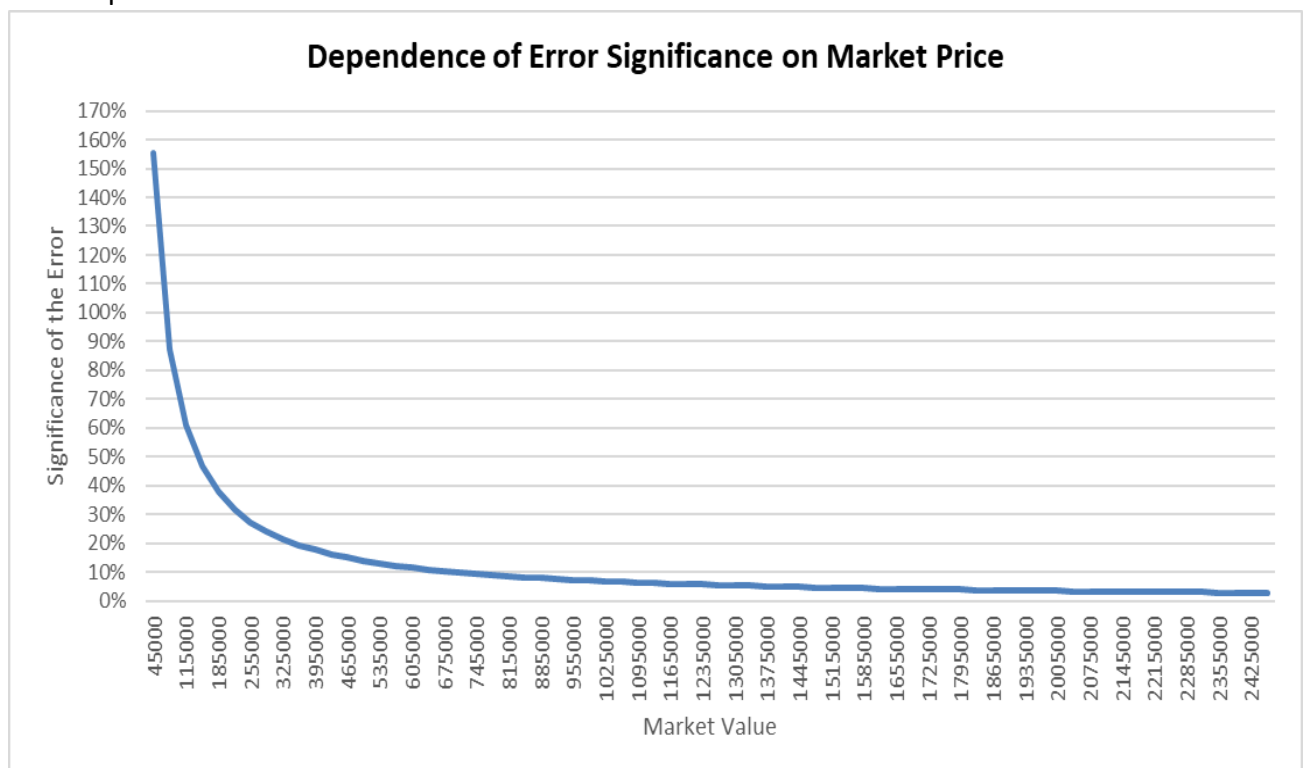


Since the previous graphs were heavily skewed by outliers, I also created a variation of these histograms, where the VALUE 2013 values were logarithmically transformed.



Based on these histograms, it can be said that the **predicted values are much more evenly distributed**, which is quite logical when using linear regression.

I also constructed a graph showing the dependence of the error significance (70,000) on the market price.



The graph demonstrates that as the price increases, the error becomes less significant.

9. Conclusion

The model has several significant issues:

- an average error of 70,000 in predicting the market value
- the adaptation of predictions to a uniform distribution

In the case of estimating a residential unit worth 700 000, the error can reach about 10%. However, **as the price increases, the significance of the error becomes less prominent.**

Therefore, using the model, **it is acceptable to make predictions of the market value of more expensive housing, but it performs poorly for more budget-friendly or average offerings.**

For a more comprehensive conclusion, you can refer to the 'Summary Report' sheet or the Summary.pdf document.