Assignment 4: Predicting Current Market Value

Assignment

Objective

- Predict 2013 Current Market Value from 2011 variables and assess accuracy of the model

Deliverables

- Build a regression model using 2011 variables to predict 2013 VALUE values
- Evaluate the effectiveness of the model
- Extend analysis to other years of data

Data Prep

- Merge variables from 2011 dataset with 2013 dataset using the CONTROL variable
- Remove all rows where the CONTROL variable is not in both datasets
- Remove all rows where VALUE < $1000
- Remove all rows corresponding to rental units (OWNRENT = 2)
- Remove all rows corresponding to non-single-family units (STRUCTURETYPE <> 1 and TYPE <> 1)
- Randomly select 1000 rows and set aside for the purpose of testing the model

Data Transformations

The **METRO3** variable was transformed into **CCITY**, using the following rule:

- If METRO3 = 1 then CCITY = 1 else CCITY = 0

The **REGION** variable was transformed into **REGNE**, **REGMW**, and **REGS**, using the following rules:

- If REGION = 1 then REGNE = 1 else REGNE = 0
- If REGION = 2 then REGMW = 1 else REGMW = 0
- If REGION = 2 then REGS = 1 else REGS = 0

LN Transformations

The following variables were transformed using the LN function to improve the fit of the regression model:

- LMED
- FMR
- ZINC2
- ZSMHC
- UTILITY
- COSTMED

Variables Used

- METRO3: metropolitan status
- REGION: census region
- LMED: area median income
- FMR: fair market monthly rent
- BEDRMS: number of bedrooms
- BUILT: year built
- ZINC2: annual household income
- ZSMHC: monthly housing costs, not including mortgage
- UTILITY: monthly utility costs
- COSTMED: monthly mortgage payment, assuming median interest

Descriptive Statistics

	VALUE	LMED	FMR	BEDRMS	BUILT	ZINC2	ZSMHC	UTILITY	COSTMED
Mean	258,582	68,208	1,281	3	1968	88,996	1,360	251	1,850
Standard Error	1,583	71	2	0	0	484	6	1	10
Median	190,000	64,810	1,204	3	1970	67,535	1,098	227	1,408
Mode	150,000	79,200	1,394	3	1950	99,974	532	192	795
Standard Dev	279,970	12,511	397	1	27	85,621	1,078	120	1,712
Sample Var	78,383,215,171	156,515,149	157,755	1	710	7,330,876,989	1,162,110	14,488	2,932,572
Kurtosis	33.5790	1	1	1	(1)	12	8	4	31
Skewness	4.8474	1	1	0	(0)	3	2	2	5
Range	2,510,000	76,800	3,030	7	94	1,061,920	10,658	1,243	17,049
Minimum	10,000	38,500	481	0	1919	1	9	6	106
Maximum	2,520,000	115,300	3,511	7	2013	1,061,921	10,667	1,249	17,155
Sum	8,092,310,000	2,134,583,829	40,075,246	101,988	61,577,607	2,785,135,933	42,551,732	7,870,340	57,891,349
Count	31,295	31,295	31,295	31,295	31,295	31,295	31,295	31,295	31,295

Regression Model

Our initial analysis used the variables from our previous assignment:

$$LN(VALUE\_2013) = \beta_0 + \beta_1(CCITY\_2011) + \beta_2(REGNE\_2011) + \beta_3(REGMW\_2011) + \beta_4(REGS\_2011) + \beta_5(LN(LMED\_2011)) + \beta_6(LN(FMR\_2011)) + \beta_7(BEDRMS\_2011) + \beta_8(BUILT\_2011) + \beta_9(LN(ZINC2\_2011)) + \beta_{10}(LN(ZSMHC\_2011)) + \beta_{11}(LN(UTILITY\_2011)) + \beta_{12}(LN(COSTMED\_2011))$$

Pair-Wise Correlation

	CCITY	REGNE	REGMW	REGS	LMED	FMR	BEDRMS	BUILT	LN(ZINC2)	LN(ZSMHC)	LN(UTILITY)	LN(COSTMED)
CCITY	1.0000											
REGNE	-0.0054	1.0000										
REGMW	0.0168	-0.2617	1.0000									
REGS	-0.0108	-0.3471	-0.4563	1.0000								
LMED	0.0680	0.0031	-0.0412	0.0022	1.0000							
FMR	0.0914	-0.0217	0.0024	-0.0267	0.6599	1.0000						
BEDRMS	-0.0240	0.0038	0.0054	-0.0333	0.1032	0.4688	1.0000					
BUILT	-0.1296	-0.0143	-0.0257	0.0232	-0.1216	0.0383	0.1536	1.0000				
LN(ZINC2)	-0.0022	0.0366	-0.0055	-0.0618	-0.0506	-0.0222	0.0102	-0.0358	1.0000			
LN(ZSMHC)	0.0286	-0.0268	-0.0267	0.0228	0.3177	0.4387	0.3281	0.2041	-0.0375	1.0000		
LN(UTILITY)	0.0471	-0.0206	0.0198	-0.0125	0.1748	0.2914	0.3343	-0.0020	0.0198	0.4554	1.0000	
LN(COSTMED)	-0.0151	-0.0156	-0.0226	-0.0023	0.3911	0.5596	0.3671	0.1731	-0.0345	0.5956	0.4061	1.0000

We do not have to consider correlation, as we have no values with a correlation > 90%

Regression Statistics

Multiple R	0.773441565
R Square	0.59821185
Adjusted R Square	0.597968433
Standard Error	0.492912916
Observations	19820

	df	**SS**	**MS**	**F**	**Significance F**
Regression	12	7165.01319	597.0844325	2457.51033	0
Residual	19807	4812.370962	0.242963142		
Total	19819	11977.38415			

The R square value for this regression model is ~ 0.60, which is significantly lower than our regression model using only data from the 2013 dataset.

Regression Statistics

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.984486548	0.421228618	-4.711186431	2.4795E-06	-2.810129921	-1.158843175	-2.810129921	-1.158843175
CCITY	-0.047931181	0.008729191	-5.490907438	4.04787E-08	-0.065041128	-0.030821235	-0.065041128	-0.030821235
REGNE	-0.06822267	0.013326215	-5.119433286	3.09323E-07	-0.094343169	-0.042102172	-0.094343169	-0.042102172
REGMW	-0.101768525	0.013932044	-7.304637185	2.88554E-13	-0.129076498	-0.074460552	-0.129076498	-0.074460552
REGS	-0.111449269	0.011272547	-9.886786899	5.37018E-23	-0.133544405	-0.089354133	-0.133544405	-0.089354133
LMED	0.227021281	0.034367215	6.605751559	4.05553E-11	0.159658662	0.2943839	0.159658662	0.2943839
FMR	0.266228694	0.02407804	11.05690879	2.45676E-28	0.219033719	0.31342367	0.219033719	0.31342367
BEDRMS	0.029476651	0.005689088	5.18126122	2.22551E-07	0.018325562	0.040627739	0.018325562	0.040627739
BUILT	0.001940186	0.000146625	13.23234396	8.42276E-40	0.00165279	0.002227582	0.00165279	0.002227582
LN(ZINC2)	0.048826418	0.004160379	11.73604956	1.06068E-31	0.040671726	0.05698111	0.040671726	0.05698111
LN(ZSMHC)	0.0244417	0.006727642	3.633025963	0.000280825	0.011254957	0.037628443	0.011254957	0.037628443
LN(UTILITY)	-0.099650298	0.00940044	-10.60059932	3.48006E-26	-0.118075947	-0.081224648	-0.118075947	-0.081224648
LN(COSTMED)	0.780958372	0.008104494	96.36115148	0	0.765072885	0.796843859	0.765072885	0.796843859

No variables have a p-value > 0.05, meaning they are all statistically significant in our model

Performance of the Model

Using our 1000 data rows set aside for testing:

Mean VALUE	$247,720.00
Mean Absolute Deviation	$71,463.04

Our model does not perform very well for predicting 2013 current market value.

Other Models

A variety of other models were built and evaluated. Some examples:

$$LN(VALUE\_2013) = \beta_0 + \beta_1(LN(LMED\_2011)) + \beta_2(LN(FMR\_2011)) + \beta_3(LN(ZINC2\_2011)) + \beta_4(ZSHMC\_2011) + \beta_5(BEDRMS\_2011) + \beta_6(ROOMS\_2011) + \beta_7(OTHERCOST\_2011)$$

R-square: 0.596068227 **Mean Absolute Deviation:** $71,463.04 **Mean VALUE:** $247,720.00

$$LN(VALUE\_2013) = \beta_0 + \beta_1(LN(VALUE\_2011)) + \beta_2(LN(VALUE\_2009)) + \beta_3(LN(VALUE\_2007)) + \beta_4(LN(VALUE\_2005))$$

R-square: 0.659113576 **Mean Absolute Deviation:** $66,072.19 **Mean VALUE:** $245,660.00

$$LN(VALUE\_2013) = \beta_0 + \beta_1(LN(VALUE\_2011)) + \beta_2(LN(VALUE\_2009)) + \beta_3(BUILT) + \beta_4(REGNE) + \beta_5(REGMW) + \beta_6(REGS)$$

R-square: 0.640461294 **Mean Absolute Deviation:** $69,124.82 **Mean VALUE:** $245,721.00

Summary

- Using the same variables from our regression analysis in assignment 3 was not an accurate model for predicting 2013 values using 2011 data
- A variety of other models were tested using various years of data, none of which were particularly good at predicting 2013 current market values
- Further exploration of available data might yield more promising results