Assignment 3: Current Market Value Regression Model

Assignment

Objective

- Conduct regression analysis on the VALUE (Current Market Value) variable from the 2013 dataset

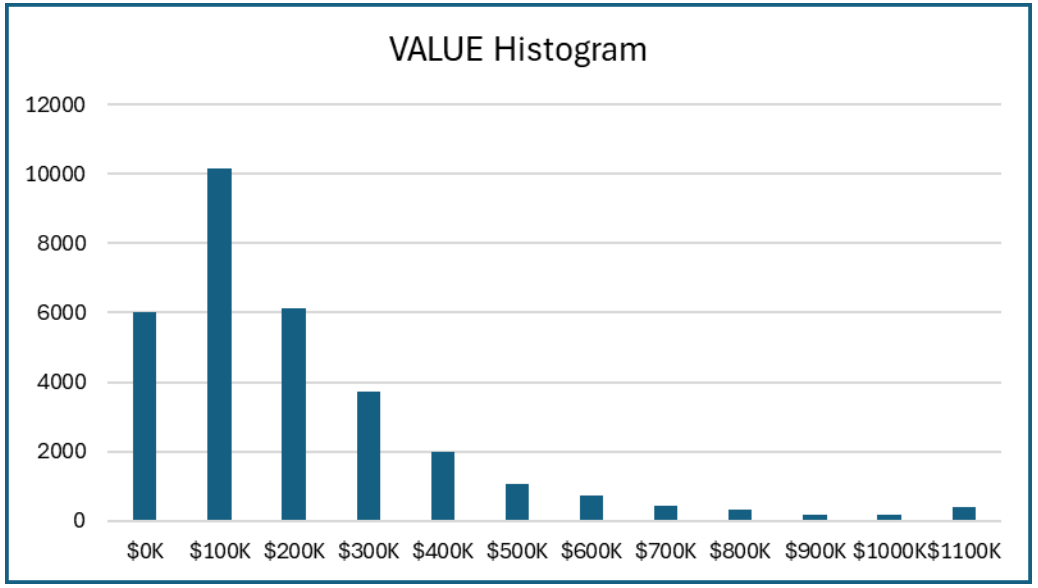
Deliverables

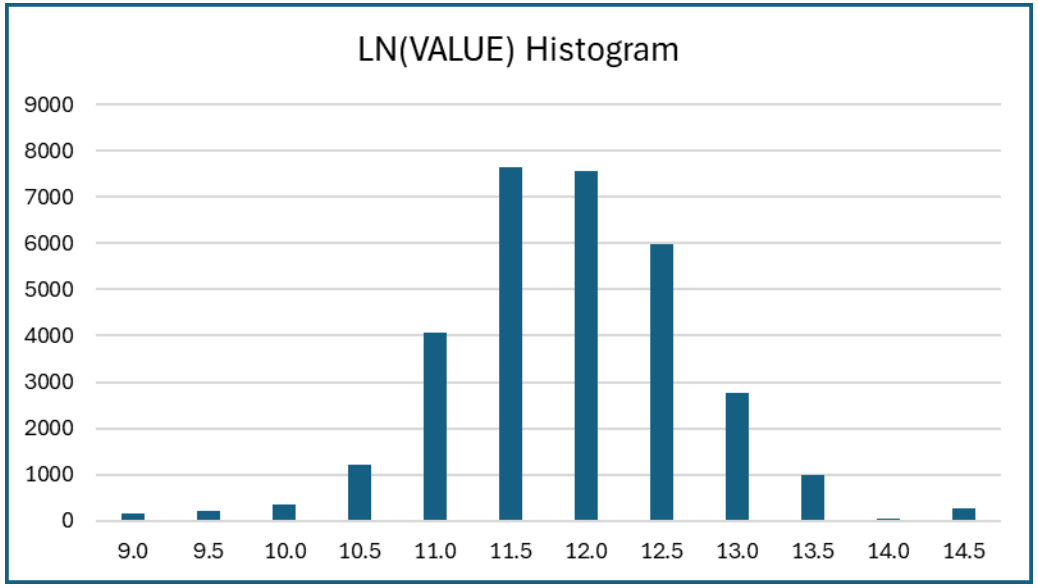
- List the variables used in the regression analysis
- Discuss any data transformations
- Explain the estimated regression model

Data Prep

- 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)

VALUE Analysis





Since the LN of VALUE is closer to a normal distribution, we will use LN(VALUE) in our regression model

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

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

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

The following regression model was used in the analysis of VALUE:

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

Running this model using 2013 data resulted in the following estimate:

$$LN(VALUE) = 3.4692 - 0.0156\ (CCITY) + 0.0117\ (REGNE) - 0.0159\ (REGMW) - 0.0176\ (REGS) + 0.0387\ (LN(LMED)) - 0.0499\ (LN(FMR)) + 0.0138\ (BEDRMS) + 0.0003\ (BUILT) + 0.0087\ (LN(ZINC2)) - 0.0248\ (LN(ZSMHC)) - 0.2252\ (LN(UTILITY)) + 1.2731\ (LN(COSTMED))$$

Pair-Wise Correlation

	CCITY	REGNE	REGMW	REGS	LMED	FMR	BEDRMS	BUILT	LN(ZINC2)	LN(ZSMHC)	LN(UTILITY)	LN(COSTMED)
CCITY	1.0000											
REGNE	-0.0568	1.0000										
REGMW	-0.0123	-0.3664	1.0000									
REGS	-0.0029	-0.3709	-0.4255	1.0000								
LMED	0.0077	0.5392	-0.1120	-0.4131	1.0000							
FMR	0.0613	0.3425	-0.3799	-0.1965	0.6631	1.0000						
BEDRMS	-0.0336	0.0177	-0.0295	-0.0166	0.1075	0.5236	1.0000					
BUILT	-0.1609	-0.2019	-0.0889	0.2065	-0.1476	0.0194	0.1420	1.0000				
LN(ZINC2)	-0.0497	0.0652	-0.0237	-0.0772	0.1617	0.2427	0.2574	0.1585	1.0000			
LN(ZSMHC)	-0.0187	0.1945	-0.0901	-0.1626	0.3577	0.4647	0.3419	0.1806	0.4601	1.0000		
LN(UTILITY)	0.0257	0.2072	-0.1282	-0.0475	0.1990	0.3279	0.3370	-0.0043	0.2462	0.4446	1.0000	
LN(COSTMED)	-0.0524	0.2183	-0.2371	-0.1362	0.3918	0.5702	0.3778	0.1758	0.4010	0.6384	0.4126	1.0000

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

Regression Statistics

Multiple R	0.98553898
R Square	0.971287081
Adjusted R Square	0.971276066
Standard Error	0.137384864
Observations	31295

	df	**SS**	**MS**	**F**	**Significance F**
Regression	12	19972.96546	1664.413788	88182.72714	0
Residual	31282	590.4352679	0.018874601		
Total	31294	20563.40073			

The R square value for this regression model is ~ 0.97, which is extremely high. This indicates that our regression model is a good fit to our sample data.

Regression Statistics

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3.46916519	0.095467791	36.33859287	3.131E-283	3.282044518	3.656285862	3.282044518	3.656285862
CCITY	-0.015612459	0.001943217	-8.03433433	9.73694E-16	-0.019421242	-0.011803675	-0.019421242	-0.011803675
REGNE	0.011720663	0.002922038	4.011125372	6.05704E-05	0.005993351	0.017447974	0.005993351	0.017447974
REGMW	-0.015937329	0.003091245	-5.155635182	2.5431E-07	-0.021996292	-0.009878367	-0.021996292	-0.009878367
REGS	-0.017563063	0.0026834	-6.545079037	6.03825E-11	-0.022822634	-0.012303493	-0.022822634	-0.012303493
LMED	0.038721199	0.007979942	4.852315707	1.22615E-06	0.023080195	0.054362204	0.023080195	0.054362204
FMR	-0.049895963	0.005804867	-8.59554005	8.66651E-18	-0.061273733	-0.038518194	-0.061273733	-0.038518194
BEDRMS	0.013802911	0.001323314	10.43056157	1.97962E-25	0.011209163	0.01639666	0.011209163	0.01639666
BUILT	0.000292257	3.22967E-05	9.049145569	1.52207E-19	0.000228955	0.00035556	0.000228955	0.00035556
LN(ZINC2)	0.008673783	0.000854041	10.15617168	3.39149E-24	0.00699983	0.010347737	0.00699983	0.010347737
LN(ZSMHC)	-0.024764679	0.001466448	-16.88753222	1.06529E-63	-0.027638975	-0.021890384	-0.027638975	-0.021890384
LN(UTILITY)	-0.22522188	0.001991587	-113.0866559	0	-0.229125469	-0.22131829	-0.229125469	-0.22131829
LN(COSTMED)	1.273146673	0.001741784	730.9442484	0	1.269732707	1.276560638	1.269732707	1.276560638

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

Interpretation of Model

- β_0: does not have a meaningful interpretation because it would not be practical for a home to have zero monthly expenses and zero fair market rent
- β_1: houses with a ***central city location*** tend to have a value that is ***1.56% less*** than houses located outside the central city, with all other variables remaining the same
- β_2: houses located in the ***Northeast*** census region tend to have a value that is ***1.17% more*** than houses located in the West census region, with all other variables remaining the same
- β_3: houses located in the ***Midwest*** census region tend to have a value that is ***1.59% less*** than houses located in the West census region, with all other variables remaining the same
- β_4: houses located in the ***South*** census region tend to have a value that is ***1.76% less*** than houses located in the West census region, with all other variables remaining the same
- β_5: for every 1% increase in ***area median income***, house values tend to ***increase by 0.04%***, with all other variables remaining the same

Interpretation of Model

- β_6: for every 1% increase in ***fair market rent***, house values tend to ***decrease by 0.05%***, with all other variables remaining the same
- β_7: for every additional ***bedroom***, house values tend to ***increase by 1.38%***, with all other variables remaining the same
- β_8: for every 1 year added to the ***year built***, house values tend to ***increase by 0.03%***, with all other variables remaining the same
- β_9: for every 1% increase in ***annual household income***, house values tend to ***increase by 0.01%***, with all other variables remaining the same
- β_{10}: for every 1% increase in ***monthly housing costs***, house values tend to ***decrease by 0.02%***, with all other variables remaining the same

Interpretation of Model

- β_{11}: for every 1% increase in ***utility costs***, house values tend to ***decrease by 0.23%***, with all other variables remaining the same
- β_{12}: for every 1% increase in ***monthly mortgage payments*** (assuming median interest), house values tend to ***increase by 1.27%***, all other variables remaining the same

Summary

The following regression model provides a statistically sound analysis of the current market value for single-family houses based on our 2013 dataset:

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