Angelo J. Vita, Peyton J. Hall

Dr. Iresha Premarathna

STAT 311-50

27 November 2024

An Investigation of Factors Affecting the Sales Price

of 1728 Single-family Homes in Saratoga NY.

Introduction:

This Case Study involves a partial investigation of the factors that affect the sales price of 1,728 Single-family Homes in Saratoga, NY. It is an analysis of the data and could demonstrate that regression analysis could be a powerful tool for appraising the markets in Saratoga. Realtors use experience and local knowledge to subjectively value a home based on its characteristics (size, amenities, location, etc.) and the price of similar homes nearby. Regression analysis provides an alternative approach that more objectively models local home prices using these same data. Thus, the data provided in the "Saratoga" table is used to develop a model for predicting the value of homes in Saratoga, NY.

The sales data were obtained and have been modified slightly for this assignment. The homes have many varying features:

Price = Sale price in dollars (quantitative)

1. lotSize = Lot size in acres (quantitative)
2. age = Age of home in years (quantitative)
3. landValue = value of the land in dollars (quantitative)
4. livingArea = square footage of the living area of the home (quantitative)
5. pctCollege = percent of neighborhood that graduated college (quantitative)
6. bedrooms = number of bedrooms (quantitative)
7. fireplaces = number of fireplaces (quantitative)
8. bathrooms = number of bathrooms (quantitative)
9. rooms = total number of rooms (quantitative)
10. heating = type of heating (qualitative)
11. fuel = source of heating (qualitative)
12. sewer = type of sewer system (qualitative)
13. waterfront = yes if waterfront property (qualitative)
14. newConstruction = yes if new construction (qualitative)
15. centralAir = yes if home has central air (qualitative)

Our objectives for this study are:

1. To acquire the prediction equation relating all the qualitative and quantitative variables to the sales price and determine whether the data supply is sufficient evidence to indicate these variables contribute information for the prediction of sales price.
2. Note: ID number will not be included as it only serves as identification for the House in question.

Data and Methodology:

This data obtained was from Saratoga NY Home Prices.

Since the multiple variables each will have a differing impact on the model we will first relate a first-order model (linear) of all the variables to sales price. Model 1 will assume that the impact of all the variables will be independent of the other variables. So our Model 1 should look like $E(y)= β0+β1x1+ β2x2+ β3x3+ β4x4+ β5x5 + β6x6+ β7x7+ β8x8+ β9x9+ β10x10+ β11x11+ β12x12 + β13x13+ β14x14+ β15x15+ β16x16+ β17x17+ β18x18$. where x10=1 if electric 0 if not, where x11=1 if hot air =1 if not, where x12=1 if electric 0 if not, where x13=1 if gas 0 if not, where x14=1 if no sewer 0 if not, where x15=1 if public/commercial 0 if not, where x16=1 if no waterfront, 0 if waterfront where x17=1 if not new construction 0 if new construction, x18=1 if not central air 0 if central air.

The second model is the effects of lot size, age, and living area might not be linear[1] due to the fact that those three factors have decreasing returns once a house reaches a certain size, and age. Model 2 has a second-order response for x1, x2 and x4 with squared variables assuming they will be the same regardless of the qualitative and other quantitative variables, and with no interaction term for any of the variables giving the Model 2: $E(y)= β0+β1x1+ β2x2+ β3x3 + β4x4+ β5x1^2 + β6x2^2+ β7x4^2+ β8x5 + β9x6+ β10x7+ β11x8+ β12x9+ β13x10+ β14x11+ β15x12+ β16x13+ β17x14+ β18x15+ β19x16+ β20x17+ β21x18$. where x10=1 if electric 0 if not, where x11=1 if hot air =1 if not, where x12=1 if electric 0 if not, where x13=1 if gas 0 if not, where x14=1 if no sewer 0 if not, where x15=1 if public/commercial 0 if not, where x16=1 if no waterfront, 0 if waterfront where x17=1 if not new construction 0 if new construction, x18=1 if not central air 0 if central air. Model 2 may have a problem since it assumes there is no difference in lot size, age, and living area variables for bedrooms and bathroom numbers. And since it is well known that the number of bathrooms and bedrooms can greatly increase house value[2,3].

Model 3 adds interaction terms between the lot size, age, and living area of the house and the number of bathrooms and bedrooms, so between x1,x2,x4, and the x6 and x8 terms. These changes will cause changes in y for increase in x1,2 and 4 price should increase as well meaning our Model 3 will be: $E(y)= β0+β1x1+ β2x2+ β3x3 + β4x4+ β5x1^2 + β6x2^2+ β7x4^2+ β8x5 + β9x6+ β10x7+ β11x8+ β12x9+ β13x10+ β14x11+ β15x12+ β16x13+ β17x14+ β18x15+ β19x16+ β20x17+ β21x18+ β22x1x2+ β23x1x4,+ β24x2x4+ β25x1x2x4+ β26x1x6+ β27x1x8+ β28x2x6+ β29x2x8+ β30x4x6+ β31x4x8+ β32x1x2x6+ β33x1x2x8+ β34x1x4x6+ β35x1x4x8+ β36x2x4x6+ β37x2x4x8+ β38x1x2x4x6+ β39x1x2x4x8$. where x10=1 if electric 0 if not, where x11=1 if hot air =1 if not, where x12=1 if electric 0 if not, where x13=1 if gas 0 if not, where x14=1 if no sewer 0 if not, where x15=1 if public/commercial 0 if not, where x16=1 if no waterfront, 0 if waterfront where x17=1 if not new construction 0 if new construction, x18=1 if not central air 0 if central air. Model 3 should be the most accurate, as the other variables, such as number of fire places, percent of College, and the qualitative should have lower importance on the Model. But there is 1 variable that should have a major impact of the price and that is land value, for naturally the more valuable the land the more valuable the property on that land meaning that we could have made a potential flaw in our reasoning by not including the importance of a variable[1,2].

Model 4 adds another second order term, this time x3, land value, and adds interactions between it and x1,x2,x4. Because the increased land value should be interacting with the lot size, age of the house and the living area. These changes to the model will cause changes in y for an increase in x1-x4 there should be a increase in price. Our Model 4 will be then: $E(y)= β0+β1x1+ β2x2+ β3x3 + β4x4+ β5x1^2 + β6x2^2+ β7x3^2+ β8x4^2+ β9x5 + β10x6+ β11x7+ β12x8+ β13x9+ β14x10+ β15x11+ β16x12+ β17x13+ β18x14+ β19x15+ β20x16+ β21x17+ β22x18+ β23x1x2+ β24x1x3,+ β25x1x4+ β26x2x3+ β27x2x4+ β28x3x4+ β29x1x2x3+ β30x1x2x4+ β31x1x3x4+ β32x2x3x4+ β33x1x2x3x4+ β34x1x6+ β35x1x8+ β36x2x6+ β37x2x8+ β38x3x6+ β39x3x8+ β40x4x6+ β41x4x8+ β42x1x2x6+ β43x1x2x8+ β44x1x3x6+ β45x1x3x8+ β46x1x4x6+ β47x1x4x8+ β48x2x3x6+ β49x2x3x8+ β50x2x4x6+ β51x2x4x8+ β52x3x4x6+ β53x3x4x8+ β54x1x2x3x6+ β55x1x2x3x8+ β56x1x2x4x6+ β57x1x2x4x8+ β58x1x3x4x6+ β59x1x3x4x8+ β60x2x3x4x6+ β61x2x3x4x8+ β62x1x2x3x4x6+ β63x1x2x3x4x8$. Where x10=1 if electric 0 if not, where x11=1 if hot air =1 if not, where x12=1 if electric 0 if not, where x13=1 if gas 0 if not, where x14=1 if no sewer 0 if not, where x15=1 if public/commercial 0 if not, where x16=1 if no waterfront, 0 if waterfront where x17=1 if not new construction 0 if new construction, x18=1 if not central air 0 if central air. This should take into account all major factors and their interactions with each other for the sales price as it allows for price changes to carry from variable to variable depending on their respective values.

Fitting Models 1-4 to the data and comparing the models using the nested model F test and conducting each test at $α=0.05$.

Results:

Our goal was to develop a model that can accurately and reliably predict the 1728 Saratoga NY Home Prices. Below is the generated data from the models created above. We will be using a partial F-test to see which model is more useful for predicting sales price compared to the predecessor. This is more useful as it avoids a type 2 error possibility. Moreover, t-tests should not unduly influence our analysis as seen in previous Case studies, a set of terms can contribute information for y prediction while having none of their t-values be statistically significant. This is due to the t-test focusing on a single-term contribution while all other terms are retained. Therefore, no terms may be statistically significant, even when the whole set contributes to the prediction of y. The regression analysis is seen below.

Model 1: H_0 β1-18=0 F value is 176.33. That is statistically significant at an $α=0.05$ level. Consequently, there is considerable evidence that the overall model contributes information to the prediction of y. At least one of the 18 factors contributes to the prediction of sales price.

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	93818.297	20401.45	4.60	<.0001*
lotSize	7560.5625	2246.892	3.36	0.0008*
age	-120.6641	58.67697	-2.06	0.0399*
landValue	0.9289156	0.047792	19.44	<.0001*
livingArea	70.001945	4.651268	15.05	<.0001*
pctCollege	-118.9697	152.0288	-0.78	0.4340
bedrooms	-7818.628	2586.486	-3.02	0.0025*
fireplaces	873.79145	3005.345	0.29	0.7713
bathrooms	23133.533	3395.037	6.81	<.0001*
rooms	2999.1419	971.6762	3.09	0.0021*
heating[electric]	10846.825	12868.37	0.84	0.3994
heating[hot air]	10777.627	4245.909	2.54	0.0112*
fuel[electric]	-5545.983	12900.82	-0.43	0.6673
fuel[gas]	4804.2642	5041.742	0.95	0.3408
sewer[none]	-4986.572	17139.35	-0.29	0.7711
sewer[public/commercial]	-2510.922	3687.653	-0.68	0.4960
waterfront[No]	-119938.3	15556.91	-7.71	<.0001*
newConstruction[No]	45856.008	7369.839	6.22	<.0001*
centralAir[No]	-9692.729	3504.027	-2.77	0.0057*

Summary of Fit

RSquare	0.653758
RSquare Adj	0.65005
Root Mean Square Error	58306.72
Mean of Response	211421.8
Observations (or Sum Wgts)	1700

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	18	1.0791e+13	5.995e+11	176.3323
Error	1681	5.7149e+12	3.3997e+9	Prob > F
C. Total	1699	1.6505e+13		<.0001*

If you examine the t-tests for the individual parameters you will see that most of the β's are statistically significant, except the β5, β7, β10, and β12-15. The failures of these variables to show their importance in mean sale price demonstrates the pitfall of relying on the results of t-tests in a regression analysis. We would expect these variables to have an impact on sales price because one could argue that the lack of sewers or differing fuel types would result in a different sales price or that people may want them for these features or may not. Why are the t-tests not statistically significant then? Both are correct but there is an interaction likely between the many variables but not the ones highlighted in our model. These effects would be cancelled because we do not have a complete interaction term for the model giving the impression it is unimportant. But this is information for the future. We now have to determine if Model 2 is better than Model 1.

Model 2 results:

Summary of Fit	
RSquare	0.656228
RSquare Adj	0.651926
Root Mean Square Error	58150.24
Mean of Response	211421.8
Observations (or Sum Wgts)	1700

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	21	1.0831e+13	5.158e+11	152.5308
Error	1678	5.6741e+12	3.3815e+9	Prob > F
C. Total	1699	1.6505e+13		<.0001*

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	120267.75	22224.41	5.41	<.0001*
lotSize	10682.55	4565.73	2.34	0.0194*
age	-450.7479	162.4829	-2.77	0.0056*
landValue	0.9167972	0.048277	18.99	<.0001*
livingArea	40.912149	11.47058	3.57	0.0004*
lotSize*lotSize	-477.3235	595.5704	-0.80	0.4230
age*age	2.2955515	1.064612	2.16	0.0312*
livingArea*livingArea	0.0064096	0.002453	2.61	0.0091*
pctCollege	-35.86349	153.6393	-0.23	0.8155
bedrooms	-6041.602	2651.741	-2.28	0.0228*
fireplaces	1273.8244	3020.846	0.42	0.6733
bathrooms	22286.722	3468.547	6.43	<.0001*
rooms	2876.031	970.6045	2.96	0.0031*
heating[electric]	5997.7116	12995.91	0.46	0.6445
heating[hot air]	9079.4286	4296.635	2.11	0.0347*
fuel[electric]	-2990.667	13150.99	-0.23	0.8201
fuel[gas]	5043.1158	5102.657	0.99	0.3231
sewer[none]	-6832.873	17163.09	-0.40	0.6906
sewer[public/commercial]	-2028.651	3801.378	-0.53	0.5936
waterfront[No]	-122502.5	15532.93	-7.89	<.0001*
newConstruction[No]	48842.793	7500.22	6.51	<.0001*
centralAir[No]	-10073.52	3502.692	-2.88	0.0041*

Are lot size, age, and living area related to sales price in a curvilinear manner, basically use a second order model? This time we only need to use the null hypothesis of β5-7=0. The F statistic for this test, based on this equation: Running a nested F test: $F = \frac{(\mathrm{SSE}_1 - \mathrm{SSE}_2)/\#\beta\text{'s in } H_0}{\mathrm{MSE}_2}$ Will give us the

Sum of Squares	40777031517
Numerator DF	3
F Ratio	4.0196779243
Prob > F	0.0073167061

F-statistic of 4.02. And with a p-value of less than 0.05 we reject H_0. Meaning there is evidence to indicate that Model 2 contributes more to prediction of y than Model 1. Meaning there is evidence of curvature in the response relating mean sale price E(y) to age, lot size, and living area. Also seen is how these interaction terms are all significant save for lot size. You will recall that the difference between Model 1 and Model 2 is that Model 2 allows for second order surfaces, one for age, one for lot size and one for living area. Model 1 employs just a linear model for all the second order terms. Now adding interactions terms to the Model 3 we can determine if Model 3 is better than Model 2.

Model 3 results:

Summary of Fit

RSquare	0.673218
RSquare Adj	0.665541
Root Mean Square Error	57001.66
Mean of Response	211421.8
Observations (or Sum Wgts)	1700

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	39	1.1112e+13	2.849e+11	87.6882
Error	1660	5.3937e+12	3.2492e+9	Prob > F
C. Total	1699	1.6505e+13		<.0001*

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	178470.91	33761.62	5.29	<.0001*
lotSize	-113679.9	29318.62	-3.88	0.0001*
age	-2646.264	605.8058	-4.37	<.0001*
landValue	0.9009278	0.048415	18.61	<.0001*
livingArea	-11.33549	19.4686	-0.58	0.5605
lotSize*lotSize	553.42276	655.8827	0.84	0.3989
age*age	4.3814095	1.162943	3.77	0.0002*
livingArea*livingArea	0.0079139	0.004982	1.59	0.1124
pctCollege	29.61443	152.4407	0.19	0.8460
bedrooms	5110.6155	11003.22	0.46	0.6424
fireplaces	1228.5029	2979.407	0.41	0.6801
bathrooms	-22969.53	15710.65	-1.46	0.1439
rooms	2586.9037	960.8116	2.69	0.0072*
heating[electric]	7953.6486	12784.91	0.62	0.5340
heating[hot air]	7462.9898	4275.361	1.75	0.0811
fuel[electric]	-3759.584	12950.44	-0.29	0.7716
fuel[gas]	5172.0766	5087.048	1.02	0.3094
sewer[none]	-10286.62	16897.03	-0.61	0.5428
sewer[public/commercial]	-1368.011	3769.947	-0.36	0.7167
waterfront[No]	-124605.6	15270.81	-8.16	<.0001*
newConstruction[No]	55283.495	7592.658	7.28	<.0001*
centralAir[No]	-10384.28	3463.956	-3.00	0.0028*
lotSize*age	4208.5977	1145.249	3.67	0.0002*
lotSize*livingArea	90.933782	17.95765	5.06	<.0001*
age*livingArea	1.4905465	0.339044	4.40	<.0001*
lotSize*age*livingArea	-2.864055	0.634302	-4.52	<.0001*
lotSize*bedrooms	5672.7254	11037.09	0.51	0.6073
lotSize*bathrooms	48193.071	16540.15	2.91	0.0036*
age*bedrooms	419.38651	197.9828	2.12	0.0343*
age*bathrooms	74.715008	319.8658	0.23	0.8153
livingArea*bedrooms	-4.121843	5.603191	-0.74	0.4621
livingArea*bathrooms	27.689699	8.273865	3.35	0.0008*
lotSize*age*bedrooms	-643.4151	207.6105	-3.10	0.0020*
lotSize*age*bathrooms	-554.7021	419.133	-1.32	0.1859
lotSize*livingArea*bedrooms	-4.978885	3.82264	-1.30	0.1929
lotSize*livingArea*bathrooms	-31.46746	7.788439	-4.04	<.0001*
age*livingArea*bedrooms	-0.326015	0.107988	-3.02	0.0026*
age*livingArea*bathrooms	-0.023268	0.148555	-0.16	0.8756
lotSize*age*livingArea*bedrooms	0.4334585	0.099271	4.37	<.0001*
lotSize*age*livingArea*bathrooms	0.3488861	0.191913	1.82	0.0693

Running a nested F-test for Model 3 $F = \frac{(\mathrm{SSE}_2 - \mathrm{SSE}_3)/\#\beta\text{'s in } H_0}{\mathrm{MSE}_3}$ based on the previous equation gives us the F-statistic of 4.79. And with a p-value of less than 0.05 we can reject the null hypothesis and say that Model 3 is better at the prediction of y than Model 2.

Sum of Squares	280420495457
Numerator DF	18
F Ratio	4.7947087612
Prob > F	1.360093e-10

And with a p-value below 0.05 (our significance level). We reject H_0 and conclude there is evidence to indicate that we need interactions terms to relate E(y) to x1,x2,and x4 to x6 and x8.

Model 4 results:

Summary of Fit

RSquare	0.697685
RSquare Adj	0.686044
Root Mean Square Error	55226.85
Mean of Response	211421.8
Observations (or Sum Wgts)	1700

Indicator Function Parameterization

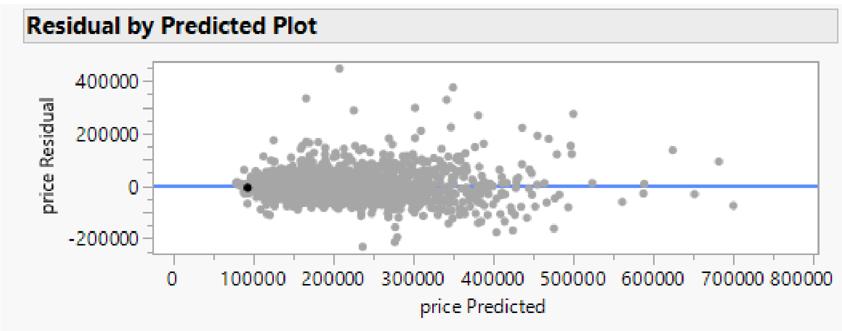
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	205186.42	48602.98	4.22	<.0001*
lotSize	-240351.6	56272.17	-4.27	<.0001*
age	-1158.945	940.4893	-1.23	0.2180
landValue	1.1099964	1.148215	0.97	0.3338
livingArea	-37.78713	31.39707	-1.20	0.2289
lotSize*lotSize	1653.3895	1039.051	1.59	0.1117
age*age	3.87302	1.166042	3.32	0.0009*
landValue*landValue	-7.98e-8	5.429e-7	-0.15	0.8832
livingArea*livingArea	0.0257346	0.006077	4.24	<.0001*
pctCollege	-26.46784	152.1171	-0.17	0.8619
bedrooms	41539.015	19202.24	2.16	0.0307*
fireplaces	3451.1111	2948.608	1.17	0.2420
bathrooms	-77937.49	22568.69	-3.45	0.0006*
rooms	2109.8569	941.7311	2.24	0.0252*
heating[electric]	9657.9453	12421.26	0.78	0.4370
heating[hot air]	9360.361	4207.693	2.22	0.0262*
fuel[electric]	-1298.066	12589.7	-0.10	0.9179
fuel[gas]	7110.9395	4997.866	1.42	0.1550
sewer[none]	-6917.46	16556.71	-0.42	0.6761
sewer[public/commercial]	1150.5467	3878.622	0.30	0.7668
waterfront[No]	-141535.4	15336.49	-9.23	<.0001*
newConstruction[No]	48889.482	7593.861	6.44	<.0001*
centralAir[No]	-10364.76	3379.581	-3.07	0.0022*
lotSize*age	5286.9414	1948.072	2.71	0.0067*
lotSize*landValue	2.100637	1.818712	1.16	0.2483
lotSize*livingArea	160.12496	33.40497	4.79	<.0001*
age*landValue	-0.026222	0.021067	-1.24	0.2134
age*livingArea	0.2940573	0.61823	0.48	0.6344
landValue*livingArea	3.5882e-5	0.000591	0.06	0.9516
lotSize*age*landValue	-0.053953	0.06444	-0.84	0.4026
lotSize*age*livingArea	-3.33377	1.094013	-3.05	0.0023*
lotSize*landValue*livingArea	-0.0014	0.000803	-1.74	0.0814
age*landValue*livingArea	0.0000114	1.083e-5	1.05	0.2924
lotSize*age*landValue*livingArea	3.8687e-5	3.214e-5	1.20	0.2289
lotSize*bedrooms	50014.591	28679.18	1.74	0.0814
lotSize*bathrooms	60753.718	30508.37	1.99	0.0466*
age*bedrooms	-285.9684	339.0685	-0.84	0.3991
age*bathrooms	800.05825	550.1092	1.45	0.1460
landValue*bedrooms	-1.258085	0.406658	-3.09	0.0020*
landValue*bathrooms	2.0635775	0.449017	4.60	<.0001*
livingArea*bedrooms	-20.31565	10.55703	-1.92	0.0545
livingArea*bathrooms	49.836947	12.93606	3.85	0.0001*
lotSize*age*bedrooms	-1232.411	640.231	-1.92	0.0544
lotSize*age*bathrooms	-300.1393	874.3243	-0.34	0.7314
lotSize*landValue*bedrooms	-0.169169	0.681764	-0.25	0.8041
lotSize*landValue*bathrooms	-0.813127	0.709226	-1.15	0.2518
lotSize*livingArea*bedrooms	-33.36558	13.90523	-2.40	0.0165*
lotSize*livingArea*bathrooms	-34.28251	16.34088	-2.10	0.0361*
age*landValue*bedrooms	0.0183611	0.006953	2.64	0.0084*
age*landValue*bathrooms	-0.011639	0.010666	-1.09	0.2753
age*livingArea*bedrooms	0.09031	0.199844	0.45	0.6514
age*livingArea*bathrooms	-0.420574	0.284737	-1.48	0.1399
landValue*livingArea*bedrooms	0.0003837	0.000175	2.19	0.0288*
landValue*livingArea*bathrooms	-0.000791	0.00021	-3.76	0.0002*
lotSize*age*landValue*bedrooms	0.0260338	0.023871	1.09	0.2756
lotSize*age*landValue*bathrooms	-0.027121	0.024546	-1.10	0.2694
lotSize*age*livingArea*bedrooms	0.7296906	0.258989	2.82	0.0049*
lotSize*age*livingArea*bathrooms	0.3879375	0.409837	0.95	0.3440
lotSize*landValue*livingArea*bedrooms	0.0003565	0.000259	1.38	0.1684
lotSize*landValue*livingArea*bathrooms	0.0002278	0.000275	0.83	0.4085
age*landValue*livingArea*bedrooms	-6.234e-6	2.813e-6	-2.22	0.0268*
age*landValue*livingArea*bathrooms	5.7164e-6	4.978e-6	1.15	0.2510
lotSize*age*landValue*livingArea*bedrooms	-0.000017	1.013e-5	-1.68	0.0938
lotSize*age*landValue*livingArea*bathrooms	9.0325e-6	1.045e-5	0.86	0.3875

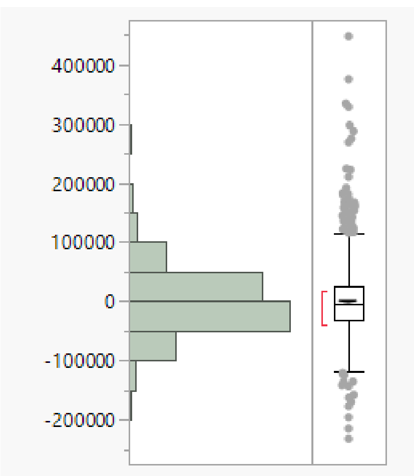
For our nested F-test with the following equation $\frac{(\mathrm{SSE}_3 - \mathrm{SSE}_4)/\#\beta'\text{s in } H_0}{\mathrm{MSE}_4}$, we get the F-statistic of 5.52 and with a p-value of less than 0.05 we can reject the null hypothesis and state that Model 4 is better at prediction of y compared to Model 3.

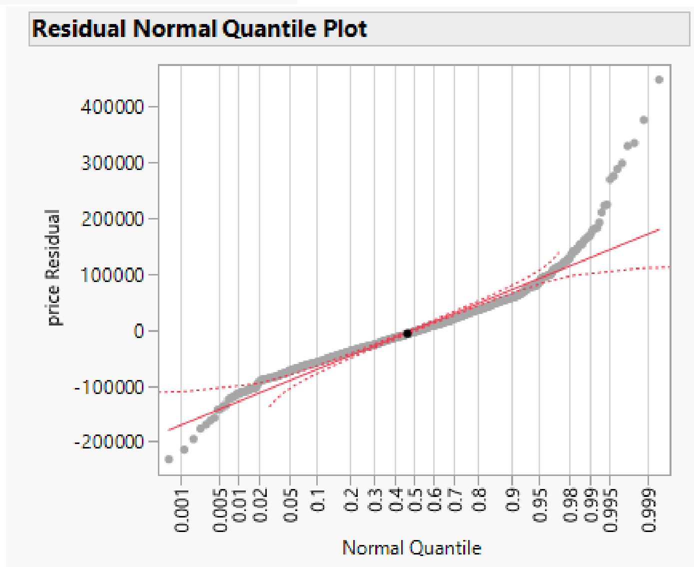
Sum of Squares	403845616162
Numerator DF	24
F Ratio	5.5170070773
Prob > F	3.32048e-16

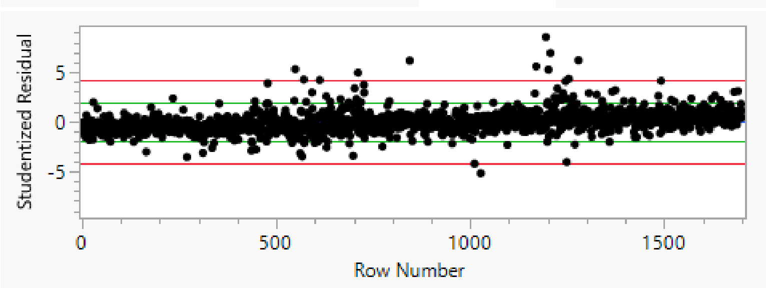
Having checked we can conclude that Model 4 is our best model best Rsquared value of almost 70%, implying our model explains 70% of our data. We can examine the prediction equation and see what it tells us about the relationship between E(y) and our 24 factors. But first we examine the residuals to determine whether the least squares assumption about random error is satisfied.

For our analysis of the residuals four assumptions about random error term must be borne in mind: 1. The mean is 0. 2. The variance (σ2) is constant for all settings of the independent variables. 3. The errors follow a normal distribution. 4. The errors are independent. If one or more of these assumptions are violated, any data derived from the Model 4 regression analysis may be wrong.







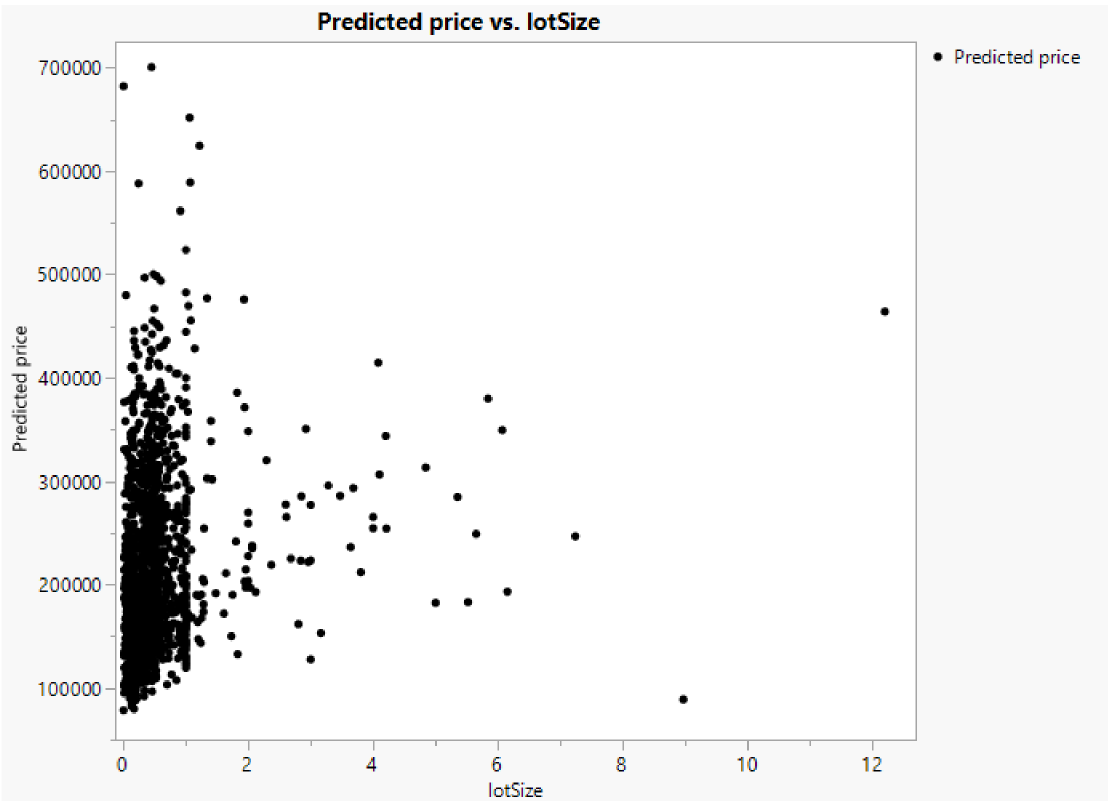


It is unlikely that the first assumption has been violated since the method of least squares guarantees the mean of residuals is 0, and assumption 4 since the sales price data is not a time series means that the errors would be independent of the data. Verifying 2 and 3 requires a complete examination of the residuals for Model 4. Seen above we can plot the residuals against the predicted values if they were not constant a cone shape would appear in the plot, but, other than about 10 outliers, we see the spread only increasing as the predicted values increase. Other than that the residuals appear randomly scattered around 0. Thus Assumption 2 (constant variance) is satisfied. To check assumption 3 we will have to generate a histogram, also seen above, to see if it follows a normal distribution and it clearly does. This is good because it means that even with outliers we possess a normal distribution.

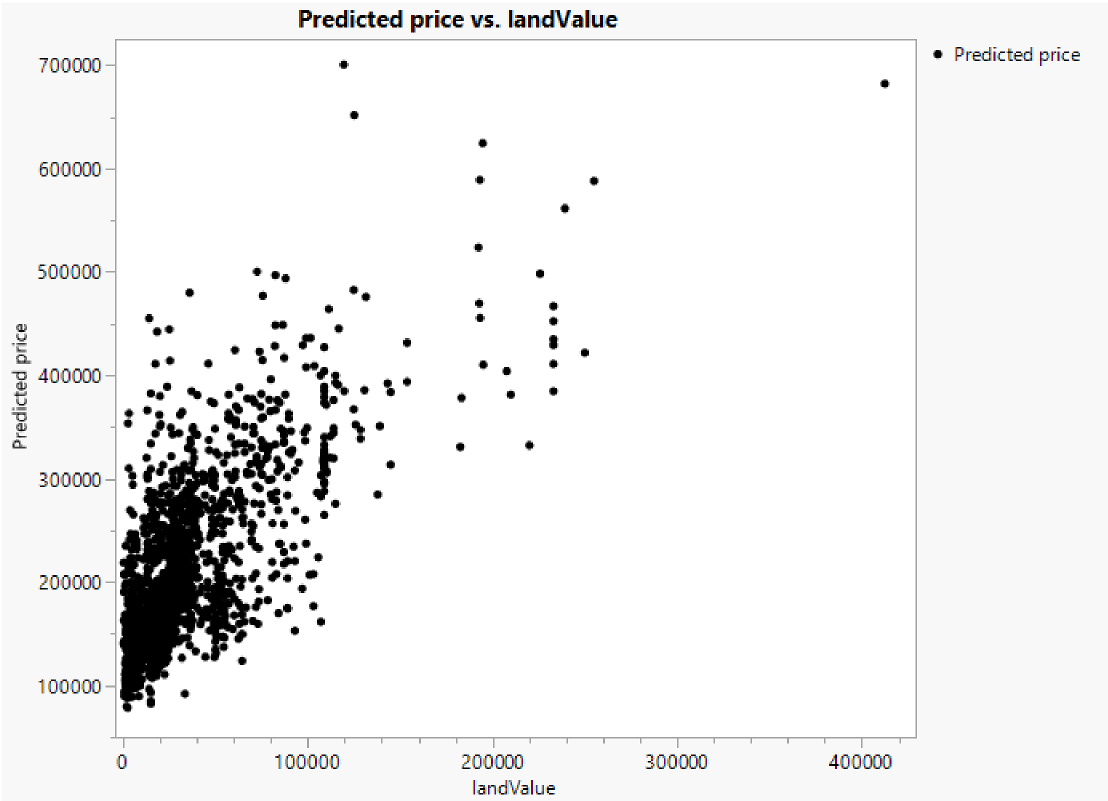
Prediction Expression

160035.83569
+ -240351.5594 • lotSize
+ -1158.945348 • age
+ 1.109996365 • landValue
+ -37.7871277 • livingArea
+ 1653.3894714 • lotSize • lotSize
+ 3.8730200451 • age • age
+ -7.979658e-8 • landValue • landValue
+ 0.0257345866 • livingArea • livingArea
+ -26.46784421 • pctCollege
+ 41539.015106 • bedrooms
+ 3451.1111439 • fireplaces
+ -77937.48524 • bathrooms
+ 2109.8568957 • rooms
+Match(heating) ("electric" ⇒ 3318.5098526; "hot air" ⇒ 3020.9255864; "hot water/steam" ⇒ -6339.435439; else ⇒ .)
+Match(fuel) ("electric" ⇒ -3235.690328; "gas" ⇒ 5173.3149289; "oil" ⇒ -1937.624601; else ⇒ .)
+Match(sewer) ("none" ⇒ -4995.155675; "public/commercial" ⇒ 3072.8512091; "septic" ⇒ 1922.3044659; else ⇒ .)
+Match(waterfront) ("No" ⇒ -70767.69841; "Yes" ⇒ 70767.698412; else ⇒ .)
+Match(newConstruction) ("No" ⇒ 24444.740978; "Yes" ⇒ -24444.74098; else ⇒ .)
+Match(centralAir) ("No" ⇒ -5182.379875; "Yes" ⇒ 5182.3798745; else ⇒ .)
+ 5286.9414336 • age • lotSize
+ 2.1006369828 • landValue • lotSize
+ 160.12495835 • livingArea • lotSize
+ -0.02622241 • landValue • age
+ 0.294057259 • livingArea • age
+ 0.0000358816 • livingArea • landValue
+ -0.053952693 • landValue • age • lotSize
+ -3.333770401 • livingArea • age • lotSize
+ -0.001399844 • livingArea • landValue • lotSize
+ 0.0000114005 • livingArea • landValue • age
+ 0.0000386868 • livingArea • landValue • age • lotSize
+ 50014.591281 • bedrooms • lotSize
+ 60753.71829 • bathrooms • lotSize
+ -285.9684466 • bedrooms • age
+ 800.0582496 • bathrooms • age
+ -1.258084933 • bedrooms • landValue
+ 2.0635774724 • bathrooms • landValue
+ -20.31564981 • bedrooms • livingArea
+ 49.836947334 • bathrooms • livingArea
+ -1232.411315 • bedrooms • age • lotSize
+ -300.1393193 • bathrooms • age • lotSize
+ -0.169169441 • bedrooms • landValue • lotSize
+ -0.813127455 • bathrooms • landValue • lotSize
+ -33.36557816 • bedrooms • livingArea • lotSize
+ -34.28250701 • bathrooms • livingArea • lotSize
+ 0.018361104 • bedrooms • landValue • age
+ -0.011639479 • bathrooms • landValue • age
+ 0.0903100438 • bedrooms • livingArea • age
+ -0.420574257 • bathrooms • livingArea • age
+ 0.0003836779 • bedrooms • livingArea • landValue
+ -0.000790591 • bathrooms • livingArea • landValue
+ 0.0260337917 • bedrooms • landValue • age • lotSize
+ -0.027120989 • bathrooms • landValue • age • lotSize
+ 0.7296905637 • bedrooms • livingArea • age • lotSize
+ 0.3879374623 • bathrooms • livingArea • age • lotSize
+ 0.0003565276 • bedrooms • livingArea • landValue • lotSize
+ 0.0002277575 • bathrooms • livingArea • landValue • lotSize
+ -6.233579e-6 • bedrooms • livingArea • landValue • age
+ 5.7164043e-6 • bathrooms • livingArea • landValue • age
+ -1.698653e-5 • bedrooms • livingArea • landValue • age • lotSize
+ 9.0325091e-6 • bathrooms • livingArea • landValue • age • lotSize

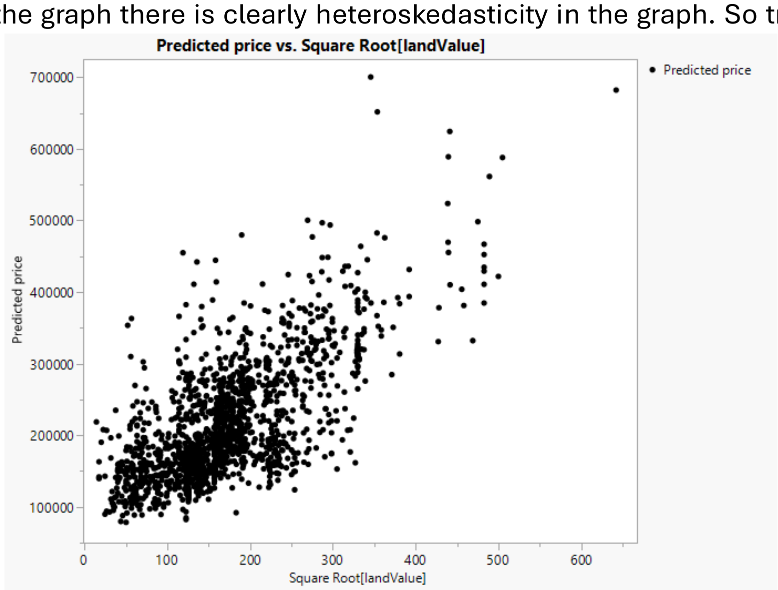
For our prediction equation of Model 4 thus the effect is seen in the that whenever we assign a value of 1 to one of the dummy variables for floor height it will increase the estimated mean sale price by a fixed amount, meaning that it will push the E(y) value up or down depending on the β parameter associated with the dummy variable. For example, with our prediction equation we can see that when there is a no waterfront (β20=1), then the value of E(y) will decrease by $70767.70 regardless of age, lot size, living area, or land value. The effect of age, lot size, land value, and living area (x1-4) can be determined by plotting y-hat as a function of each of the variables for given values of the other. For example suppose we wish to determine the relationship between y-hat and x1. With x2-4=0. The prediction curve for all the houses relating y-hat to lot size x1, can be graphed showing us the overall effect this variable has on the predicted values of the houses.



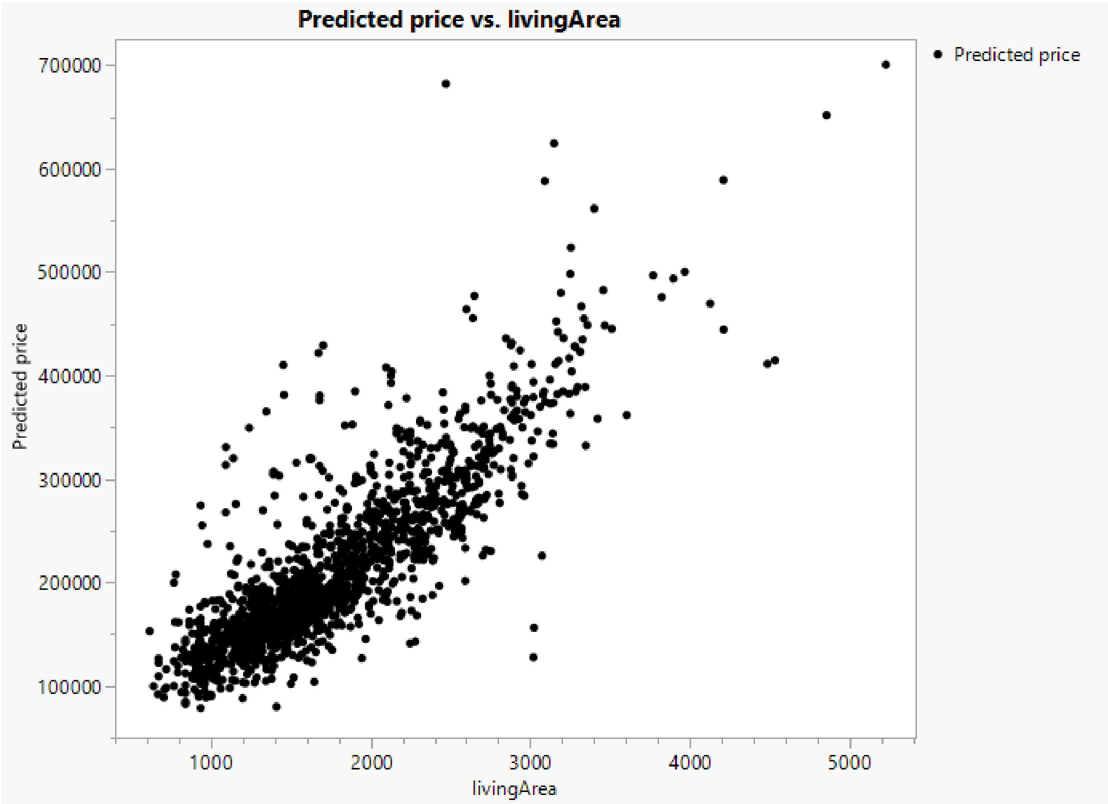
As we can the values are mostly concentrated on the left side but there is a slight upward trend the farther right one go in the graph. Generally the larger a property the more value it will have. Which is pretty common trend in the real estate market[4]. What is noteworty is that the increase in price is not a major and easily visible trend meaning that the value of the land must have some impact on this. Which we can see when we set the other variables as constants as we did with x1, and have x3=1 and x1,x2,x4=0.



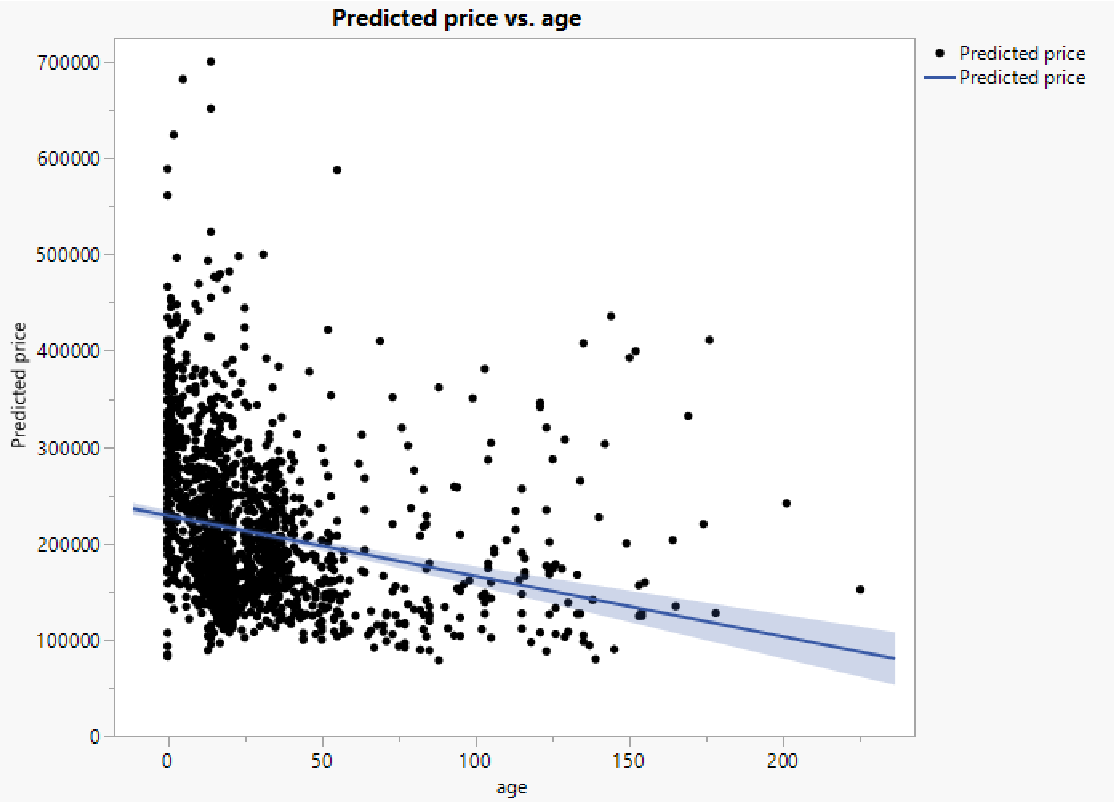
As we can see there is a clear upward trend to the predicted price related to land value. However as seen in the graph there is clearly heteroskedasticity in the graph. So transforming it with a sqrt(x2) gives us



And seen is the trend more clearly visible, of the increasing land value equals a higher predicted price. Which is consistent with realtor experience and data[4]. But living area has always contributed positively to housing prices[1,2,3,4]. So for our next x varible we will see the effect living area has on predicted price. So x4=1 and x1-3=0. Giving us the graph



Further transformation of this model, log, sqrt, squared, lead to further heteroskedasticity. But from this graph there is a clear upward trend of predicted price vs. living area. Which is consistent with realtor experience and data[1.2.3.4]. But what about age? Age can have an impact on the housing price[1]. But when we graph it we get



Unfortunately further transformations skew the data more. But the clear slight downward trend is visible, but it is very slight and small which our reference confirms as age can have charm but also further maintenance, so it depends on the house in question meaning age has a very minimal impact on mean sales price.

<u>Conclusion:</u>

With the data gathered from our generated models we can determine the impact of the many variables on mean sales price and how they affected it. Clearly as predicted the variables of living area and land value have a significant impact on the price of the 1728 Saratoga home with a strong positive collinearity. Furthermore, we can say that variable such as age, lot size have a impact but a significantly lesser one compared to living are and land value which clearly have a massive impact on the sales price. These are variables that can be used to help further real estate development proving that Model 4 does the best when it comes to predicting mean sales price E(y) because the many initial predictors of sales price are correctly modelled to the significance of their impact which leads to a more accurate model. This is because rather than the one quantitative variable which limits the model and crams the entire effect into one variable, the many transformation accurately depict the sales trend, highlighting the complex relationship between the variables.

References:

1. https://www.compmort.com/what-determines-the-value-of-a-home/#:~:text=Property%20size%20and%20usable%20space,-Lot%20size%20refers&text=Consider%20the%20following%3A,are%20more%20appealing%20to%20buyers.
2. https://money.com/how-to-price-a-home/#:~:text=Quantity%20of%20bathrooms%20is%20more%20important%20than%20quality&text=If%20an%20average%2Dsized%20home,price%20up%20thousands%20of%20dollars.
3. https://www.homelight.com/blog/how-much-value-does-a-bedroom-add/
4. https://www.investopedia.com/articles/mortgages-real-estate/08/housing-appreciation.asp

GitHub Link:

1. https://github.com/Peytonjohnhall/STAT-311