Does Size Matter: An Estimation of the Relationship Between Size & House Prices in Nassau County

Kanwarpartap Singh Brar

This paper analyzes data regarding how the price of a home changes in response to variables that affect a property's market value.

Table of Contents

1. Introduction

The real estate housing market is a very complex and dynamic market that tends to fluctuate year-round. A multitude of different factors and variables can cause these fluctuations. Some of these factors include local and global economic conditions, government policies, interest rates, and housing supply and demand during a given period. However, even though the housing market changes all the time, homes need to be appraised accordingly. A home's appraisal value is basically the value that assesses how much the property is worth in the given market. One of the major factors that plays a role in property value is the house size and how much liveable square footage is available in the property.

Whether you are a home buyer or a home seller, you want to make sure that your home is appraised fairly. One of the ways you can do this is by researching and looking at how other homes in your area are appraised based on specific factors and characteristics. One of the major factors that can affect the appraisal value and price of a home is the size of the home. The size of a home is measured in square footage (sq ft). There are two measurements of square footage that are commonly accounted for when a home is appraised. The first measure is the lot size of the home. The lot size of the home refers to the total area of land on which the home is situated. This measurement includes the property's outdoor space but does not account for the total livable square footage. The livable square footage is the interior square footage of a home, which accounts for all of the interior living space. This typically includes bedrooms, living rooms, dining rooms, bathrooms, and any other livable room in a home. However, getting data on the livable square footage is typically difficult because the way to measure the livable square footage varies based on local building codes and standards.

In this research paper, I measure how the total interior liveable area affects the appraisal price of homes in Nassau County. As mentioned before, how big a house is affects the overall value of the home. Therefore, I use this understanding to analyze and assess the significance of house size, measured in liveable square footage, on property appraisals in Nassau County. While my population is Nassau County, the data in this report has been drawn from a sample of 3 different towns in Nassau County. West Hempstead, Garden City, and New Hyde Park respectively, allowing for a total of 99 observations (n = 99). However, as a disclaimer since not all of the available data/observations offer information on the amount of liveable square footage, I will also be including another independent variable measuring how the lot size of a home tends to affect its price in Nassau County. Doing this will help predict and analyze home appraisal prices in the current market based on how much liveable square footage and how large the lot size of a home is. For further details on the methodology used in this report, head to the methodology section *(Section 4).*

In the past, similar research has been done that measures how different factors affect the value of a home. The peer-reviewed literature concludes that size would have a statistically significant effect on home value. Some argue that view has the greatest impact on the value of a home, in that a view adds the greatest premium to the appraisal value of a home. On the other hand, there are also similar research studies that have been done that show a statistically significant negative effect on home prices from relatively larger home sizes. All of these studies have similarities with the research and results being presented in this report, however, they do vary based on methodology. As mentioned before, real estate housing pricing can be determined by a multitude of different variables which can have an effect on how pricing in a market can be analyzed. Overall, based on the previous literature, home size does affect the valuation of a home, however, it does depend on location. That is precisely what the purpose of this research and report is, to see how the size affects the pricing in a specific location (Nassau County).

Further information on previous literature reviews can be accessed in the literature review section of this report (*Section 2)*, as well as in the citation section (*Section 7*) where all citations from previous literature reviews mentioned in this report can be found in MLA format.

Sections 5 and 4 of this report show us that the size of a home does have a positive relationship with its price. We can see that the square footage of the interior of a home is statistically significant in relation to the price of the home, whereas in comparison, even though the lot size is positively correlated with the price of a home, it is not statistically significant in the robust multivariate log-linear benchmark model which is presented in ***Table 5.1***.

2. Literature Review

There are plenty of different studies that measure how a home's price is affected. A study presented by RePEc (Research Papers in Economics) measured how views can affect the transaction price of a home. Unlike the study conducted in this research paper, the study by RepEc presented its results based on the transaction price of a home. The transaction price of the home is the agreed-upon value of the home during its official sale, whereas the appraisal price is the estimated value of the property. More importantly, this study's results indicated that "square footage and lot size significantly affect a home's value… and having a very desirable view adds an 89.9% premium to the value of a home" (Bond et al., 2002).

Similar to the study mentioned above another research article measuring the impact of relative size on home values concluded that the size of a house is statistically significant. In this study, instead of using the square footage of a home as a variable to measure the size of a home, the researchers used other house-related characteristics to measure the size of a home. For example, variables such as number of bedrooms (*BEDS*), bathrooms (*TBATHS*), and number of fireplaces (*FIRENUM*). The results of this study concluded that the larger the house the higher the price it will be appraised at, however, it also shows that the larger the house, the more likely it is to sell at a discount when comparing it to the appraised value. Whereas a smaller home will sell at a higher price when compared to its appraised value. "These results support that relative to the average range of sale prices by neighborhood, larger houses sell at discounts while relatively smaller houses sell at premiums" (Asabere & Huffman, 2013). The researchers assume that these results depend on the housing market at the time and what the buyer's motivations and needs are. This leads back to the idea mentioned in the previous section, that if you are a home buyer or seller, it is important to understand and analyze how the home of interest is appraised, so you can make your best judgment on the eventual transaction price accordingly.

In contrast to the two studies mentioned above, this paper specifically delves into a specific area of interest, which is Nassau County, and it uses liveable square footage and lot size as the two main independent variables to measure how they affect the dependent variable which is the house appraisal price.

3. Behavioral Hypothesis

This is a multivariate analysis that calculates how the size of a home affects the appraised price of the home. The hypothesis in this analysis is that as the square footage of the home increases (including both liveable sq ft and lot size), the price of the home will also increase. This would conclude that the size of a home is statistically significant when dealing with its relationship to the price of the home. To test this I will do a F-test with the null hypothesis as $\beta_1=\beta_2=0$.

4. Methodology

The data collected is sourced from Zillow.com, which is an online real estate marketplace known for offering property value information, as well as recognized as a valuable source for real estate data and insights.

The population being analyzed is Nassau County, and the samples are drawn from three different towns in Nassau County. The respective towns in the sample being measured include West Hempstead, Garden City, and New Hyde Park. All three of which are towns that belong within Nassau County on Long Island, New York. The reason behind only measuring three towns is that if more samples were to be included, the data set would have been too large and difficult to measure. Just including these three towns gives 99 total observations, letting n = 99 (n = # of observations). The number of observations out of the 99 that recorded the *LSQFT* is 45 (n = 45). Adding more towns to the dataset is something that is discussed in the conclusion section (*Section 6*). The data needed for each variable was manually retrieved from each observation by doing research on a specific property and collecting the necessary information needed for the variables that are being used in the regressions. The data set includes one dependent variable which measures the appraised price, and the rest of the data retrieved were independent and control variables measuring the liveable square footage (*LSQFT*), lot size (*LOT*), average quality of school district (*SCHOOL*), the number of beds (*BDS*), the number of bathrooms (*BA*), a dummy variable which measured whether or not the property was a new construction (*NEW_CONST*), the walk score of the neighborhood (*WLK_SCR*), the transit score of the neighborhood (*TRNSIT_SCR*), and the bike score of the neighborhood (*BIKE_SCR*). The average quality of a school district was achieved by gathering the rating from each school in the observed district and calculating its average using the formula **m = sum of ratings/# of ratings**.

The reasoning behind including *BDS* & *BA* in the model lies in recognizing them as integral components influencing the size of a house. This means that bedrooms and bathrooms are components of a house that take up liveable square footage, the number of beds and baths are to impact the price of the home. The rationale for including *NEW_CONST, SCHOOL, WLK_SCR, TRNSIT_SCR,* and *BIKE_SCR* is that these variables are key features that can contribute to the property's market value, and their inclusion can enhance the analysis of the model and strengthen it. For example, *WLK_SCR, TRNSIT_SCR,* and *BIKE_SCR* are all examples of a neighborhood rating, and the neighborhood is a quality that can impact the price of a home.

Below is a definition of variables table (***Figure 4.1***) which provides the variable, definition of the variable, and the units of measurement of the variable in further detail.

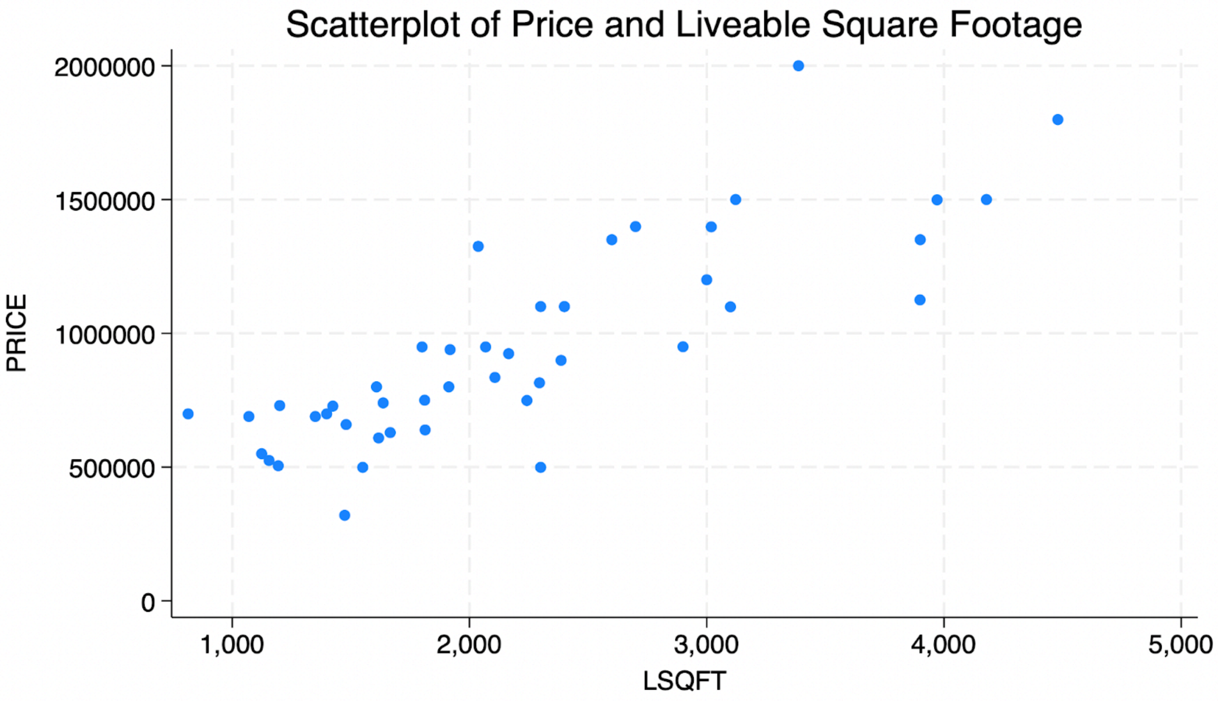
(***Figure 4.1***)

Variable	***Variable Symbol***	***Unit of Measurement***
Dependent Variable:		
Estimated appraisal price of a home	*PRICE*	U.S Dollars in 2023/%
Independent Variables:		
Liveable square footage of a home. (The habitable space within a residential property)	*LSQFT*	Square Feet (Sq ft)
Lot Size/the total area of the property	*LOT*	Square Feet (Sq ft)

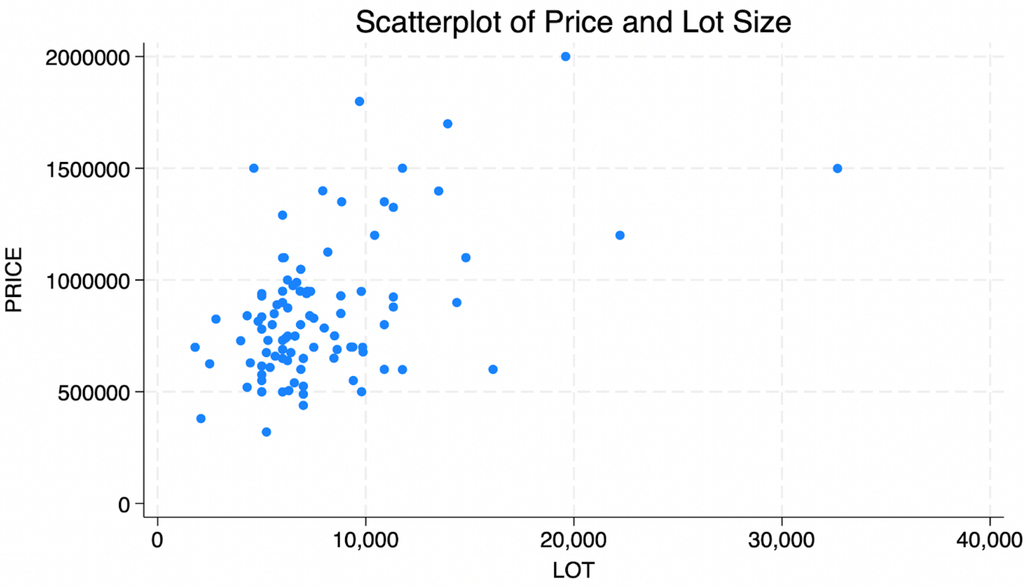
Control Variables:		
Average Quality of School District	*SCHOOL*	Number Rating (1-10) 1 is the worst 10 is the best
Number of bedrooms in the house	*BDS*	Number Rating
Number of bathrooms in the house	*BA*	Number Rating
Whether or not the house is a new construction	*NEW_CONST*	Zero (Not a new construction) One (New Construction)
Other:		
Walk Score. A numerical measure that assesses the walkability of an area	*WLK_SCR*	Rating from 0-100 0-49: Car dependent 50-69: Somewhat walkable 70-89: Very walkable 90-100: Walkers Paradise
Transit Score. A numerical measure that assesses the availability and quality of public transportation options	*TRNSIT_SCR*	Rating from 0-100 0-24: Minimal transit 25-49: Some transit 50-69: Good transit 70-89: Excellent transit 90-100: Riders paradise
Bike Score. A numerical measure that assesses the bike-friendliness of a specific area	*BIKE_SCR*	Rating from 0-100 0-49: Somewhat bikeable 50-69: Bikeable 70-89: Very bikeable 90-100: Bikers paradise

Below are two different scatterplots that present the dependent variable and the independent variables being studied. The first scatterplot in the figure (***Figure 4.2***) shows the relationship between *PRICE* and *LSQFT.* We can see that as *LSQFT* increases, there is a positive trend, and *PRICE* increases as well. This indicates there is a positive association between the variables and we can also see a linear trend. The second scatterplot in the figure (***Figure 4.3***) shows the relationship between *PRICE* and *LOT*. The figure indicates that the relationship between *PRICE* and *LOT* is also a positive relationship. This means that as the lot size of the home increases, the appraised price of the home tends to be higher. (***Figures 4.4 & 4.5***) show the trend lines in the two scatterplots. As we can see from looking at the trend lines, ***Figure 3.2*** is more of a positive linear association, whereas ***Figure 3.3*** has a positive logarithmic association.

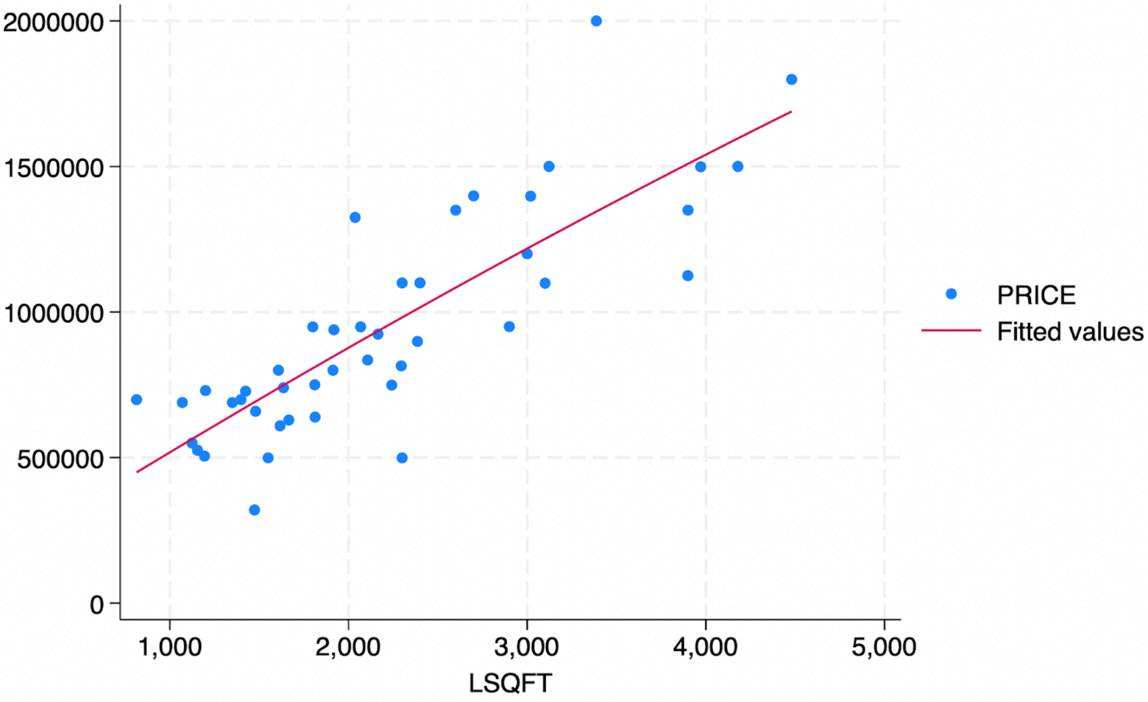
(***Figure 4.2***) n = 45



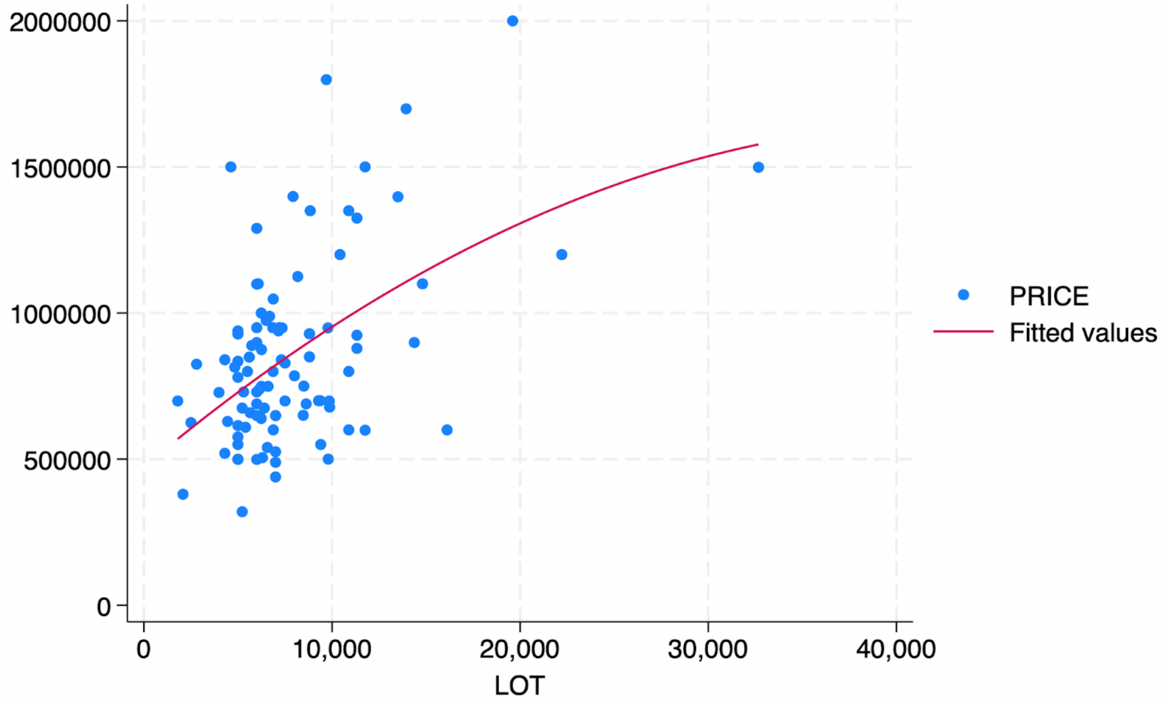
(***Figure 4.3***) n = 99



(***Figure 4.4***)



(***Figure 4.5***)



For this research, a robust multivariate log-linear regression model is estimated using STATA. 9 different models are analyzed to see how adding more regressors affects and either strengthens or weakens the model. 9 different models are analyzed to measure how *LOT* & *LSQFT* affect price both separately and together. The equations for the models analyzed are as shown below:

Model 1 (Independent Variable *LSQFT* Only):

$\boldsymbol{ln(PRICE) = \beta_0 + \beta_1(LSQFT)}$

Model 2 (Independent Variable *LOT* Only):

$\boldsymbol{ln(PRICE) = \beta_0 + \beta_1(LOT)}$

Model 3 (Both I.V *LSQFT* & *LOT*):

$\boldsymbol{ln(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2(LOT)}$

Model 4 (*LSQFT* & Control Variables):

$\boldsymbol{ln(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2(SCHOOL) + \beta_3(BDS) + \beta_4(BA) + \beta_5(NEW\_CONST)}$

Model 5 (*LOT* & Control Variables):

$\boldsymbol{ln(PRICE) = \beta_0 + \beta_1(LOT) + \beta_2(SCHOOL) + \beta_3(BDS) + \beta_4(BA) + \beta_5(NEW\_CONST)}$

Model 6 (Both I.V *LSQFT* & *LOT*, & Control Variables):

$ln(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2(LOT) + \beta_3(SCHOOL) + \beta_4(BDS) + \beta_5(BA) + \beta_6(NEW\_CONST)$

Model 7 (*LSQFT*, Control Variables, & Other):

$ln(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2(SCHOOL) + \beta_3(BDS) + \beta_4(BA) + \beta_5(NEW\_CONST) + \beta_6(WLK\_SCR) + \beta_7(TRNSIT\_SCR) + \beta_8(BIKE\_SCR)$

Model 8 (*LOT*, Control Variables, & Other):

$ln(PRICE) = \beta_0 + \beta_1(LOT) + \beta_2(SCHOOL) + \beta_3(BDS) + \beta_4(BA) + \beta_5(NEW\_CONST) + \beta_6(WLK\_SCR) + \beta_7(TRNSIT\_SCR) + \beta_8(BIKE\_SCR)$

Model 9 (Both I.V *LSQFT* & *LOT*, Control Variables, & Other):

$ln(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2(LOT) + \beta_3(SCHOOL) + \beta_4(BDS) + \beta_5(BA) + \beta_6(NEW\_CONST) + \beta_7(WLK\_SCR) + \beta_8(TRNSIT\_SCR) + \beta_9(BIKE\_SCR)$

The rationale behind using a robust log-linear multivariate regression is due to the non-linear relationship we see in the scatterplots in *Figure 4.5*. Alongside that, it is more beneficial to use a log-linear regression model because the coefficients can be interpreted as percentage changes. An increase in X by 1 unit changes Y by $\beta_1(100)$%. Using a log-linear regression model is valuable because it allows an analysis of how variations in a home size impact its price, expressed in a percentage change. The importance of this is that house prices change year-to-year depending on the real estate market, fluctuations occur often so it is more useful to use a log-linear model. Moreover, this approach can support comparisons over a period of time in a real estate market and is not just bound to the current market in which the data was analyzed. When it comes to using a robust model, its significance is to avoid any heteroskedasticity, and a robust regression provides a more accurate measurement of standard errors which helps with that case.

Alongside analyzing the regressions based on the models above, because in ***Figure 4.5*** we see that *LOT* is not linear, I measure *LOT* logarithmically using *ln(LOT)*. This is to increase the validity of the data and try to have the data avoid filling the exponential curve. All the models remain the same except for *PRICE,* which is not measured logarithmically and *LOT* is now measured using *ln(LOT)*. The comparison between the two different sets of models is explained in *Section 5.* The equations for the analysis of these new models are shown below:

Model 1 (Independent Variable *LSQFT* Only):

$$(PRICE) = \beta_0 + \beta_1(LSQFT)$$

Model 2 (Independent Variable *LOT* Only):

$$(PRICE) = \beta_0 + \beta_1 ln(LOT)$$

Model 3 (Both I.V *LSQFT* & *LOT*):

$$(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2 ln(LOT)$$

Model 4 (*LSQFT* & Control Variables):

$$(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2(SCHOOL) + \beta_3(BDS) + \beta_4(BA) + \beta_5(NEW\_CONST)$$

Model 5 (*LOT* & Control Variables):

$$(PRICE) = \beta_0 + \beta_1 ln(LOT) + \beta_2(SCHOOL) + \beta_3(BDS) + \beta_4(BA) + \beta_5(NEW\_CONST)$$

Model 6 (Both I.V *LSQFT* & *LOT*, & Control Variables):

$$(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2 ln(LOT) + \beta_3(SCHOOL) + \beta_4(BDS) + \beta_5(BA) + \beta_6(NEW\_CONST)$$

Model 7 (*LSQFT*, Control Variables, & Other):

$$(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2(SCHOOL) + \beta_3(BDS) + \beta_4(BA) + \beta_5(NEW\_CONST) + \beta_6(WLK\_SCR) + \beta_7(TRNSIT\_SCR) + \beta_8(BIKE\_SCR)$$

Model 8 (*LOT*, Control Variables, & Other):

$$(PRICE) = \beta_0 + \beta_1 ln(LOT) + \beta_2(SCHOOL) + \beta_3(BDS) + \beta_4(BA) + \beta_5(NEW\_CONST) + \beta_6(WLK\_SCR) + \beta_7(TRNSIT\_SCR) + \beta_8(BIKE\_SCR)$$

Model 9 (Both I.V *LSQFT* & *LOT*, Control Variables, & Other):

$$(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2 ln(LOT) + \beta_3(SCHOOL) + \beta_4(BDS) + \beta_5(BA) + \beta_6(NEW\_CONST) + \beta_7(WLK\_SCR) + \beta_8(TRNSIT\_SCR) + \beta_9(BIKE\_SCR)$$

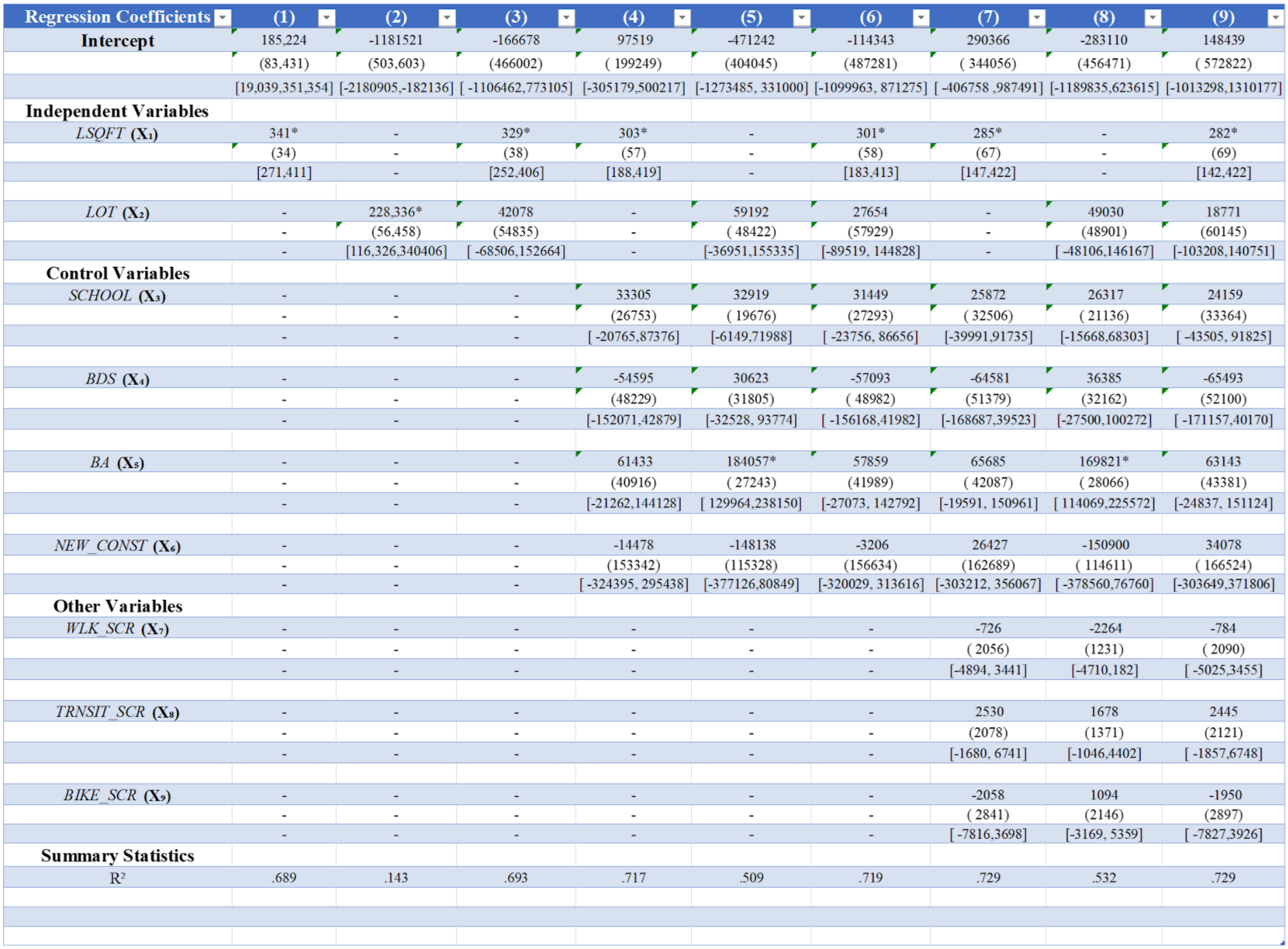
5. Results

This section highlights the findings and analysis of the nine robust multivariate log-linear models mentioned above. Below is a comparative regression analysis table that indicates the intercept and coefficient of each model and regressor (***Table 5.1***). Also included is the standard error for the coefficient as well as the confidence interval. A "*" next to the coefficient indicates that the regressor is statistically significant at a 5% significance level.

(***Table 5.1 Comparative Regression Analysis based on first set of models)***

Regression Coefficients	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
Intercept	12.926	13.302	12.906	12.794	12.624	12.843	12.976	12.694	13.060
	(.094)	(.066)	(.093)	(.232)	(.162)	(.231)	(.393)	(.237)	(.395)
	[12.736,13.116]	[13.172,13.433]	[12.717,13.094]	[12.326,13.263]	[12.300,12.946]	[12.376,13.310]	[12.179,13.72]	[12.222,13.165]	[12.258,13.862]
Independent Variables									
LSQFT **(X_1)**	.0003*	-	.0003*	.0003*	-	.0003*	.0003*	-	.0003*
	(.00004)	-	(00005)	(.00007)	-	(.00007)	(.00008)	-	(.00008)
	[.0003,.0004]	-	[.0002,.0004]	[.0001,.0004]	-	[.0001,.0004]	[.0001,.0004]	-	[.00009,.0004]
LOT **(X_2)**	-	.00004*	.00001	-	.00001*	.00001	-	.00001	.00001
	-	(7.30e-06)	(7.61e-06)	-	(6.69e-06)	(8.35e-06)	-	(6.85e-06)	(8.64e-06)
	-	[.00002,.00005]	[-2.82e-06,.00003]	-	[9.64e-08,.00003]	[-4.79e-06,.00003]	-	[-1.72e-06,.00003]	[-6.51e-06,.00003]
Control Variables									
SCHOOL **(X_3)**	-	-	-	.023	.038	.021	.019	.036	.015
	-	-	-	(.031)	(.022)	(.031)	(.037)	(.024)	(.037)
	-	-	-	[-.041,.086]	[-.004,.081]	[-.042,.083]	[-.057,.094]	[-.010,.083]	[-.059,.091]
BDS **(X_4)**	-	-	-	-.015	.046	-.035	-.033	.048	-.050
	-	-	-	(.056)	(.035)	(.057)	(.059)	(.036)	(.059)
	-	-	-	[-.128,.098]	[-.024,.115]	[-.149,.080]	[-.152,.086]	[-.024,.119]	[-.171,.071]
BA **(X_5)**	-	-	-	.065*	.178*	.053	.071*	.170*	.059*
	-	-	-	(.031)	(.030)	(.048)	(.031)	(.032)	(.029)
	-	-	-	[.001,.128]	[.117,.238]	[-.044,.149]	[.008,.134]	[.108,.233]	[-.0002.117]
NEW_CONST **(X_6)**	-	-	-	.039	-.160	.078	.087	-.162	.120
	-	-	-	(.178)	(.127)	(.177)	(.186)	(.128)	(.186)
	-	-	-	[-.320,.399]	[-.413,.092]	[-.281,.437]	[-.290,.464]	[-.416,.093]	[-.257,.498]
Other Variables									
WLK_SCR **(X_7)**	-	-	-	-	-	-	-.0003	-.001	-.0004
	-	-	-	-	-	-	(.002)	(.001)	(.002)
	-	-	-	-	-	-	[-.005,.004]	[-.004,.001]	[-.005,.004]
TRNSIT_SCR **(X_8)**	-	-	-	-	-	-	.004	.002	.003
	-	-	-	-	-	-	(.002)	(.002)	(.002)
	-	-	-	-	-	-	[-.001,.009]	[-.002,.005]	[-.001,.008]
BIKE_SCR **(X_9)**	-	-	-	-	-	-	-.003	.0001	-.003
	-	-	-	-	-	-	(.003)	(.002)	(.003)
	-	-	-	-	-	-	[-.009,.004]	[-.005,.005]	[-.009,.003]
Summary Statistics									
R^2	.639	.211	.661	.659	.509	.677	.683	.518	.697

(***Table 5.2 Comparative Regression Analysis based on second set of models***)



Regression Coefficients	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Intercept	185,224	-1181521	-166678	97519	-471242	-114343	290366	-283110	148439
	(83,431)	(503,603)	(466002)	(199249)	(404045)	(487281)	(344056)	(456471)	(572822)
	[19,039,351,354]	[-2180905,-182136]	[-1106462,773105]	[-305179,500217]	[-1273485, 331000]	[-1099963, 871275]	[-406758 ,987491]	[-1189835,623615]	[-1013298,1310177]
Independent Variables									
LSQFT (**X_1**)	341*	-	329*	303*	-	301*	285*	-	282*
	(34)	-	(38)	(57)	-	(58)	(67)	-	(69)
	[271,411]	-	[252,406]	[188,419]	-	[183,413]	[147,422]	-	[142,422]
LOT (**X_2**)	-	228,336*	42078	-	59192	27654	-	49030	18771
	-	(56,458)	(54835)	-	(48422)	(57929)	-	(48901)	(60145)
	-	[116,326,340406]	[-68506,152664]	-	[-36951,155335]	[-89519, 144828]	-	[-48106,146167]	[-103208,140751]
Control Variables									
SCHOOL (**X_3**)	-	-	-	33305	32919	31449	25872	26317	24159
	-	-	-	(26753)	(19676)	(27293)	(32506)	(21136)	(33364)
	-	-	-	[-20765,87376]	[-6149,71988]	[-23756, 86656]	[-39991,91735]	[-15668,68303]	[-43505, 91825]
BDS (**X_4**)	-	-	-	-54595	30623	-57093	-64581	36385	-65493
	-	-	-	(48229)	(31805)	(48982)	(51379)	(32162)	(52100)
	-	-	-	[-152071,42879]	[-32528, 93774]	[-156168,41982]	[-168687,39523]	[-27500,100272]	[-171157,40170]
BA (**X_5**)	-	-	-	61433	184057*	57859	65685	169821*	63143
	-	-	-	(40916)	(27243)	(41989)	(42087)	(28066)	(43381)
	-	-	-	[-21262,144128]	[129964,238150]	[-27073, 142792]	[-19591, 150961]	[114069,225572]	[-24837, 151124]
NEW_CONST (**X_6**)	-	-	-	-14478	-148138	-3206	26427	-150900	34078
	-	-	-	(153342)	(115328)	(156634)	(162689)	(114611)	(166524)
	-	-	-	[-324395, 295438]	[-377126,80849]	[-320029, 313616]	[-303212, 356067]	[-378560,76760]	[-303649,371806]
Other Variables									
WLK_SCR (**X_7**)	-	-	-	-	-	-	-726	-2264	-784
	-	-	-	-	-	-	(2056)	(1231)	(2090)
	-	-	-	-	-	-	[-4894, 3441]	[-4710,182]	[-5025,3455]
TRNSIT_SCR (**X_8**)	-	-	-	-	-	-	2530	1678	2445
	-	-	-	-	-	-	(2078)	(1371)	(2121)
	-	-	-	-	-	-	[-1680, 6741]	[-1046,4402]	[-1857,6748]
BIKE_SCR (**X_9**)	-	-	-	-	-	-	-2058	1094	-1950
	-	-	-	-	-	-	(2841)	(2146)	(2897)
	-	-	-	-	-	-	[-7816,3698]	[-3169, 5359]	[-7827,3926]
Summary Statistics									
R^2	.689	.143	.693	.717	.509	.719	.729	.532	.729

Looking at the results in *Table 5.1*, we can see that *LSQFT* was statistically significant in all the models it was measured in (model 1, model 3, model 4, model 6, model 7, model 9). In contrast, *LOT* was statistically significant solely in model 2 and model 5. When looking at the rest of the control variables, none of them were statistically significant except for *BA* in all models besides model 6. The R^2 value for all models remained relatively consistent except in model 2 where it decreased drastically. The decrease suggests a weaker fit of the data in the regression model when *LSQFT* is not considered, underlining its importance in explaining variance in the dependent variable *PRICE*. We also see this when we look at the models that don't include LSQFT, the R^2 is always lower. It is also shown that the coefficients of the independent variables do not change much across the different models, which strengthens the behavioral hypothesis above in *section 3*. This consistency suggests that the relationships between the independent variables and *PRICE* are robust and relatively unchanged. Going back to the null hypothesis in *section 3*, we can see that ***Table 5.1*** shows that in model 3, the F-statistic is 49.02 which is greater than the critical value at a 5% significance level, allowing for a rejection of the null hypothesis. The rejection of the null hypothesis signifies that the overall regression model has significant differences and relationships between the variables.

Looking at the results in ***Table 5.2***, where *LOT* was measured using a natural log, we see that *LOT* is only statistically significant in one of the models, model 2, which was also the only model in which the R^2 decreased when you compare the results to the models in ***Table 5.1***. Comparing the two different tables, we see that the R^2 and the goodness of fit of the models improve when *LOT* is measured as ln(*LOT*). The benchmark model will also remain the same as model 9. When analyzing the results of the two different tables, the data tells us the same thing. Which is that size does matter, *LSQFT* and *LOT* both have a positive impact on *PRICE*, however, *LSQFT* plays a more significant role because lot size does not hold as much significance when compared to the livable square footage. The interpretations of the coefficients can be seen below and are based on the data in ***Table 5.1***, and the following is the equation for the benchmark model in ***Table 5.2***:

$$(PRICE) = \beta_0 + \beta_1(LSQFT) + \beta_2 \ln(LOT) + \beta_3(SCHOOL) + \beta_4(BDS) + \beta_5(BA) + \beta_6(NEW\_CONST) + \beta_7(WLK\_SCR) + \beta_8(TRNSIT\_SCR) + \beta_9(BIKE\_SCR)$$

Furthermore, after a careful analysis of the results, it can be concluded that the benchmark specification of the data is model 9 in ***Table 5.1***. Even though the R^2 of model 9 in ***Table 5.2*** is higher. The reasoning in choosing model 9 in ***Table 5.1*** over model 9 in ***Table 5.2*** is due to its inclusion of a greater number of variables that demonstrate statistical significant at the 5% level, allowing for a better representation of the population. Additionally, the marginal increase of approximately 3% in R^2 between the two models is negligible.

Approximately 69.7% of the observed variance in *PRICE* can be explained by the variation in *LSQFT, LOT, SCHOOL, BDS, BA, NEW_CONST, WLK_SCORE, TRNSIT_SCR,* and *BIKE_SCR* in model 9 from ***Table 5.1***. Alongside the justification of the high R^2, model 9 also contains all of the independent and control variables present in all models, thus mitigating the risk of omitted variable bias with the addition of more variables. Even though most of the variables were not statistically significant, model 9 includes *BA* and *LSQFT* as the two statistically significant variables. The equation for the benchmark model is as follows:

ln(PRICE) = 13.060 + .0003(LSQFT) + .00001(LOT) + .015(SCHOOL) - .050(BDS) + .059(BA) + .120(NEW_CONST) - .0004(WLK_SCR) + .003(TRNSIT_SCR) - .003(BIKE_SCR)

The interpretations of the coefficients are as follows. The constant is not interpreted because it does not make sense in the context of this analysis.

(X_1): The benchmark, model 9, suggests that a change in *LSQFT* by 1 sq ft increases *PRICE* by .03%, with the result being significant at the 5% level. The 95% confidence interval can be viewed in ***Table 5.1***. This can be considered as a relatively large impact especially when considering the price of a home. For example, homes are sold in the dollar amounts of hundreds of thousands, and a .03% of hundreds of thousands can accumulate to a large amount of change when you are increasing by .03% per square foot.

(X_2): The benchmark, model 9, suggests that a change in *LOT* by 1 sq ft increases *PRICE* by .001%, with the result being not statistically significant at the 5% level. The 95% confidence interval can be viewed in ***Table 5.1***. The increase in square footage of the lot does not have a relatively large impact on the price of a home, especially when compared to the increase in square footage of *LSQFT*.

(X_3): The benchmark, model 9, suggests that an increase in *SCHOOL* by 1 rating, increases *PRICE* by 1.5%, with the result being not statistically significant at the 5% level. The 95% confidence interval can be viewed in ***Table 5.1***. Even though this impact is relatively large, however, due to the fact that *SCHOOL* is not statistically significant, the impact is not significant enough to be considered beyond what could reasonably occur by random chance.

(X_4): The benchmark, model 9, suggests that an increase in *BDS* by 1 bedroom, decreases *PRICE* by 5%, with the result being not statistically significant at the 5% level. The 95% confidence interval can be viewed in ***Table 5.1***. Even though this impact is relatively large, due to the fact that *BDS* is not statistically significant, the impact is not significant enough to be considered beyond what could reasonably occur by random chance.

(X_5): The benchmark, model 9, suggests that an increase in *BA* by 1 bathroom, increases *PRICE* by 5.9%, with the result being statistically significant at the 5% level. The 95% confidence interval can be viewed in ***Table 5.1***. This can be considered as a relatively large impact especially when considering the price of a home, as well as the fact that *BA* is statistically significant.

(X_6): The benchmark, model 9, suggests that if the house is a new construction, *PRICE* increases by 12%, with the result being not statistically significant at the 5% level. The 95% confidence interval can be viewed in ***Table 5.1***. Even though this impact is relatively large, due to the fact that *NEW_CONST* is not statistically significant, the impact is not significant enough to be considered beyond what could reasonably occur by random chance.

(X_7): The benchmark, model 9, suggests that an increase in *WLK_SCR* by 1 point, decreases *PRICE* by .04%, with the result being not statistically significant at the 5% level. The 95% confidence interval can be viewed in ***Table 5.1***. Even though this does have a relative impact on the price of a home, due to the fact that *WLK_SCR* is not statistically significant, the impact is not significant enough to be considered beyond what could reasonably occur by random chance.

(X_8): The benchmark, model 9, suggests that an increase in *TRNSIT_SCR* by 1 point, increases *PRICE* by .3%, with the result being not statistically significant at the 5% level. The 95% confidence interval can be viewed in ***Table 5.1***. Even though this does have a relative impact on the price of a home, due to the fact that *TRNSIT_SCR* is not statistically significant, the impact is not significant enough to be considered beyond what could reasonably occur by random chance.

(X_9): The benchmark, model 9, suggests that an increase in *BIKE_SCR* by 1 point, decreases *PRICE* by .3%, with the result being not statistically significant at the 5% level. The 95% confidence interval can be viewed in ***Table 5.1***. Even though this does have a relative impact on the price of a home, due to the fact that *BIKE_SCR* is not statistically significant, the impact is not significant enough to be considered beyond what could reasonably occur by random chance.

6. Conclusion

Regarding the limitations of this research, there is a multitude of more variables that can impact the price of a home that was not included in this research but can be included in future works. For example, the age of the home, amenities, as well as neighborhood demographics such as income levels and age of the local population. Adding more variables such as these will further help reduce the risk of omitted variable bias. Another limitation was the number of observations that provided the *LSQFT*. Due to approximately 55% of the observations not providing data on the interior square footage of a home, this made it difficult to properly control for and measure *LSQFT* in the sample space. Lastly, to enhance this research, it is recommended to broaden the scope of the sample size by including a larger number of towns in Nassau County to obtain a larger data set. Increasing the sample size will contribute to a more precise representation of the population, helping strengthen the statistical validity of the results and increasing the understanding of the different factors influencing home prices in Nassau County.

In conclusion, in both sets of models, this study suggests that an increase in house size, in square footage, does have a positive increase on the price/appraisal value of the home. However, the impact of an increase in lot size is not nearly as significant as the impact of the liveable square footage of a house. The results in the data show that the livable square footage is more significant and has far more of an impact in percentage terms on a home's price. This can be due to various reasons as a buyer/seller's demands and needs fluctuate continuously in a market. But a somewhat understandable assumption could be due to the fact that the liveable square footage is the space in the house that you are able to utilize and that is what provides more utility to a buyer/seller. In contrast, the entirety of the lot may include areas that are not usable or beneficial, diminishing its impact on property value. Based on this analysis of data we can conclude that there is a somewhat strong goodness of fit in the benchmark model of the data, in the fact that 69.7% of the observed variance in *PRICE* can be explained by the variation in the independent variables considered. Overall, size does matter, at least when it comes to the interior of a home.

7. Citations

Bond, Michael, et al. "Residential real estate prices: A room with A view." *Journal of Real Estate Research*, vol. 23, no. 1–2, 2002, pp. 129–138, https://doi.org/10.1080/10835547.2002.12091077.

Asabere, Paul. (2013). The Impact of Relative Size on Home ValuesThe Appraisal Journal, Winter 2013. The Appraisal Journal.