SPSS Group Project - Case Study 2

Modelling The Sales Prices of Residential Properties in Four Neighbourhoods

Group 2

Francesco Figliano #501117218

Kevin Modi # 501022942

Course: CQMS 442 DA0 - Multiple Regressions for Business

Professor: Pauline Fu

Date: July 15, 2023

The Problem

The purpose of this case study is to examine the relationship between the mean sale price **E(y)** of a property and the following independent variables:

1. The appraised value of the property's land
2. The appraised value of the property's improvements
3. The neighbourhood in which the property is listed

Research Objective

The objective of this case study is concerned with the relationship between the sales price and the appraised value of a property. The study ultimately has two goals:

1. This report will assess the extent to which the data suggest a correlation between appraised land values and sales price improvements. Specifically, the aim is to determine whether the data provide substantial evidence to support the notion that these variables offer valuable information for predicting the sales price.
2. To determine if appraisers utilize the same appraisal criteria for various types of neighbourhoods by acquiring the prediction equation concerning the land's appraised value and improvements to sales price which will assess the relationship against various neighbourhoods.

Data: (TAMSALES-ALL.xlsx)

The textbook provided the data used in this case study and was downloaded from D2L. The data set was initially sourced from the property appraiser's office in Hillsborough County, Florida, and consists of the appraised land values, improvement values, and sale prices for 412 randomly selected residential properties sold in Tampa, Florida, between January 2017 and January 2018. Below is more information regarding the data.

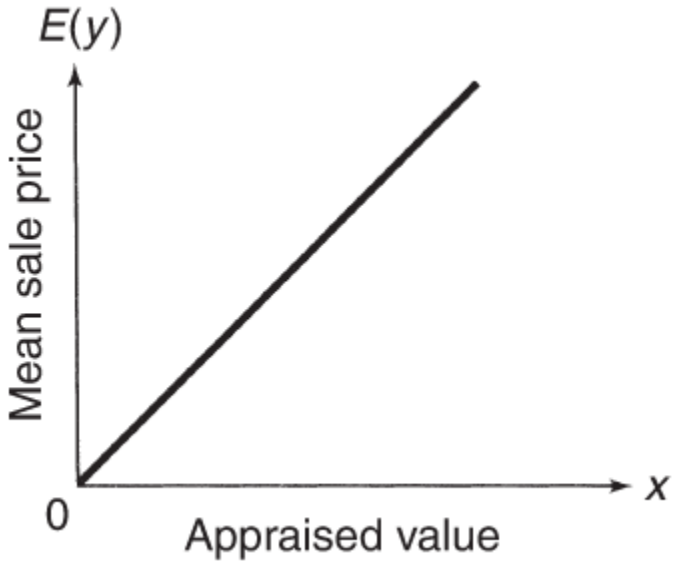
- Four relatively homogenous neighbourhoods were selected, each similar but varying sociologically and in property types and values.
    - Town & Country, Cheval, North Dale, and Davis Isles
- All values are in thousands of dollars USD.

This subset of sales and appraisal data for these four neighbourhoods was used to develop a predictive equation that relates sales price to appraised land and improvement values.

The Models & an Analysis of the Models

The Theoretical Model

The model we designed seeks to accurately predict value ŷ and correlate it to the mean sales price of y. Upon completing the model, we understood how closely our independent variables (appraisal value of land, improvements valuation, & neighbourhood location) reflect on the selling price of residential properties on the market in the data sets area.



The objective of our research would be to study the different pricing models to accurately determine the most useful in predicting the selling price of the 412 randomly selected homes in Tampa, Florida, from January 2017 to January 2018.

A combination of the land's value and improvement value makes up the independent variable of the appraised value. In an ideal situation, the appraised value would be equal to the mean sales price of the home represented by a straight line with a slope of 1.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.972[a]	.945	.945	111550.998

a. Predictors: (Constant), TOTVAL

b. Dependent Variable: SALES

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.713E+13	1	8.713E+13	7001.726	<.001[b]
	Residual	5.102E+12	410	12443625100		
	Total	9.223E+13	411			

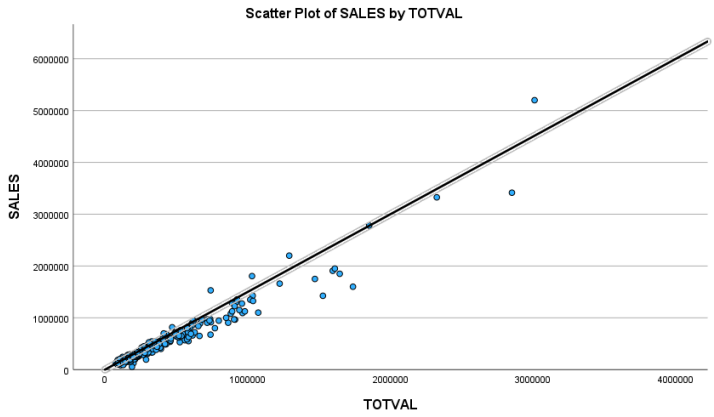
a. Dependent Variable: SALES

b. Predictors: (Constant), TOTVAL

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-11256.699	7643.953		-1.473	.142
	TOTVAL	1.326	.016	.972	83.676	<.001

a. Dependent Variable: SALES



The scatter plot constructed through SPSS allows insight into the relationship between sales and appraisal value of the land and the scatterplot is based on all 412 observations in the data set. Despite many of the concerns in this report, the model is statistically fit to be a linear model based on the scatterplot and the ANOVA table. The p-value and F-statistic in the ANOVA table illustrate that there is statistical significance due to the very low p-value (<0.01) and extensive F statistic (7000.726). Although, the variation must be acknowledged, which may have resulted from market trends and over/under-sold houses due to economic conditions. Further research and exploration through various models would allow us to achieve a better result to verify or dismiss the objectives in this report.

The Hypothesized Regression Models

This section will seek to relate the sale price (y) to three independent variables. The independent variables are the qualitative factor, neighbourhood, and the two quantitative factors, appraised land value and appraised improvement value. We will deploy four models to assist in making inferences.

Model 1

The first model is a first-order or, in other words, a linear model that will evaluate the relationship between Sale Price E(y) and the two independent variables: appraised value of the land which is X_1 in thousands of dollars, and apprised improvement value X_2 in thousands of dollars. The first-order model is identical for all neighbourhoods, and it will trace a response place for a mean of the sale price and the two independent variables X_1 and X_2. The model below

assumes that the sale price change for every unit increase in X_1 is constant for a fixed X_2. The model also assumes that the change in y for every one unit increase in X_2 is constant for a fixed X_1. This relationship ultimately indicates that the model is identical for all neighbourhoods.

$$E(y) = \beta_0 + \underbrace{\hat{\beta}_1 x_1}_{\text{Appraised land value}} + \underbrace{\hat{\beta}_2 x_2}_{\text{Appraised improvement value}}$$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.977[a]	.954	.954	102116.873

a. Predictors: (Constant), IMP, LAND

b. Dependent Variable: SALES

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.796E+13	2	4.398E+13	4217.730	<.001[b]
	Residual	4.265E+12	409	10427855709		
	Total	9.223E+13	411			

a. Dependent Variable: SALES

b. Predictors: (Constant), IMP, LAND

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1980.676	7073.680		-.280	.780
	LAND	1.649	.039	.578	42.401	<.001
	IMP	1.095	.030	.504	36.948	<.001

a. Dependent Variable: SALES

The SPSS output data suggests that the model is statistically significant due to the values displayed within the ANOVA table. Additionally, the regression statistics from the coefficients table and model summary suggest that the model fits the data well. In conclusion, when we examine the ANOVA table, we find that if it reveals a statistically significant model and we observe significant coefficients and a high R-squared value in the regression statistics presented in the model summary, we can confidently state that the model is statistically significant and fits the data quite well.

Model 2

Model 2 is also a first-order model, which assumes that the relationship between **E(y)** and $\mathbf{x_1}$ and $\mathbf{x_2}$ are still first-order but that the planes' y-intercept changes depending on the neighbourhood used in the model.

$$E(y) = \beta_0 + \underbrace{\hat{\beta}_1 x_1}_{\text{Appraised land value}} + \underbrace{\hat{\beta}_2 x_2}_{\text{Appraised improvement value}} + \underbrace{\beta_3 x_3 + \hat{\beta}_4 x_4 + \beta_5 x_5}_{\text{Main effect terms for neighborhoods}}$$

where

The base level is = Town & Country

$$x_1 = \text{Appraised land value}$$

$$x_2 = \text{Appraised improvement value}$$

$$x_3 = \begin{cases} 1 & \text{if Cheval neighborhood} \\ 0 & \text{if not} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if Davis Isles neighborhood} \\ 0 & \text{if not} \end{cases}$$

$$x_5 = \begin{cases} 1 & \text{if North Dale neighborhood} \\ 0 & \text{if not} \end{cases}$$

The fourth neighbourhood, Town & Country, is the base level. This model predicts E(y) for T&C when $x_3 = x_4 = x_5 = 0$. Model 2 aids us in studying the change in variable y for increases in x_1 or x_2 that vary depending on which neighbourhood.

Furthermore, there are no interactions between independent variables x_1 and x_2 and the terms of the neighbourhood. With that being said, model 2 assumes that fluctuations in sales price y for every dollar increase in x_1 and x_2 are not reliant on the neighbourhood.

As for whether or not this model has appropriate applications, we concluded that this model is appropriate if an appraiser establishes a value based on the mean sales price and x_1 and x_2 that differs in two or more neighbourhoods that remained constant for values of x_1 and x_2.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.979[a]	.958	.958	97251.531

a. Predictors: (Constant), x5, IMP_x2, x3, x4, LAND_x1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.839E+13	5	1.768E+13	1869.109	<.001[b]
	Residual	3.840E+12	406	9457860278.5		
	Total	9.223E+13	411			

a. Dependent Variable: SALES_E(y)

b. Predictors: (Constant), x5, IMP_x2, x3, x4, LAND_x1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3560.405	8133.658		.438	.662
	LAND_x1	1.982	.062	.695	31.828	<.001
	IMP_x2	1.029	.031	.474	32.737	<.001
	x3	-14269.552	13622.438	-.013	-1.048	.295
	x4	-145508.176	22346.382	-.124	-6.511	<.001
	x5	913.034	13611.543	.001	.067	.947

a. Dependent Variable: SALES_E(y)

Based on the SPSS output, the data indicates that the model has statistical significance based on the values presented in the ANOVA and model summary. However, upon analyzing the coefficients table in the regression statistics, it is evident that the two coefficients are not statistically significant. Nevertheless, overall the model demonstrates a satisfactory fit to the data. Thus, the model possesses statistical significance and exhibits a reasonably good fit for the data.

Model 3

Model 3 is also another first-order model. However, it is very similar to model 2 because we have now added "interaction terms" for the dummy variables. These terms correspond to each neighbourhood, and the variables x_1 and x_2. Model 3 allows changes in y, via increases in x_1 or x_2, to change with a given neighbourhood.

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1}_{\text{Appraised land value}} + \underbrace{\beta_2 x_2}_{\text{Appraised improvement value}} + \underbrace{\beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5}_{\text{Main effect terms for neighborhoods}} + \underbrace{\beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \beta_8 x_1 x_5}_{\text{Interaction, appraised land by neighborhood}} + \underbrace{\beta_9 x_2 x_3 + \beta_{10} x_2 x_4 + \beta_{11} x_2 x_5}_{\text{Interaction, appraised improvement by neighborhood}}$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.981[a]	.962	.961	93997.732

a. Predictors: (Constant), x2x5, IMP_x2, x3, x4, x1x3, LAND_x1, x1x5, x2x3, x5, x2x4, x1x4

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.869E+13	11	8.063E+12	912.577	<.001[b]
	Residual	3.534E+12	400	8835573562.2		
	Total	9.223E+13	411			

a. Dependent Variable: SALES_E(y)

b. Predictors: (Constant), x2x5, IMP_x2, x3, x4, x1x3, LAND_x1, x1x5, x2x3, x5, x2x4, x1x4

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	52009.011	31891.044		1.631	.104
	LAND_x1	1.074	.904	.377	1.189	.235
	IMP_x2	.874	.207	.402	4.220	<.001
	x3	22486.671	38038.092	.020	.591	.555
	x4	-214462.773	38604.926	-.183	-5.555	<.001
	x5	73237.848	72558.417	.060	1.009	.313
	x1x3	.601	.954	.062	.630	.529
	x1x4	.897	.906	.343	.990	.323
	x1x5	.065	1.296	.003	.050	.960
	x2x3	-.026	.230	-.008	-.112	.911
	x2x4	.225	.210	.106	1.074	.283
	x2x5	-.312	.349	-.044	-.895	.371

a. Dependent Variable: SALES_E(y)

According to the SPSS output, the data implies that the model holds statistical significance based on the information presented within the ANOVA table. However, upon scrutinizing the coefficients table and model summary in the regression statistics, it becomes apparent that only

two coefficients out of several exhibit statistical significance. Nonetheless, the model showcases a satisfactory fit to the data when considering the overall picture. Thus, we can confidently conclude that the model is statistically significant, albeit with only a subset of coefficients displaying significance, and it demonstrates a commendable fit to the data.

Model 4

Interaction model in x_1 and x2 differs between neighbourhoods.

Interaction model in x_1 and x_2 | Main effect terms for neighborhoods

$$E(y) = \beta_0 + \beta_1 x_1 + \hat{\beta}_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_3 + \hat{\beta}_5 x_4 + \beta_6 x_5$$
$$+\beta_7 x_1 x_3 + \beta_8 x_1 x_4 + \beta_9 x_1 x_5 + \beta_{10} x_2 x_3$$
$$+\beta_{11} x_2 x_4 + \beta_{12} x_2 x_5 + \beta_{13} x_1 x_2 x_3$$
$$+\beta_{14} x_1 x_2 x_4 + \beta_{15} x_1 x_2 x_5$$

Interaction terms: x_1, x_2, and $x_1 x_2$ terms by neighborhood

Unlike models 1–3, Model 4 allows for a change in y for increases in x1 to depend on x2 and vice versa. Example: the difference in the sale price for a $1000 increase in appraised land value in the base-level neighbourhood (T & C) is $\beta_1 + \beta_3 x_2$. Model 4 also accounts for these sales prices to differ from neighbourhood to neighbourhood (due to the inclusion of the interaction terms).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.983[a]	.966	.964	89404.251

a. Predictors: (Constant), x1x2x5, IMP_x2, x3, x4, x1x2x3, x1x2x4, x5, LAND_x1, x1x3, x2x5, x2x3, x2x4, x1x5, x1x4

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.906E+13	14	6.361E+12	795.823	<.001[b]
	Residual	3.173E+12	397	7993120070.2		
	Total	9.223E+13	411			

a. Dependent Variable: SALES_E(y)

b. Predictors: (Constant), x1x2x5, IMP_x2, x3, x4, x1x2x3, x1x2x4, x5, LAND_x1, x1x3, x2x5, x2x3, x2x4, x1x5, x1x4

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	52009.011	30332.592		1.715	.087
	LAND_x1	1.074	.860	.377	1.250	.212
	IMP_x2	.874	.197	.402	4.437	<.001
	x3	-24412.295	52021.468	-.021	-.469	.639
	x4	4703.330	49500.942	.004	.095	.924
	x5	80882.174	241290.364	.066	.335	.738
	x1x3	.882	.935	.091	.943	.346
	x1x4	.288	.867	.110	.332	.740
	x1x5	-.098	5.086	-.004	-.019	.985
	x2x3	.149	.259	.047	.574	.567
	x2x4	-.058	.204	-.027	-.285	.776
	x2x5	-.360	1.472	-.051	-.245	.807
	x1x2x3	-8.889e-7	.000	-.054	-1.255	.210
	x1x2x4	6.259e-7	.000	.231	6.602	<.001
	x1x2x5	1.006e-6	.000	.007	.033	.974

a. Dependent Variable: SALES_E(y)

Based on the SPSS output, the data suggests that the model is statistically significant, as indicated by the values in the ANOVA table. However, when we look at the coefficients table and model summary in the regression statistics, we find that only two coefficients are statistically significant. Nevertheless, the model still fits the data well. To sum it up, if the ANOVA table shows a statistically significant model and we have significant coefficients and a high R-squared value in the regression statistics, we can confidently say that the model is statistically significant and fits the data nicely.

Model Comparative Analysis

Summary of Regressions of the Models

Model	MSE	R^2_a	s
1	10427855709	0.954	102116.873
2	9457860278.5	0.958	97251.531
3	88355735622	0.961	93997.732
4	7993120070.2	0.964	89404.251

We will be utilizing a conservative approach to compare the models in which $\alpha = 0.01$

Test #1: Model 1 vs Model 2

H0: $\beta 3 = \beta 4 = \beta 5 = 0$
Ha: at least one β parameter is non-zero

$$F = \frac{(SSE_R - SSE_C)/\text{Number of } \beta \text{ parameters in } H_0}{MSE_C}$$

Where the reduced model is model 1 and the complete model is model 2

$$= \frac{(SSE_1 - SSE_2)/3}{MSE_2}$$

$$= \frac{(4.265E{+}12 - 3.840E{+}12)/3}{9457860278.5}$$

$$= 14.9787227233$$

Rejection Region for the F-test is when $F > F\alpha$
$F\alpha = 3.83006537$ ($d_{f1} = 3$, $d_{f2} = 407$)
Therefore, reject H0 since $F > F\alpha$, the terms contribute to the prediction of y

Test #2: Model 2 vs Model 3

H0: $\beta 6 = \beta 7 = \beta 8 = \beta 9 = \beta 10 = \beta 11 = 0$
Ha: at least one β parameter is non-zero

$$F = \frac{(SSE_R - SSE_C)/\text{Number of } \beta \text{ parameters in } H_0}{MSE_C}$$

Where the reduced model is model 2 and the complete model is model 3

$$= \frac{(SSE_2 - SSE_3)/6}{MSE_3}$$

$$= \frac{(3.840E+12 - 3.534E+12)/6}{88355735622}$$

= 0.865818155

Rejection Region for the F-test is when F > $F\alpha$

$F\alpha$ = 2.84712279 (d_{f1} = 6, d_{f2} = 401)

Therefore, do not reject H0 since F < $F\alpha$, the terms do not contribute to the prediction of y indicating that there is insufficient evidence to support the null hypothesis

Test #3: Model 3 vs Model 4

H0: $\beta 3 = \beta 13 = \beta 14 = \beta 15 = 0$

Ha: atleast one β parameter is non-zero

$$F = \frac{(SSE_R - SSE_C)/\text{Number of } \beta \text{ parameters in } H_0}{MSE_C}$$

Where the reduced model is model 3 and the complete model is model 4

$$= \frac{(SSE_3 - SSE_4)/4}{MSE_4}$$

$$= \frac{(3.534E+12 - 3.173E+12)/4}{7993120070.2}$$

= 11.2909601266

Rejection Region for the F-test is when F > $F\alpha$

$F\alpha$ = 3.36671267 (d_{f1} = 4, d_{f2} = 397)

Therefore, reject H0 since F > $F\alpha$, the terms contribute to the prediction of y indicating that there is sufficient evidence to support null hypothesis.

Interpreting the Prediction Equation

$$E(y) = \beta_0 + \underbrace{\hat{\beta}_1 x_1}_{\text{Appraised land value}} + \underbrace{\hat{\beta}_2 x_2}_{\text{Appraised improvement value}} + \underbrace{\beta_3 x_3 + \hat{\beta}_4 x_4 + \beta_5 x_5}_{\text{Main effect terms for neighborhoods}}$$

Substituting the estimate parameters from model 2 yields the following prediction equation:

$$\hat{y} = 3560.405 + 1.982x1 + 1.029x2 - 14269.552x3 - 145508.176x4 + 913.034x5$$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3560.405	8133.658		.438	.662
	LAND_x1	1.982	.062	.695	31.828	<.001
	IMP_x2	1.029	.031	.474	32.737	<.001
	x3	-14269.552	13622.438	-.013	-1.048	.295
	x4	-145508.176	22346.382	-.124	-6.511	<.001
	x5	913.034	13611.543	.001	.067	.947

a. Dependent Variable: SALES_E(y)

The model above predicts four response surfaces which allow us to have a deeper look into the effect on sale and appraisal value based on the neighbourhoods. We obtain the following models below by substituting the appropriate dummy variables:

Town & Country (x3 = 0, x4 = 0, x5 = 0)

$$\hat{y} = 3560.405 + 1.982x1 + 1.029x2$$

Cheval (x3 = 1, x4 = 0, x5 = 0)

$$\hat{y} = 3560.405 + 1.982x1 + 1.029x2 - 14269.552x3$$

Davis Isles (x3 = 0, x4 = 1, x5 = 0)

$$\hat{y} = 3560.405 + 1.982x1 + 1.029x2 - 145508.176x4$$

North Dale (x3 = 0, x4 = 0, x5 = 1)

$$\hat{y} = 3560.405 + 1.982x1 + 1.029x2 + 913.034x5$$

We can further analyze the equation by holding one independent variable fixed such as x_1 and focus on the slope of the line relating y to a different variable such as x_2. Evidence suggests that there is variance in the sale price and appraised value across the neighbourhoods based on the rate of higher sales price increases with more expensive neighbourhoods tending to be more significant. Also, there is no interaction with the appraised land value when there is an increase in sale price with improvements.

Conclusion

The findings from our regression analyses reveal an interesting trend. We discovered that the relationship between property sale prices and appraised values differs significantly across various neighbourhoods. This implies that property price factors vary significantly depending on the specific neighbourhood context.

Moreover, the prediction intervals obtained from our analyses indicate substantial room for improvement in the methods utilized to determine appraised property values. The wide prediction intervals suggest that the current appraisal criteria may not adequately capture all the pertinent factors influencing property prices, resulting in less accurate predictions.

Furthermore, future research and appraisal practices must consider each neighbourhood's distinct characteristics and dynamics when evaluating property values. This could involve incorporating additional variables or refining the existing appraisal criteria to more accurately capture the unique factors influencing property prices within different neighbourhoods.

By delving deeper into the interesting relationships between appraised values and sale prices across neighbourhoods, stakeholders in the real estate industry can make better-informed decisions regarding property valuation, investment strategies, and pricing. Ultimately, enhancing the accuracy of appraised values will foster a more transparent and efficient real estate market, benefiting both buyers and sellers in their decision-making processes.