The Problem

This report will determine the statistical relationship between the mean sales price of a home and several independent variables. The variables tested include the appraised land value of the property, appraised value of improvements on the property, and the neighborhood in which the property is listed. The goal of this report is to measure the change in the mean sales price as appraised land value and appraised value of improvements change, and to determine if appraisers change their criteria depending on the neighborhood the house is in. The data of 8 different neighborhoods (Hyde Park, Cheval, Hunter's Green, Davis Isles, Avila, Carrollwood, Tampa Palms, and Town & Country) will be used in the study, which was obtained through an observational study.

The Data

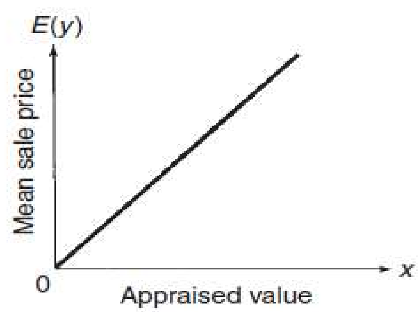
TAMSALES8

The data for the study were provided by the property appraiser's office of Hillsborough County, Florida. It consists of the appraised land and improvement values and sale prices for residential properties sold in the city of Tampa, Florida, from May 2008 to June 2009. Eight neighborhoods (Hyde Park, Cheval, Hunter's Green, Davis Isles, Avila, Carrollwood, Tampa Palms, and Town & Country), each relatively similar but different sociologically with differing property types and values. The subset of sales and appraisal data pertinent to these eight neighborhoods— a total of 350 observations—was used to develop a prediction equation relating sale prices to appraised land and improvement values. The data (recorded in thousands of dollars) are saved in the TAMSALES8 file and are described in the Appendix.

Theoretical Model

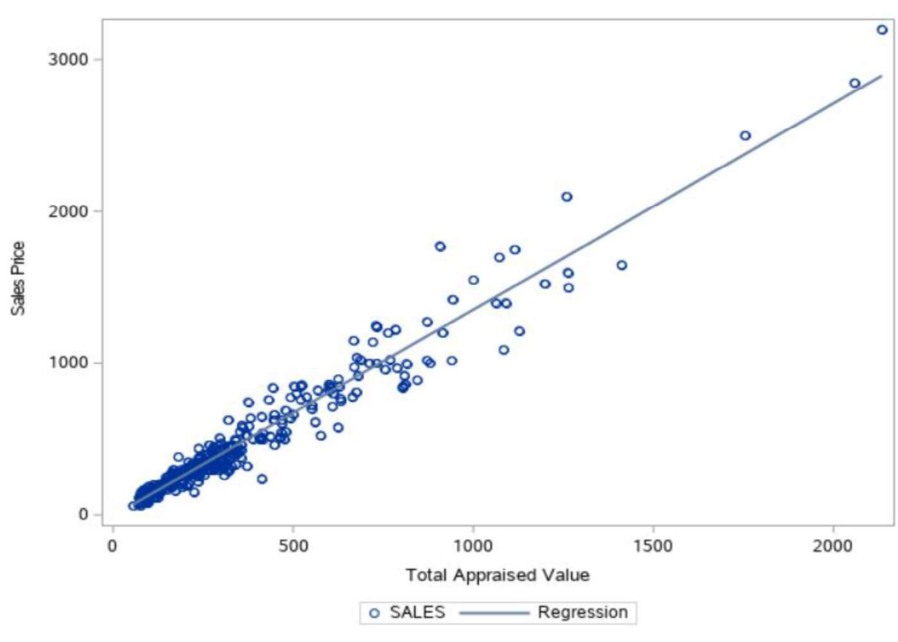
If the mean sale price (E(y)) were equal to the appraised value, then the relationship between E(y) and x (total appraised value) would be a straight line with the slope equal to 1. As seen in Figure CS1.1 below.

Figure CS1.1
Theoretical relationship between mean sale price and appraised value x



This may not be the case in reality as there are varying factors that might affect the sale price, such as inflation, overappraisal, and underappraisal of the property. In Figure CS1.2, we used SAS to conduct a scatter plot of sale price versus total appraised value for all 350 observations in the data set. Based on the figure, it appears that the theoretical model will fit the data well. We used y = sale price (in thousands of dollars) as dependent variable and considered only first-order linear models.

Figure CS1.2
SAS scatter plot of sales-appraised value



Hypothesized Regression Models

Model 1. *First-order model, identical for all neighborhoods*

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

x_1= *Appraised land value*

x_2= *Appraised improvement value*

For Model 1, we are assuming that the change in the sale price y for every \$1000 (1 unit) increase in x_1 is constant for constant x_2 value. Likewise, the change in y for every \$1000 (1 unit) increase in x_2 is constant for constant x_1 value.

Model 2. *First-order model, factoring in differences in each neighborhood*

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9$$

For Model 2, we hold Tampa Palms as the base level. This model will predict when E(y) for this neighborhood when $x_3 = \ldots = x_9 = 0$. This model assumes that the change in sale price y for every \$1000 increase in either x_1 or x_2 does not depend on the neighborhood.

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

$$x_4 = \begin{cases} 1 \text{ if Hunter's Green neighborhood} \\ 0 \text{ if not} \end{cases}$$

$$x_5 = \begin{cases} 1 \text{ if Hyde Park neighborhood} \\ 0 \text{ if not} \end{cases}$$

$$x_6 = \begin{cases} 1 \text{ if Davis Isles neighborhood} \\ 0 \text{ if not} \end{cases}$$

$$x_7 = \begin{cases} 1 \text{ if Town and Country neighborhood} \\ 0 \text{ if not} \end{cases}$$

$$x_8 = \begin{cases} 1 \text{ if Avila neighborhood} \\ 0 \text{ if not} \end{cases}$$

$$x_9 = \begin{cases} 1 \text{ if Carroll Wood neighborhood} \\ 0 \text{ if not} \end{cases}$$

Model 3. *First-order model, factoring in interactions between each neighborhood and appraised land and improvement values*

$$\begin{aligned} E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 \\ + \beta_{10} x_1 x_3 + \beta_{11} x_1 x_4 + \beta_{12} x_1 x_5 + \beta_{13} x_1 x_6 + \beta_{14} x_1 x_7 + \beta_{15} x_1 x_8 + \beta_{16} x_1 x_9 \\ + \beta_{17} x_2 x_3 + \beta_{18} x_2 x_4 + \beta_{19} x_2 x_5 + \beta_{20} x_2 x_6 + \beta_{21} x_2 x_7 + \beta_{22} x_2 x_8 + \beta_{23} x_2 x_9 \end{aligned}$$

For Model 3, we added interaction variables between the neighborhood dummy variables and x_1 as well as x_2. The interaction terms will allow for change in the sale price y for increase in x_1 or x_2 to vary depending on the neighborhood.

Model Comparisons

Model 1:

Global F-test: Model 1

H_0: $\beta_1 = \beta_2 = 0$

H_a: At least one of the Betas does not equal 0

F-Stat= 3050.68

F-value=<.0001

.0001<.05

Adjusted R^2= 0.9459

Sum of Squares Error: 3192256 ,n=350, k = 2

We reject the null hypothesis (H_0), and conclude that at least one of the slopes for the individual terms in the model does not equal 0.

The F-value (<.0001) is less than the critical value (.05) so we can determine that the model is statistically significant, there is sufficient evidence (at α = .05) to indicate that the model including appraised land value and appraised improvement value contribute information for the prediction of y.

Model 3 (complete model):
Global F-test: Model 3
H_0: $\beta_1=\beta_2=\beta_3=\ldots=\beta_{23}=0$
H_a: At least one of the Betas does not equal 0
F-Stat=703.58
F-value=<.0001
.0001<.05
Adjusted R^2=.9502
Sum of Squares Error: 2756960
n=350, k=23

We reject the null hypothesis (H_0), at least one of the slopes for the individual terms in the model does not equal 0.

The F-value (<.0001) is less than the critical value (.05) so we can determine that the model is statistically significant, there is sufficient evidence (at $\alpha = .05$) to indicate that the model including appraised land value, appraised improvement value, the main effects terms for neighborhood, and the neighborhood interaction terms contribute information for the prediction of y.

Model 2 (reduced model):
Global F-test: Model 2
H_0: $\beta_1= \beta_2= \beta_3=\ldots= \beta_9=0$
H_a: At least one of the Betas does not equal 0
F-Stat=290.81
F-value=<.0001
.0001<.05
Adjusted R^2=.9477
Sum of Squares Error: 3022927 ,n=350, g=9

We reject the null hypothesis (H_0), at least one of the slopes for the individual terms in the model does not equal 0.

The F-value (<.0001) is less than the critical value (.05) so we can determine that the model is statistically significant, there is sufficient evidence (at $\alpha = .05$) to indicate that the model including appraised land value, appraised improvement value, and the main effects terms for neighborhood contribute information for the prediction of y.

Nested F-Test: Models 2 and 3
H_0: $\beta_{10} = \ldots = \beta_{23} = 0$
H_a: At least one of the Betas in the H_0 does not equal to 0

$$\frac{(3022927 - 2756960)/(23-9)}{2756960/(350-23-1)} = 2.2464$$

F-Stat: 2.2464
P-Value: 0.0063650933
P-Value Critical: .05
0.0063650933<.05

We reject the null hypothesis (H_0) and conclude that the additional interaction terms are useful to the model.

The F-value (.0064) is less than the critical value (.05) so we can determine that the model is statistically significant, there is sufficient evidence (at $\alpha = .05$) to indicate that the neighborhood interaction terms of Model 3 contribute information for the prediction of y.

Interpreting the Prediction Equation

The estimates of the Model 3 parameters in the Appendix (*Model 3 output) were substituted into the prediction equation:

$$\hat{y} = -21.76+2.40x_1+1.25x_2+31.91x_3-43.82x_4-86.57x_5-39.18x_6+23.87_7x_7+492.65x_8+62.43x_9$$
$$-1.81x_1x_3-0.81x_1x_4-0.703x_1x_5-0.806x_1x_6+0.63x_1x_7-3.16_5x_1x_8-0.35x_1x_9$$
$$+0.33x_2x_3+0.325x_2x_4+0.122x_2x_5-0.089x_2x_6-0.393x_2x_7+0.327x_2x_8-0.352x_2x_9$$

Cheval (x4,x5,x6,x7,x8,x9=0) (x3=1):

$$\hat{y}= -21.76 + 2.40x_1 + 1.25x_2 + 31.91x_3-1.81x_1x_3+0.33x_2x_3$$

Hunter's Green (x3,x5,x6,x7,x8,x9=0) (x4=1):

$$\hat{y}= -21.76 + 2.40x_1 + 1.25x_2-43.82x_4 -0.81x_1x_4+0.325x_2x_4$$

Hyde Park (x3,x4,x6,x7,x8,x9=0) (x5=1):

$$\hat{y}= -21.76 + 2.40x_1 + 1.25x_2-86.57x_5 -0.703x_1x_5+0.122x_2x_5$$

Davis Isles (x3,x4,x5,x7,x8,x9=0) (x6=1):

$$\hat{y}= -21.76 + 2.40x_1 + 1.25x_2-39.18x_6 -0.806x_1x_6-0.089x_2x_6$$

Town & Country (x3,x4,x5,x6,x8,x9=0) (x7=1):

$$\hat{y}= -21.76 + 2.40x_1 + 1.25x_2 +23.87x_7 +0.63x_1x_7-0.393x_2x_7$$

Avila (x3,x4,x5,x6,x7,x9=0) (x8=1):

$$\hat{y}= -21.76 + 2.40x_1 + 1.25x_2+492.65x_8 -3.16_5x_1x_8+0.327x_2x_8$$

Carrollwood (x3,x4,x5,x6,x7,x8=0) (x9=1):

$$\hat{y}= -21.76 + 2.40x_1 + 1.25x_2+62.43x_9-0.35x_1x_9-0.352x_2x_9$$

Tampa Palms (x3,x4,x5,x6,x7,x8,x9=0):

$$\hat{y}= -21.76 + 2.40x_1 + 1.25x_2$$

The amount of sales price that would increase with appraised values is different for each neighborhood.

Predicting the Sale Price of a Property

Figure CS1.3
Graph of predicted sales for land value of $100,000

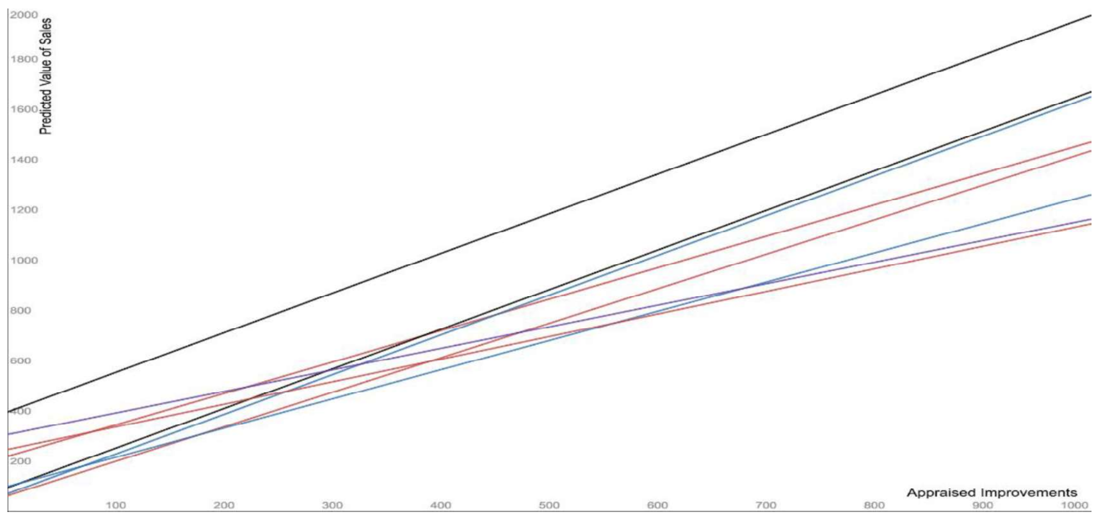


Table CS1.1 Predicted Increase in Sales Price for $1000 Increase in Appraised Improvements

	Neighborhood							
	Tampa Palms	Cheval	Hunter's Green	Hyde Park	Davis Isles	Town& Country	Avila	Carroll Woods
Predicted Increase in Sales Price for $1000 increase in appraised improvements	$2400	$590	$1590	$1697	$1594	$3003	-$760	$2005

Based on Model 3, we acquired that the adjusted R^2 for the model is .9502, which indicates that the model accounts for approximately 95% of the sample variability in the sale price values. This would normally indicate that the model is a good fit to the data. However, since the s = 91.96, it can be interpreted that approximately 95% of the predicted sale price will be within (2s)($1000) = (2*91.96)($1000) = $183,920 of their actual values. This is a large standard deviation that would lead to large errors of prediction for residential properties if the model was actually used.

Conclusions

This report sought to create a prediction equation to estimate the sale price of a home based on its appraised land value, appraised value of improvements, and the neighborhood in which it is located. We are able to adjust the resulting prediction equation to the neighborhood of the property.

In conducting our analysis, we looked at multiple models that vary in complexity. Our first model strictly used the land value and improvement value, regardless of neighborhood, to predict the resulting sale price. This model was effective at capturing the resulting sales price, but failed to incorporate the effect of the neighborhood of the property. As a result, we expanded the model by adding in terms to take this into account. Using statistical tests, we found this model to be more effective at predicting the actual sales price of the properties. That is, including the effect of the location of the property significantly increased the accuracy of the predictions of the model. Taking this a step further, we created a third model that included all the terms from model 3 as well as additional terms that affect the magnitude of the change in sales price for a given increase in either appraised value or improvement value based on which neighborhood the property is in. Running tests on this new model, we again see an improvement in predicting the actual sales price of the properties.

As stated earlier, the full model 3 can be simplified for each individual neighborhood, so each neighborhood effectively has its own prediction equation. This is useful in practice as the full model consists of 24 terms, so the simplified neighborhood versions are much easier to handle.
As suggested by the significance of adding in the additional neighborhood terms, our analysis indicates that the relationships between appraised values and sales price are not linear; rather, the neighborhood in which the property is located influences the effects of the appraised values on sales price. This is important to know for the purpose of our study, as appraisers and home buyers alike may use these results to get a better perspective and properly predict the price of prospective properties.

While the statistical tests indicate our third model captures the majority of the variation in sale price, they also show that the models have a large standard deviation of 92. This number means that we are 95% sure that our estimate for a certain property will be within $92,000 of the model’s output. This implies a total range of $184,000 that we are confident that the true price is in. This wide interval suggests the model may not actually be effective in practice. In the future, a more precise model could be developed using additional variables such as some that are more specific to either the property itself or to the overall housing market. We believe this model would a more accurate predictor of the true sales price of a property.

Appendix

TAMSALES8 (n=350 observations)		
Variable	Type	Description
SALES	Numeric	Sales price (thousands of dollars)
LAND	Numeric	Land value (thousands of dollars)
IMP	Numeric	Value of improvements (thousands of dollars)
NBHD	Character	Neighborhood (HYDEPARK, DAVISISLES, CHEVAL, HUNTERSGREEN, AVILA, CRLLWOODVILL, TAMPAPALMS, TOWN&CNTRY)

*Model 1 Output

The REG Procedure
Model: MODEL1
Dependent Variable: SALES

Number of Observations Read	350
Number of Observations Used	350

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	56130007	28065004	3050.68	<.0001
Error	347	3192256	9199.58432		
Corrected Total	349	59322263			

Root MSE	95.91446	R-Square	0.9462
Dependent Mean	465.15171	Adj R-Sq	0.9459
Coeff Var	20.62004		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.44550	7.92334	-0.81	0.4165
LAND	1	1.33833	0.04399	30.42	<.0001
IMP	1	1.37137	0.02755	49.78	<.0001

*Model 2 Output

The REG Procedure
Model: MODEL1
Dependent Variable: SALES

Number of Observations Read	350
Number of Observations Used	350

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	56299336	6255482	703.58	<.0001
Error	340	3022927	8890.96186		
Corrected Total	349	59322263			

Root MSE	94.29190	R-Square	0.9490
Dependent Mean	465.15171	Adj R-Sq	0.9477
Coeff Var	20.27121		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.44404	13.06627	0.57	0.5692
LAND	1	1.58782	0.07542	21.05	<.0001
IMP	1	1.33777	0.03128	42.77	<.0001
Cheval	1	-32.80474	18.07859	-1.81	0.0705
HUNTERSG	1	-21.73928	16.68278	-1.30	0.1934
Hydepark	1	-80.00018	23.44244	-3.41	0.0007
DAVISISL	1	-115.63119	27.57599	-4.19	<.0001
TOWNCNT	1	-12.59040	17.45186	-0.72	0.4711
Avila	1	-61.14301	34.55276	-1.77	0.0777
CarrollW	1	-18.71894	20.01939	-0.94	0.3504

*Model 3 Output

The REG Procedure
Model: MODEL1
Dependent Variable: SALES

Number of Observations Read	350
Number of Observations Used	350

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	56565303	2459361	290.81	<.0001
Error	326	2756960	8456.93314		
Corrected Total	349	59322263			

Root MSE	91.96159	R-Square	0.9535
Dependent Mean	465.15171	Adj R-Sq	0.9502
Coeff Var	19.77023		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.75618	21.52546	-1.01	0.3129
LAND	1	2.40247	0.47521	5.06	<.0001
IMP	1	1.25040	0.08989	13.91	<.0001
Cheval	1	31.90950	41.85159	0.76	0.4463
HUNTERSG	1	-43.82323	35.49288	-1.23	0.2178
Hydepark	1	-86.56526	66.60744	-1.30	0.1946
DAVISISL	1	-39.17994	37.83979	-1.04	0.3012
TOWNCNT	1	23.87393	95.47336	0.25	0.8027
Avila	1	492.64837	353.47474	1.39	0.1643
CarrollW	1	62.42872	67.25058	0.93	0.3539
landchev	1	-1.81487	0.80473	-2.26	0.0248
landHunt	1	-0.81006	0.84136	-0.96	0.3364
landHyde	1	-0.70318	0.58463	-1.20	0.2299
landDavis	1	-0.80614	0.48170	-1.67	0.0952
landTown	1	0.62972	4.91961	0.13	0.8982
landAvil	1	-3.16261	1.40774	-2.25	0.0253
landCarr	1	-0.35221	1.67099	-0.21	0.8332
impchev	1	0.33076	0.23405	1.41	0.1585
impHunt	1	0.32548	0.16285	2.00	0.0465
impHyde	1	0.12159	0.14022	0.87	0.3865
impDavis	1	-0.08863	0.10466	-0.85	0.3977
impTown	1	-0.39308	0.81657	-0.48	0.6306
impAvil	1	0.32668	0.13418	2.43	0.0154
impCarr	1	-0.35219	0.47408	-0.74	0.4581

SAS Codes Used

```
*loads the TAMSALES8.txt file
data cs1;
infile '/folders/myfolders/TAMSALES8.txt' dlm='09'x firstobs=2;
input FOLIO SALES LNSALES LAND IMP TOTVAL NBHD $;
run;

*Figure CS 1.2
proc sgplot data=cs1;
scatter y=sales x=totval;
yaxis label = "Sales Price";
xaxis label = "Total Appraised Value";
reg x=totval y=sales;
run;
*Model 1
proc reg data=cs1 plots=none;
model sales = land imp;
run;

*Model 1, First order model, assumed that sales price differs depending on the neighborhood
makes TAMPAPALMS to be the base;

data model2;
set cs1;
Cheval = 0;
HUNTERSG = 0;
Hydepark = 0;
DAVISISL = 0;
TOWNCNT = 0;
Avila = 0;
CarrollW = 0;
if NBHD = 'CHEVAL' then Cheval = 1;
if NBHD = 'HUNTERSG' then HUNTERSG = 1;
if NBHD = 'HYDEPARK' then Hydepark = 1;
if NBHD = 'DAVISISL' then DAVISISL = 1;
if NBHD = 'TOWN&CNT' then TOWNCNT = 1;
if NBHD = 'AVILA' then Avila = 1;
if NBHD = 'CARROLLW' then CarrollW = 1;
run;

proc reg data = model2 plots=none;
model sales = land imp Cheval Huntersg Hydepark Davisisl towncnt avila carrollw;
Run;
```

```
*Model 3, with interaction terms;

data model3;
set model2;
landchev = land*Cheval;
landHunt = land*Huntersg;
landHyde = land*Hydepark;
landDavis = land*Davisisl;
landTown=land*Towncnt;
landAvil=land*Avila;
landCarr=land*CarrollW;
impchev=imp*Cheval;
impHunt=imp*Huntersg;
impHyde=imp*Hydepark;
impDavis=imp*Davisisl;
impTown=imp*Towncnt;
impAvil=imp*Avila;
impCarr=imp*CarrollW;
run;

proc reg data = model3 plots=none;
model sales = land imp Cheval Huntersg Hydepark Davisisl towncnt avila carrollw
landchev landHunt landHyde landDavis landTown landAvil landCarr impChev impHunt
impHyde impDavis impTown impAvil impCarr;
run;

*Nested F test for models 2 and 3
data nestedFtest;
Fstat = (2.2464);
pvalue = SDF('F',Fstat,23 - 9 ,350 - 23 - 1);
proc print data=nestedFtest;
Run;
```