Analysis of Sale Prices

Timothy Kim, Max Dodge, Teja Ganti, Graham Gilliam

Introduction

The real estate market is one of the most consequential sectors of the US economy and has far reaching consequences not only domestically but also abroad as the 2008 financial crisis has taught the international community. Being able to predict home prices can also be of use to homeowners and real estate agents if a simple and robust problem can be found that gives a ballpark estimate of the final sale price. Many variables can explain the final sale price of a house with some of the most obvious being considered in this report. Value comes from the land and buildings, but also as the old adage goes: location, location, location. These are the three primary elements upon which our model will be based. Appraisal data can provide relatively good estimates for the first two aspects, and the value of location can be determined empirically. If the mean value of the property can be predicted that could be considered the fair market price. The objectives of this study are as follows:

1. Determine that a relationship exists between land value, improvement value, neighborhood, and sale price. Does the data provide evidence that this information contributes to the sale price?
2. Develop a model and prediction equation relating the variables of land value, improvement value, and neighborhood to the sale price of a house. Determine the accuracy and effectiveness of this model for different neighborhoods: does the appraisal criteria differ between neighborhoods?

Data Summary

The property appraisal office of Hillsborough County, Florida provided the data for use in this study. The independent variables in this data set relevant to our study are as follows:

1. Land Value [Land]: The appraised value of solely the land of the property in thousands of dollars
2. Improvement Value [Imp]: The appraised value of the buildings and other structures on the property in thousands of dollars
3. Neighborhood [NBHD]: A categorical variable consisting of eight levels representing neighborhoods that are relatively internally homogeneous but in property types and value as well as possessing some socioeconomic differences. The levels are found below:
    a. Hyde Park, Cheval, Hunter's Green, Davis Isles, Avila, Carrollwood, Tampa Palms, Town & Country

Table 1.1

Variable	N	Mean	Median	Standard Deviation	Minimum	Maximum
Sale Price	350	465.151	328.450	412.286	59.100	3200.00
Land Value	350	115.840	59.340	131.572	16.560	1004.59
Improvement Value	350	230.838	164.370	210.071	31.930	1714.98
Total Value	350	346.678	260.895	294.885	55.770	2134.23

Hypothesized Models

If the appraisals were completely accurate the sale price of a house could be perfectly predicted by adding the Land and Improvement values together. However, it can be seen that the total value does not explain all of the variance in sales price. Appraisals can be inaccurate or out of date and usually reflect the opinion of tax assessors.

Theoretical Models

Where: x_1 = Land Value x_2 = Improvement Value

$x_3 = 1$ if Hunter's Green; 0 otherwise $x_4 = 1$ if Hyde Park; 0 otherwise

$x_5 = 1$ if Davis Isles; 0 otherwise $x_6 = 1$ if Town & Country; 0 otherwise

$x_7 = 1$ if Avila; 0 otherwise $x_8 = 1$ if Carrollwood; 0 otherwise $x_9 = 1$ if Tampa Palms; 0 otherwise

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

The model above assumes constant differences between neighborhoods and that the land value and improvement value does not depend on neighborhood.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_5 + \beta_7 x_6 + \beta_8 x_7 + \beta_9 x_8 + \beta_{10} x_9$$
$$+ \beta_{11} x_1 x_3 + \beta_{12} x_1 x_4 + \beta_{13} x_1 x_5 + \beta_{14} x_1 x_6 + \beta_{15} x_1 x_7 + \beta_{16} x_1 x_8 + \beta_{17} x_1 x_9$$
$$+ \beta_{18} x_2 x_3 + \beta_{19} x_2 x_4 + \beta_{20} x_2 x_5 + \beta_{21} x_2 x_6 + \beta_{22} x_2 x_7 + \beta_{24} x_2 x_8 + \beta_{25} x_2 x_9$$
$$+ \beta_{26} x_3 x_1 x_2 + \beta_{27} x_4 x_1 x_2 + \beta_{28} x_5 x_1 x_2 + \beta_{29} x_6 x_1 x_2 + \beta_{30} x_7 x_1 x_2 + \beta_{31} x_8 x_1 x_2 + \beta_{32} x_9 x_1 x_2$$

The model above assumes that changes in y due to changes in x_1 or x_2 to vary depending on the neighborhood. Furthermore, it allows for changes in y due to x_1 to depend on x_2 and vice versa. Due to neighborhood interaction terms the change in y will vary between neighborhoods.

Analysis

In order to better understand the two models ability to explain the data, the two models were first compared by MSE, R^2_{adj} and s. The output was placed in Table 1.2.

Table 1.2

Model	MSE	$R^{2\ adj}$	s
Model 1	8890.96	0.9477	94.2919
Model 2	7627.38	0.9551	87.33491

Model 2 appears to have both a smaller MSE (87 < 94) and a higher R^2_{adj} (0.95>0.94). This suggests that Model 2 has significant variables that Model 1 lacks. Looking at the two models, the difference between the two lies in the interaction terms. In order to test the significance of the interaction terms contained in Model 2, a partial F-test was run in SAS, with a null hypothesis that the interaction terms offer no significance, in other words that the interaction terms' parameters are equal to zero: H_0: $\beta_3 = \beta_{11} = \beta_{12} = \ldots = \beta_{32} = 0$. The reduced model was Model 1, with the complete model as Model 2. The test statistic to used by SAS is F = ((SSE_R - SSE_C)/(number of β parameters in H_O)/ MSE_C. The output of this partial-F Test was put into Table 1.3. The F-Value ends up being 3.56 with 22 degrees of freedom in the numerator and 318 degrees of freedom in the denominator. The corresponding p-value, or the probability of finding a F-Value higher than that under the null hypothesis, was <.0001.

Table 1.3

Source	DF	Mean Square	F Value	Pr > F
Numerator	22	27155	3.56	<.0001
Denominator	318	7627.38664		

At an alpha = 0.05, our p-value of <.0001 suggests that there is evidence to indicate that the addition of the interaction terms between the neighborhood and land value and/or improvement value and the interaction between land value variables contribute significantly to the prediction of Sales Price. Next, the model was actually run in SAS with confidence intervals for the parameter estimates and displayed in Table 1.4

Table 1.4

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	155.24943	92.25047	1.68	0.0934	-26.24893	336.74779
LAND	1	-0.82722	1.03879	-0.80	0.4264	-2.87099	1.21654
IMP	1	0.96092	0.41999	2.29	0.0228	0.13460	1.78723
huntval	1	-132.46668	111.72040	-1.19	0.2366	-352.27119	87.33783
hydeval	1	-265.90544	123.52900	-2.15	0.0321	-508.94282	-22.86807
davisval	1	-60.17087	101.65499	-0.59	0.5543	-260.17217	139.83043
townval	1	-385.93772	438.28678	-0.88	0.3792	-1248.24589	476.37046
avilaval	1	-1610.50458	836.81128	-1.92	0.0552	-3256.89055	35.88139
carolval	1	-31.41548	202.98198	-0.15	0.8771	-430.77278	367.94182
tampaval	1	-96.94217	100.34251	-0.97	0.3347	-294.36123	100.47689
landimp	1	0.00518	0.00306	1.69	0.0915	-0.00084040	0.01119
landhunt	1	0.93615	1.55909	0.60	0.5486	-2.13129	4.00358
landhyde	1	2.53421	1.10364	2.30	0.0223	0.36285	4.70556
landdavis	1	2.01231	1.04464	1.93	0.0550	-0.04297	4.06759
landtown	1	15.06094	20.72807	0.73	0.4680	-25.72054	55.84242
landavila	1	6.10553	2.89100	2.11	0.0355	0.41763	11.79343
landcarr	1	1.42362	3.50235	0.41	0.6847	-5.46708	8.31432
landtamp	1	2.08939	1.23049	1.70	0.0905	-0.33154	4.51033
imphunt	1	0.25249	0.49776	0.51	0.6123	-0.72682	1.23181
imphyde	1	0.41920	0.47457	0.88	0.3777	-0.51450	1.35290
impdavis	1	-0.19767	0.43033	-0.46	0.6463	-1.04433	0.64899
imptown	1	3.11688	5.86587	0.53	0.5955	-8.42393	14.65769
impavila	1	3.42264	1.18977	2.88	0.0043	1.08182	5.76345
impcarr	1	-0.59328	1.24639	-0.48	0.6344	-3.04549	1.85893
imptamp	1	-0.02108	0.44816	-0.05	0.9625	-0.90281	0.86064
limphunt	1	-0.00024653	0.00441	-0.06	0.9555	-0.00893	0.00844
limphyde	1	-0.00520	0.00310	-1.68	0.0948	-0.01130	0.00090472
limpdavis	1	-0.00428	0.00306	-1.40	0.1635	-0.01030	0.00175
limptown	1	-0.15820	0.27560	-0.57	0.5664	-0.70042	0.38403
limpavila	1	-0.01378	0.00457	-3.01	0.0028	-0.02277	-0.00478
limpcarr	1	0.00369	0.01841	0.20	0.8414	-0.03254	0.03991
limptamp	1	-0.00188	0.00336	-0.56	0.5751	-0.00849	0.00472

Even with some variables appearing to have non-significant values, we still include them in the model as using T-tests isn't effective on models with such a high degree of interaction- there is likely multicollinearity or other interactions and not including some points would mess up the model's effectiveness. The large confidence intervals are indicative of the fairly large root MSE (87) from Table 1.2.

Results and Conclusion

Interpreting the Prediction Equation

Substituting the parameter point estimates into the prediction equation for Model 2 yields:

$$E(y) = 155.24 - 0.82722x_1 + 0.96092x_2 + 0.00518x_1x_2 - 132.46668x_3 - 265.90544x_4 - 60.17087x_5 - 385.93772x_6 - 1610.50458x_7 - 31.41548x_8 - 96.94217x_9 + 0.93615x_1x_3 + 2.53421x_1x_4 + 2.01231x_1x_5 + 15.06094x_1x_6 + 6.10553x_1x_7 + 1.42362x_1x_8 + 2.08939x_1x_9 + 0.25249x_2x_3 - 0.41920x_2x_4 - 0.19767x_2x_5 + 3.11688x_2x_6 + 3.42264x_2x_7 - 0.59328x_2x_8 - 0.02108x_2x_9 - 0.00024x_3x_1x_2 - 0.00520x_4x_1x_2 - 0.00428x_5x_1x_2 - 0.15820x_6x_1x_2 - 0.01378x_7x_1x_2 + 0.00369x_8x_1x_2 - 0.00188x_9x_1x_2 .$$

The prediction equation for Model 2 can be simplified in context of each of the eight neighborhoods as essentially eight different neighborhood-unique equations. These simplified equations were solved by setting each of the coded categorical variables equal to 0 or 1 depending on the neighborhood, and then substituted into Table 1.5.

Table 1.5

Neighborhood	Prediction Equation
Cheval	$\hat{y} = 155.24 - 0.82722x_1 + 0.96092x_2 + 0.00518x_1x_2$
Hunter's Green	$\hat{y} = 22.7733 + 0.1089x_1 + 1.2134x_2 + 0.00278x_1x_2$
Hyde Park	$\hat{y} = -110.665 + 1.7069x_1 + 0.5419 - 0.00002x_1x_2$
Davis Isles	$\hat{y} = 95.069 + 1.185x_1 + 0.7632x_2 + 0.00009x_1x_2$
Town & Country	$\hat{y} = -230.697 + 14.2337x_1 + 4.077x_2 - 0.153x_1x_2$
Avila	$\hat{y} = -1455.26 + 5.278x_1 + 4.3835x_2 -0.0086x_1x_2$
Carrollwood	$\hat{y} = 123.824 + 0.596x_1 + 0.367x_2 + 0.00887x_1x_2$
Tampa Palms	$\hat{y} = 58.298 + 1.262x_1 + 0.939x_2 + 0.0033x_1x_2$

The intercept for each prediction equation can be interpreted as the expected sale price for a home with $0 land value and $0 improvement- in context this doesn't have much significant meaning as that is a very rare circumstance. To interpret the β estimates of each interaction equation, we take one independent variable as given, say land value, and focus on the resultant slope for improvement value. For example, if land value=50 ($50 thousand) for a Cheval home, a $1,000 increase in the appraised improvement value increases the average sale price by (50*.00518)+.96092=1.219, or $1,219. The prediction equations then give information about which neighborhoods are most impacted by the variables: for instance, the Town & Country neighborhood experiences large increases in average sale price for every $1000 increase in appraised land value, while Carrollwood has marginal increases. This results in some neighborhoods being under/over appreciated based on their appraisals.

Predicting the Sale Price of a Property

From Table 1.2, we found R^2_{adj} to be 0.9551. This indicates that Model 2 accounted for about 95% of all of the variability in the individual samples of the sale price value, y. This strongly indicates that the model is a good predictor for the data. However, the large root MSE, s= 87.33, indicates that individual point prediction will be varied. We would expect that about 95% of our predicted price values would fall fall within 2s= 174.66= $174,660 of the actual value. This shows that while Model 2 might be an accurate representation of the data, it wouldn't be effective for a realtor attempting to actually predict the sale price of individual properties, only the mean sale price of multiple similarly valued properties.

Conclusion

The models created from the data suggest that each of the eight neighborhoods have different relationships between property sale prices and appraised land values. Some neighborhoods are more impacted by the modeled variables than other neighborhoods, indicating that the appraisal criteria differ between neighborhoods. This discrepancy reveals certain areas that require improvement or adjustment.

Code Appendix

nebhood * PROC REG running

```
data nbhd;
infile 'C:\Users\Graham\Documents\uva\SAS\TAMSALES8.txt' dlm='09'X firstobs=2;
input FOLIO SALES LNSALES LAND IMP TOTVAL NBHD $;
run;
*loads in the initial data;
data nbhd1;
set nbhd;
cheval= 0;
huntval= 0;
hydeval = 0;
davisval= 0;
townval = 0;
avilaval= 0;
carolval = 0;
tampaval = 0;
if NBHD = 'CHEVAL' then cheval = 1;
if NBHD = 'HUNTERSG' then huntval = 1;
if NBHD = 'HYDEPARK' then hydeval = 1;
if NBHD = 'DAVISISL' then davisval = 1;
if NBHD = 'TOWN&CNT' then townval = 1;
if NBHD = 'AVILA' then avilaval = 1;
if NBHD = 'CARROLLW' then carolval = 1;
if NBHD = 'TAMPAPAL' then tampaval = 1;
run;
*codes the qualitative variables;
```

```
data nbhd2;
set nbhd1;
landimp= LAND*IMP;
landchev = LAND*cheval;
landhunt= LAND*huntval;
landhyde = LAND*hydeval;
landdavis = LAND*davisval;
landtown = LAND*townval;
landavila = LAND*avilaval;
landcarr= LAND*carolval;
landtamp= LAND*tampaval;
impchev = IMP*cheval;
imphunt= IMP*huntval;
imphyde = IMP*hydeval;
impdavis = IMP*davisval;
imptown = IMP*townval;
impavila = IMP*avilaval;
impcarr= IMP*carolval;
imptamp= IMP*tampaval;
limpchev = IMP*LAND*cheval;
limphunt= LAND*IMP*huntval;
limphyde = LAND*IMP*hydeval;
limpdavis = LAND*IMP*davisval;
limptown = LAND*IMP*townval;
limpavila = LAND*IMP*avilaval;
limpcarr= LAND*IMP*carolval;
limptamp= LAND*IMP*tampaval;
run;
*creates all the necessary interaction terms for the models;


*model 1;
proc reg data=nbhd2 plots=none;
model SALES = LAND IMP huntval hydeval davisval townval avilaval carolval tampaval;
run;

*model 2;
proc reg data=nbhd2 plots=none;
model SALES = LAND IMP huntval hydeval
davisval townval avilaval carolval tampaval
landimp landhunt landhyde landdavis landtown landavila
landcarr landtamp imphunt imphyde impdavis imptown impavila
impcarr imptamp limphunt limphyde limpdavis limptown limpavila
limpcarr limptamp;
run;


*Partial-F Test;
proc reg data=nbhd2 plots= none;
model SALES = LAND IMP huntval hydeval davisval townval
avilaval carolval tampaval landimp landhunt landhyde landdavis
landtown landavila landcarr landtamp imphunt imphyde impdavis
imptown impavila impcarr imptamp limphunt limphyde limpdavis
limptown limpavila limpcarr limptamp;
NHBD: test landimp, landhunt, landhyde, landdavis, landtown,
landavila, landcarr, landtamp, imphunt, imphyde, impdavis,
imptown, impavila, impcarr, imptamp, limphunt, limphyde,
limpdavis, limptown, limpavila, limpcarr, limptamp;
run;


*prints confidence intervals for parameters, and predictions
and confidence and prediction intervals for the sale value;
```

```
*prints confidence intervals for parameters, and predictions
and confidence and prediction intervals for the sale value;
proc reg data=nbhd2 plots=none;
model SALES = LAND IMP huntval hydeval
davisval townval avilaval carolval tampaval
landimp landhunt landhyde landdavis landtown landavila
landcarr landtamp imphunt imphyde impdavis imptown impavila
impcarr imptamp limphunt limphyde limpdavis limptown limpavila
limpcarr limptamp / p clb cli clm;
run;
```