Effect of Taxes, Bedrooms, Baths, Quadrants, Size, and Lot Size on Price of Homes in Gainesville, Florida

Brooke Dippold, Lindsey Gordon, and Hannah Ramey

Introduction

In this report, we will create a multiple linear regression model to predict sale price for homes in the Gainesville, Florida area. The population in this model is all homes and condos in Gainesville, Florida. The statistical model is analyzing the sample of 100 recent homes sales in Gainesville, Florida. The variables we are using include: taxes, bedrooms, baths, quadrant (NW, NE, SW, SE), size, lot size, and price. The quantitative explanatory variables include: taxes, bedrooms, baths, size, and lot size. The only categorical explanatory variable is quadrant. The remaining variable, price, is a quantitative response variable. Our results show that taxes, size and lot size are significant predictors of sale price for homes in the Gainesville, Florida area.

Correlation Among Variables

First we examine the correlations between variables in Table 1 and Figure 1. The table shows the the variables price, size, and lot size have a strong correlation with taxes. We also note that the variables size and lot size are strongly correlated with price.

Initial Regression Model

The statistical model we will use is:

$$\text{PRICE} = \beta_0 + \beta_{\text{TAXES}} * \text{TAXES} + \beta_{\text{BEDROOMS}} * \text{BEDROOMS} + \beta_{\text{BATHS}} * \text{BATHS} + \beta_{\text{SIZES}} * \text{SIZES} + \beta_{\text{LOT SIZE}} * \text{LOTSIZE} + \varepsilon$$

where ε is the residual error and is assumed to be normally distributed with mean zero, standard deviation σ_ε, and is independent of all of the explanatory variables.

The hypotheses we will be testing for this model are:

$$H_0 : \beta_{\text{TAXES}} = \beta_{\text{BEDROOMS}} = \beta_{\text{BATHS}} = \beta_{\text{SIZES}} = \beta_{\text{LOT SIZE}} = 0$$

H_a: At least one slope (i.e. coefficient) is nonzero.

Running a multiple regression, we have the following prediction equation using all variables (see Table 4 in the Appendix):

$$\widehat{PRICE} = 6633.7997 + 20.643631 * TAXES - 6469.686 * BEDROOMS + 11824.488 * BATHS + 33.571428 * SIZE + 1.6162385 * LOTSIZE$$

This model is significant (F(5,94) = 61.5235, p < 0.0001) meaning at least one of the coefficients in the equation is nonzero (see Table 3 in Appendix). With $R^2 = 0.765946$ we note that 76.6% of the variation in price is explained by the linear model using all variables (see Table 2 in the Appendix).

Model Reduction

The variables bedrooms (p = 0.2264) and baths (p = 0.1096) were not significant in the initial model. Since bedrooms and baths are not significant we removed them to get the model:

$$PRICE = \beta_0 + \beta_{TAXES} * TAXES + \beta_{SIZE} * SIZE + \beta_{LOTSIZE} * LOTSIZE + \varepsilon$$

where ε is the residual error and is assumed to be normally distributed with mean zero, standard deviation σ_ε, and is independent of all of the explanatory variables.

Running a multiple regression, we have the following prediction equation for this reduced model variables (Table 7 in Appendix):

$$\widehat{PRICE} = 6305.3193 + 22.035493 * TAXES + 34.511792 * SIZE + 1.5943618 * LOTSIZE$$

This model is significant (F(3, 96) = 99.6485, p < 0.0001) meaning at least one of the coefficients in the equation is nonzero (Table 6 in Appendix). With an $R^2 = 0.756928$ we note that 75.69% of the variation is price is explained by the linear model using all variables (Table 5 in the Appendix). We note also that all of the variables in this model are significant (Table 7 in Appendix).

To determine if there is a significant reduction in R^2 from the full to the reduced model, we will run the nested *F* test. Our test statistic is given by:

$$F(2, 96) = \frac{(SSM_{Full} - SSM_{Reduced}) / (dFM_{Full} - dFM_{Reduced})}{MSE_{Full}}$$

$$F(2, 96) = \frac{(2.4084e^{11} - 2.38e^{11}) / (5 - 3)}{782,915,614} = 1.81373$$

with a p-value of 0.168587 (Fcdf(1.8373, 1E99, 2, 96)). Thus there is no significant reduction in R^2 and all of our variables have significant non-zero coefficients, we will use this model as our final model.

Model Verification

For this model, we also note that the estimate of the common standard deviation, σ_ε is $\sqrt{MSE} = \sqrt{796141802} = 28216$. We do not believe the standard error condition is satisfied to use this model because two times the smallest standard deviation is not greater than the largest standard deviation shown here: $0.491127 * 2 < 7.543277$ (Table 7 in Appendix).

Finally we examine the residuals versus the predicted values (Figure 2 in the Appendix) and the residuals versus the explanatory variables in the reduced model (Figure 3 in the Appendix). The residuals versus the predicted plot shows no discernable pattern as does the first row of Figure 3 which shows the residuals versus each of the explanatory variables. Also in Figure 4 in the Appendix there is no strong indication of non-normality of the distribution of the residuals.

Using the correlations listed in Table 1, variables were removed based off of little to no strong correlation with other variables. Therefore, bedrooms and bath were removed from the model since those variables were not strongly correlated with any other variable in the model.

Using Model

Finally, we use the reduced model to predict the price for House 1:

$$\widehat{PRICE} = 6,305.3193 + 22.035493 * 1,360 + 34.511792 * 1,240 + 1.5943618 * 18,000 = \$107,767.00$$

Since the observed value of price for House 1 was \$145,000, we see that this model underpredicts the price for House 1 by \$37,233.

Keeping all other variables constant, this model indicates a change in price of 40,817.11 dollars when the home size increases by 1,000 square feet as shown below:

$$\widehat{PRICE} = 6305.3193 + 22.035493 * 0 + 34.511792 * 1,000 + 1.5943618 * 0 = \$40,817.11$$

Keeping all other variables constant, this model indicates a change in price of 7,899 dollars when lot size increases by 1,000 square feet as shown below:

$$\widehat{PRICE} = 6305.3193 + 22.035493 * 0 + 34.511792 * 0 + 1.5943618 * 1,000 = \$7,899.68$$

Therefore buying home with 1,000 more square feet in lot size will be more "bang for the buck" because a homeowner is getting the same increase space for a smaller increase in price.

Bonus

$$\widehat{PRICE} = -4853.788 + 20.715758 * TAXES + 38.35445 * SIZE + 1.3990594 * LOTSIZE + 15050.436 * NW$$

With NW being 1 and not NW being 0, these are the predicted equations:

$$\widehat{PRICE}_{NW} = 10196.648 + 20.715758 * TAXES + 38.35445 * SIZE + 1.3990594 * LOTSIZE$$
$$\widehat{PRICE}_{NOT\ NW} = -4853.788 + 20.715758 * TAXES + 38.35445 * SIZE + 1.3990594 * LOTSIZE$$

The predicted price for NW homes is $15,050.44 more than homes not in the NW holding all other variables constant.

APPENDIX

Table 1. Table of Correlations

	Price	Taxes	Bedrooms	Baths	Size	Lot Size
Price	1.0000	0.8238	0.3634	0.5712	0.7613	0.7138
Taxes	0.8238	1.0000	0.3988	0.5505	0.7379	0.7355
Bedrooms	0.3634	0.3988	1.0000	0.4578	0.5733	0.2120
Baths	0.5712	0.5505	0.4578	1.0000	0.6408	0.3333
Size	0.7613	0.7379	0.5733	0.6408	1.0000	0.5345
Lot Size	0.7138	0.7355	0.2120	0.3333	0.5345	1.0000

Figure 1. Correlation Scatterplot Matrix

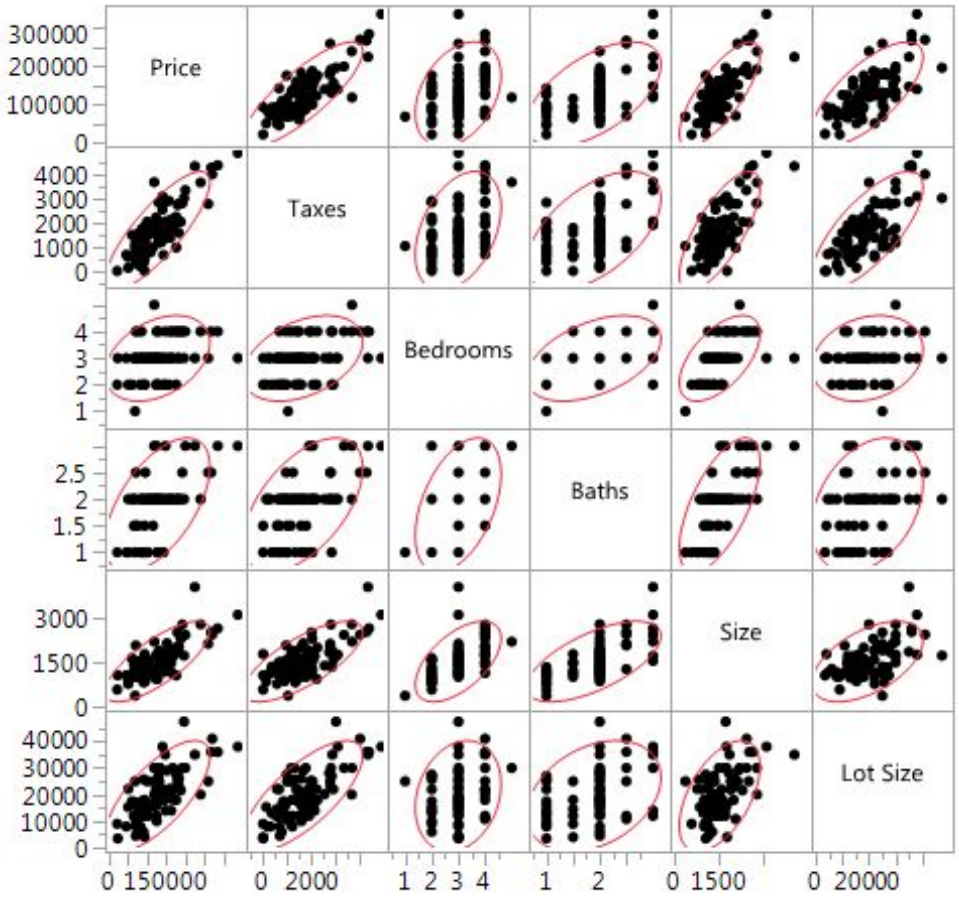

Table 2. Summary of Fit Full Model

RSquare	0.765946
RSquare Adj	0.753497
Root Mean Square Error	27980.63
Mean of Response	126698
Observations (or Sum Wgts)	100

Table 3. Analysis of Variance Full Model

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	2.4084e+11	4.817e+10	61.5235
Error	94	7.3594e+10	782915614	Prob > F
C. Total	99	3.1443e+11		<.0001*

Table 4. Parameter Estimates Full Model

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6633.7997	15834.62	0.42	0.6762
Taxes	20.643631	5.255795	3.93	0.0002*
Bedrooms	-6469.686	5313.155	-1.22	0.2264
Baths	11824.488	7320.944	1.62	0.1096
Size	33.571428	8.890447	3.78	0.0003*
Lot Size	1.6162385	0.494841	3.27	0.0015*

Table 5. Summary of Fit Reduced Model

RSquare	0.756928
RSquare Adj	0.749332
Root Mean Square Error	28215.98
Mean of Response	126698
Observations (or Sum Wgts)	100

Table 6. Analysis of Variance Reduced Model

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	2.38e+11	7.933e+10	99.6485
Error	96	7.643e+10	796141802	Prob > F
C. Total	99	3.1443e+11		<.0001*

Table 7. Parameter Estimates Reduced Model

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6305.3193	9567.273	0.66	0.5114
Taxes	22.035493	5.194517	4.24	<.0001*
Size	34.511792	7.543277	4.58	<.0001*
Lot Size	1.5943618	0.491127	3.25	0.0016*

Figure 2. Residual by Predicted Plot, Reduced Model

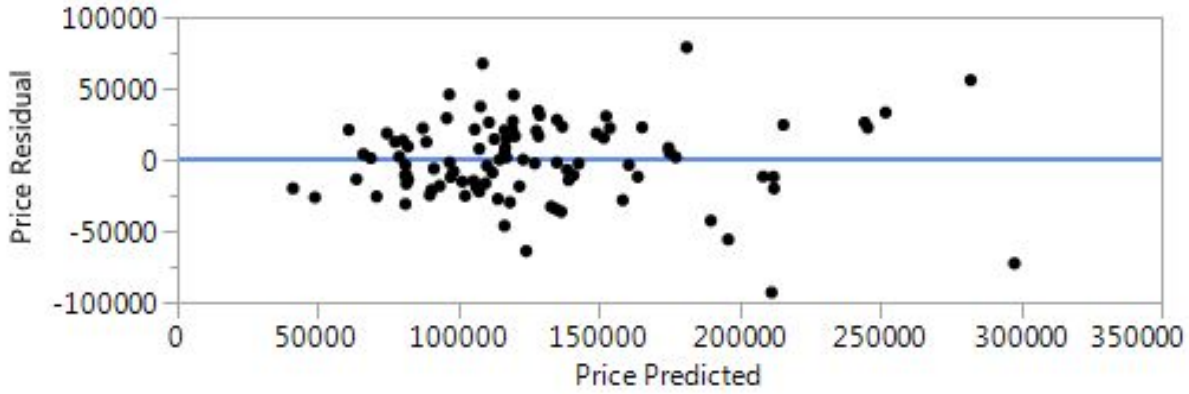

Figure 3. Scatterplot Matrix of Residuals for Reduced Model Versus Explanatory Variables

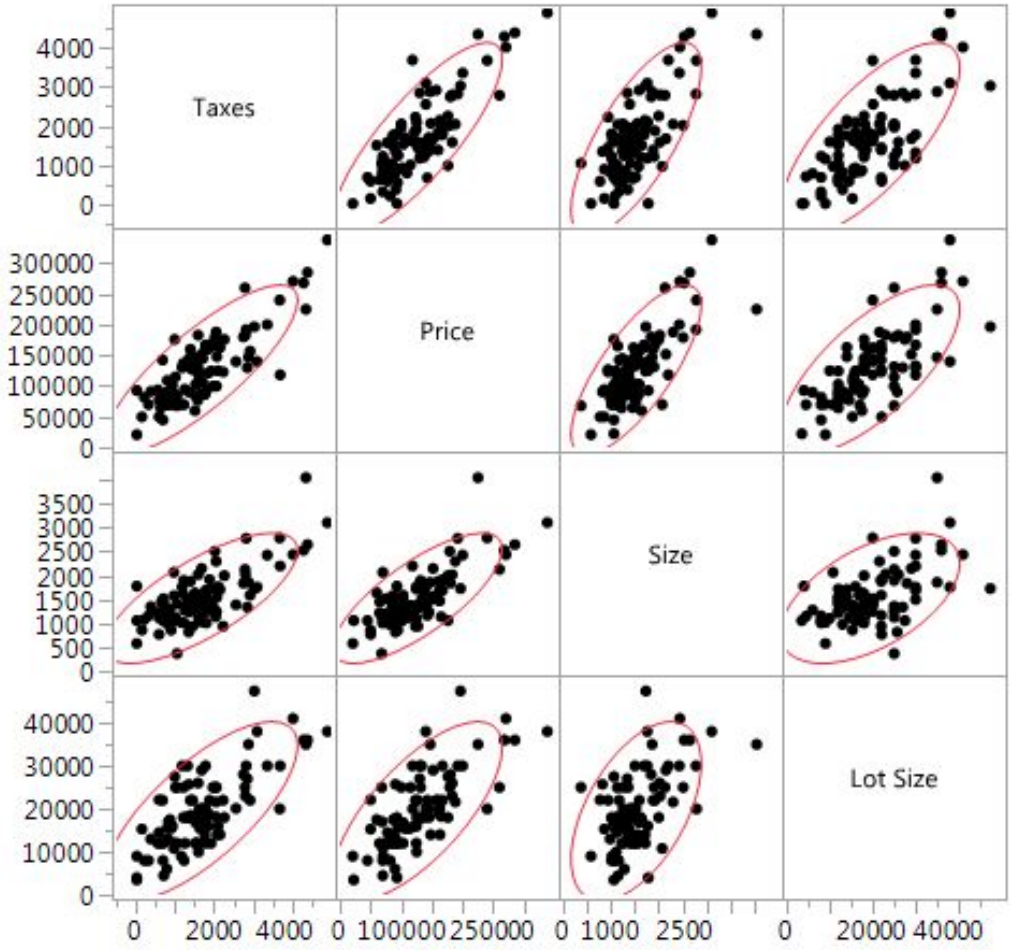

Figure 4. Distribution of Residuals, Reduced Model