GSBA 545 Regression Assignment, Fall 2021

GSBA 545 Group Regression Assignment

Simple and Multiple Linear Regression Analysis: South Bay Real Estate

This group assignment, due by Friday, December 10th, by 11:59pm Pacific Time on Blackboard, is intended to have you practice analyzing multiple regression analysis results. You are expressly forbidden from discussing anything regarding this assignment with anyone outside your team.

The data represent several properties recently listed for sale in the South Bay area not far from Los Angeles during the same time period. have included a few tables in Appendix A and regression output results in Appendix B at the end of this document, but the data can also be found on the course Blackboard site (*South Bay Real Estate.xlsx.*) The variables are:

List Price: Price ($USD) the property is currently listed for

*Ln*(List Price): Natural log transformation of the List Price variable

Home Type1: Single Family Residential, Townhouse, Condo/Coop, or Multi-Family City2: Hermosa Beach, Manhattan Beach, or Redondo Beach

Beds: Number of bedrooms

Baths: Number of bathrooms

Sqft: Square footage of the living space

Lot Size: Square footage of the lot the property is on

Year Built: Year the structure was built

Parking Spots: Number of parking spots

Days on Mkt: Number of days the property has been listed for sale

Your goal here is to predict, as accurately as possible, the listing price of properties in the South Bay.

\*\*Please note that Models 2 – 4 were created after taking the *natural log* of List Price. Answer the questions in relation to *ln*(List Price) unless I give you instructions about how to “back out” of the transformation.\*\*

1 This variable is transformed into 4 separate dummy variables, one for each property type. Each one is given a “1” if a place IS that type and “0” otherwise.

2 This variable is transformed into 3 separate dummy variables, one for each city. Each one is given a “1” if a place IS in that city and “0” otherwise.

1

GSBA 545 Regression Assignment, Fall 2021

**Descriptive Statistics Questions:**

1. List 3 insights derived from the Descriptive Statistics and Correlation Matrix in Appendix A.

* Property in the South Bay has an average listing price of $1,526,136.59 with an average number of bed/baths of 3.482/2.954, respectively
* The number of beds and square footage as well as number of beds and number of baths and of a property are positively correlated, with a correlation of 0.7164 and 0.7788, respectively.
* Square footage and the number of baths has the highest correlation of 0.8362. This is stronger than their correlation with the list price suggesting that there is potential multicollinearity.
* List price has the highest correlation coefficient with SQFT and BATHS meaning these 2 variables have the highest association with the list price

2. Which variable, on its own, would be the strongest predictor of List Price? Why?

* Square footage, on its own, would be the strongest predictor of List Price as it has the highest correlation with List Price, with a correlation coefficient of 0.5576.

**Model 1 Questions:**

3. Why have I excluded *Townhouse3* and *Redondo Beach4* from the first model?

* Because each property type is a dummy variable, and the scenario when Condo/Coop, Multi-Family and Single Family Residential are all 0, is Townhouse. This also applies to the cities. Each city is a dummy variable, and the scenario when Hermosa Beach and Manhattan Beach are both 0, is Redondo Beach.
* If a home is not Single Family Residential, Multi-Family, or a Condo/Coop, then it must be a Townhouse. Similarly, if a home is not in Hermosa or Manhattan Beach, then it must be in Redondo. Removing these parameters from the model helps to create a more accurate model with less noise.

4. What 3 items do you think need to be addressed in Model 1, and in what order?

* The F Ratio is high with a value of 19.3625 and the probability > F is less than 0.0001, so the model is useful overall, however, we likely do not need all the variables.
* There are no VIF values that exceed 10, the highest VIF value is 5.3679 for Baths, which is well below the threshold for severe multicollinearity.
* Because there are no VIF values > 10, we can remove variables with high p-values one at a time.
* We would remove the variables below, in order:
  + Single Family Residential: p-value = 0.9413
  + Multi-Family (2-4 Unit): p-value = 0.9133
  + Lot Size: p-value = 0.4538
* The residual vs fitted plot is heteroscedastic: looking at the scatter plot in Model 1, the plots spread as we move along with List Price. Try taking a natural log of list price and re-run regression.

5. Is there a statistically significant relationship between List Price and the 12 independent variables in the model? Justify your answer.

* To assess the relationship between List Price and all 12 independent variables, we will look at the Analysis of Variance section of the JMP output. The probability > F is less than 0.0001, which means there is a significant relationship between List Price and the 12 independent variables.
* The model does, however, show heteroscedasticity so the confidence interval will be too large for small list prices and too small for larger list prices. Without correcting for this, outliers will be impacted.

**Model 2 Questions:**

6. Is there a significantly different *ln*(List Price) predicted for properties that are considered to be *Multi-Family* versus *Townhouses?* Justify your answer.

* In the regression model, the probability>|t| for the multi-family variable is 0.36 which is greater than 0.05, therefore, failing to reject the null, the ln(List Price) is not statistically different for Multi-Family versus Townhouse homes.

7. What percent of the variability in *ln*(List Price) is able to be explained by the 12 independent variables?

* The R-square value for Model 2 is 0.76. Therefore, the 12 independent variables explain for 76% of the variability in ln(List Price)

**Model 3 Questions**: For this model, the City variables have been removed.

8. Does it appear that City is significantly related to *ln*(List Price)? Why?

* Yes, it appears that city is significantly related to ln(List Price) as both the dummy variables for City in model 2 had a p-value of less than 0.05. Also, the R-square value for Model 3 is lower than Model 2 and the Sum of Squares error is higher for model 3

9. Does it appear that Home Type is significantly related to *ln*(List Price)? Why?

* When home type is converted to dummy variables, it appears that if the home type is a single family residential or multi-family then the home type is significantly related to ln(List Price). However, since the p-value for condo home type is greater than 0.05, if the house is a condo, it does not significantly impact the ln(List Price)

**Model 4 Questions:**

10. Is there a statistically significant relationship between *ln*(List Price) and the 7 remaining independent variables in the model? Justify your answer.

* Yes, there is a significant relationship between the ln(List Price) and the 7 remaining variables since the Prob > F is less than 0.05.

11. What is the expected difference in *ln*(List Price) between a property in *Hermosa Beach* and one in *Redondo Beach5?*

* The expected difference between the ln(List Price) between a property in Hermosa Beach and one in Redondo Beach is 0.3598

12. What is the expected difference in *ln*(List Price) between a property in *Manhattan Beach* and one in *Redondo Beach6?*

* The expected difference between the ln(List Price) between a property in Manhattan Beach and one in Redondo Beach is 0.5183 while controlling for all the other variables

13. How does the *age* of a property related to *ln*(List Price)?

* For every additional year that the property ages, the ln(List Price) decreases by 0.0024

14. In the dataset, there is a *Single Family Residential* property in *Redondo Beach* built in 2006 with 3525 SQFT that was on the market for 391 days. What *ln*(List Price) would you predict for it?

* ln(List Price) = 17.4766 + 0.1734(1) + 0.0004(3525) - 0.0024(2006) + 0.0009(391)
* ln(List Price) = 14.5975

15. What 95% range of values would you report for the possible *ln*(List Price) of the property above?

* CI = 14.5975 + 1.96\*0.3409
* CI = [13.9293,15.2656]

16. Exponentiate your predicted value from #14. What is your predicted List Price?

* List Price = e^14.5975 = $ 2,185,816.49

17. Exponentiate the endpoints of your interval from #15. What is your interval for List Price?

* CI = e^15.2656, e^13.9293
* CI = [1,120,516.17 ; 4,261,790.78]

18. The actual List Price for the above property in *Redondo Beach* was $6,500,000. What is the residual value for it based on this model?

* Residual = Actual - Predicted = 6,500,000 - 2,185,816.49 = $ 4,314,183.51

**Final Model Questions:**

19. Of the 4 regression models above, which would you choose and why?

Model 4 because

1. R^2 is 75.5 meaning the dependent variable is explained upto 75% by the independent variables
2. Model is statistically significant as p-value< level of significance
3. No issues of multicollinearity
4. All the independent variables are statistically significant

20. Give an interpretation of each of the slope estimates for your chosen model.

* CONDO/COOP (-0.1548) : When the property is a Condo/Coop(value=1) the expected List Price value decreases by 15.48% while controlling for all other variables.
* Single Family Residential (0.1734): When the property is a Single Family Residential (value=1) the expected List Price value increases by 17.34% while controlling for all other variables.
* Hermosa Beach (0.3598): When the property is located in Hermosa Beach (value=1) the expected List Price value increases by 35.98% while controlling for all other variables.
* Manhattan Beach (0.5183): When the property is located in Manhattan Beach (value=1) the expected List Price value increases by 51.83% while controlling for all other variables.
* SQFT(.0004): When the SQFT increases by 1 unit (square foot), the expected List Price value increases by 0.04% while controlling for all other variables.
* YEAR BUILT(-.0024): When the YEAR BUILT increases by 1 unit (year), the expected List Price value decreases by 0.24% while controlling for all other variables.
* DAYS ON MARKET(.0009): When the DAYS ON MARKET increases by 1 unit (day), the expected List Price value increases by 0.09% while controlling for all other variables.

3 Note here that there are only 4 property types represented. If it is known whether a property is categorized as Single Family Residential, Condo/Coop, or Multi-Family, do we know whether it is a Townhouse?

4 Note here that there are only 3 cities represented. If it is known whether a property is in Hermosa Beach or Manhattan Beach, do we know whether it is in Redondo Beach?

5 Holding other variables constant, what is the intercept (or starting point) for a property in Redondo Beach? What is the intercept for a property in Hermosa Beach?

6 See the footnote above.

2

Appendix A.

A.1. Descriptive Statistics

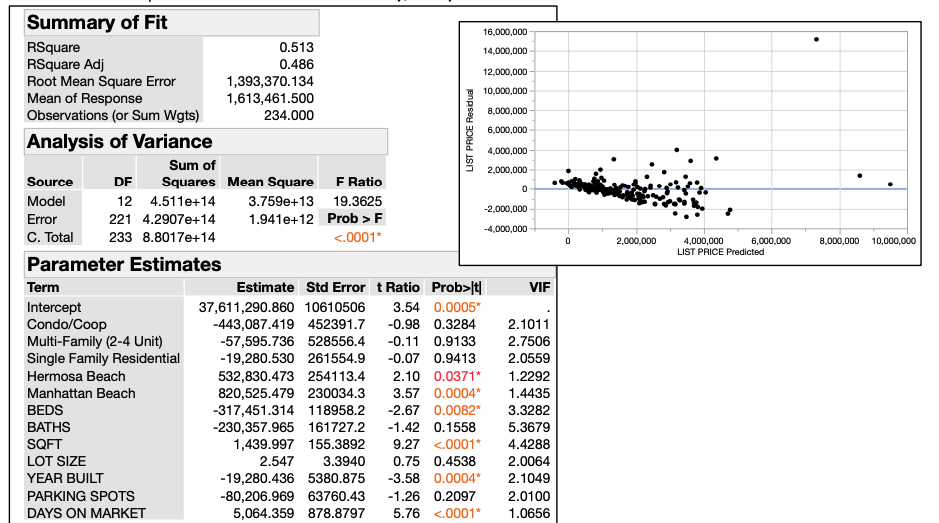
GSBA 545 Regression Assignment, Fall 2021

|  | N | Mean | Std Dev | Min | Max |
| --- | --- | --- | --- | --- | --- |
| LIST PRICE | 273 | 1,526,136.597 | 1,820,457.200 | 244,900 | 22,499,000 |
| BEDS | 272 | 3.482 | 1.396 | 1 | 9 |
| BATHS | 271 | 2.954 | 1.265 | 0.75 | 9 |
| SQFT | 272 | 2,301.610 | 1,198.358 | 384 | 9200 |
| LOT SIZE | 237 | 14,228.245 | 37,877.534 | 296 | 273001 |
| YEAR BUILT | 273 | 1,982.282 | 23.747 | 1911 | 2013 |
| PARKING SPOTS | 273 | 2.733 | 1.974 | 1 | 14 |
| DAYS ON MKT | 273 | 95.282 | 125.114 | 1 | 816 |

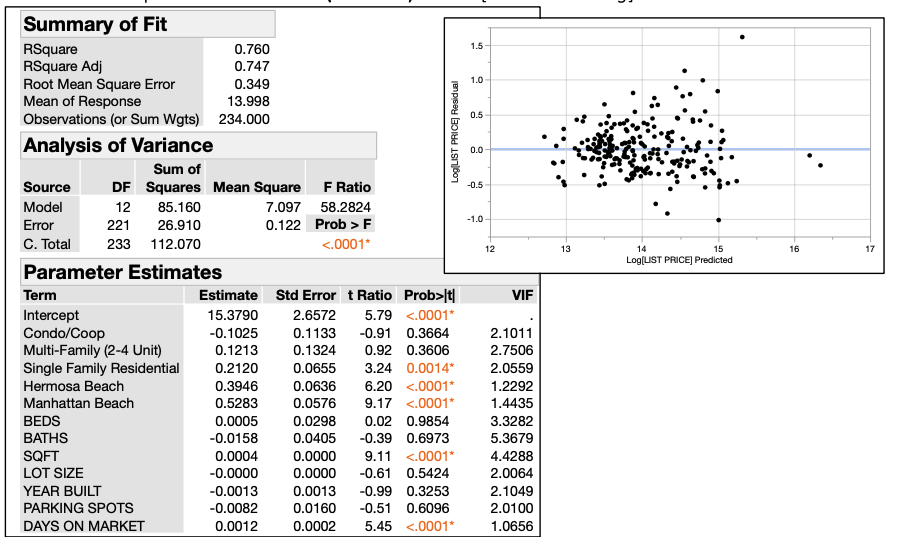
|  | N |
| --- | --- |
| Condo/Coop | 34 |
| Multi-Family (2-4 Unit) | 24 |
| Single Family Residential | 126 |
| Townhouse | 89 |
| Hermosa Beach | 56 |
| Manhattan Beach | 95 |
| Redondo Beach | 122 |

A.2 Correlations

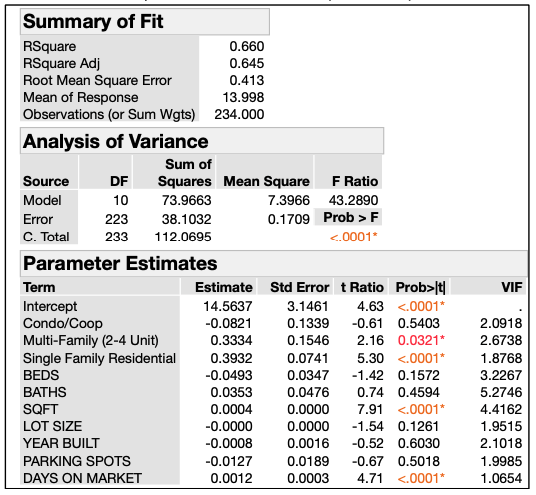
|  | LIST PRICE | BEDS | BATHS | SQFT | LOT SIZE | YEAR BUILT | PARKING  SPOTS | DAYS ON MKT |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| LIST PRICE | 1 |  |  |  |  |  |  |  |
| BEDS | 0.2958 | 1 |  |  |  |  |  |  |
| BATHS | 0.3707 | 0.7788 | 1 |  |  |  |  |  |
| SQFT | 0.5576 | 0.7164 | 0.8362 | 1 |  |  |  |  |
| LOT SIZE | -0.1514 | -0.3397 | -0.2674 | -0.2567 | 1 |  |  |  |
| YEAR BUILT | -0.0178 | 0.1076 | 0.3879 | 0.3393 | -0.0273 | 1 |  |  |
| PARKING SPOTS | 0.1912 | 0.4522 | 0.4142 | 0.388 | -0.1544 | -0.1599 | 1 |  |
| DAYS ON MKT | 0.2047 | 0.0469 | 0.0063 | -0.0488 | -0.0349 | -0.0518 | 0.0544 | 1 |

B.1. Model 1. Dependent variable is List Price ($USD)

B.2. Model 2. Dependent variable is *ln*(List Price) [\**ln* is *natural log*]

****

B.3. Model 3. Dependent variable is *ln*(List Price)



B.4. Model 4. Dependent variable is *ln*(List Price)

