**Alden Butzke**

**BA 355: Business Analytics, Case 3.2**

Use the Clean Data file available on the course webpage to answer the following questions.

1. To begin, there are 113 data points. Graph **and include here** the data with the trend line for just the square feet (as the x-variable) versus the Zestimate (as the y-variable) for all 113 data points. What is the equation of the line and the r2 value? Interpret what the y-intercept and slope tell us about the cost of a house in Durango. Note: The y-intercept is weird. Just by eyeballing the graph, do there appear to be any outliers?
2. Now let’s eliminate some of those outliers to make a more consistent data set.
   1. Use the standard Tukey’s method to determine which data are outliers for either the Zestimate, square footage or age. What are ranges are typical for the Zestimate, square footage and age? This should identify 13 total points that are outliers -- *eliminate* them from further consideration.
      1. [Hint: Use conditional formatting to find the outliers.] As a check, the average square footage of all data should now be 2353.57. Interpret the depressing news that Q1 tells new home buyers about home prices in Durango.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **sqft** | **sqft** | **age** | **age** | **price** | **price** |
| **-445** | **LL** | **-56** | **LL** | **$ (209,372.50)** | **LL** |
| **5227** | **UL** | **144** | **UL** | **$ 2,184,687.50** | **UL** |

* 1. Re-graph just square feet versus the Zestimate for these 100 data points. Redraw the graph with the regression line and equation and from your graph **and include it here**.
  2. Interpret the slope **and** y-intercept of the line from part b). According to this line, what does a square foot of housing cost in Durango?

**Each additional sqft adds about $289 dollars of value to a house price. The minimum amount a house in the data is expected to sell for is $280,650, because a house with 0 sqft is nonsensical.**

* 1. Calculate the coefficient of correlation r and the coefficient of determination r2 and interpret them both.

|  |  |
| --- | --- |
| **Corr:** | **0.72329212** |
| **R^2** | **0.52315149** |
|  |  |

**The correlation coefficient is equal to the sqrt of R^2. The correlation coefficient means is measured between -1 and 1, where values closer to 1 indicate a strong positive relationship, values close to -1 indicate a strong negative relationship, and near 0 indicates no relationship.**

**Thus, 0.7 indicates a strong positive relationship.**

**The R^2 explains how well the variation in X’s explain the variation in Y’s, thus it is measured between 0 and 1, where values closer to 1 indicate more variation in Y is explained by X. Thus, 0.52 is interpreted as 52% of the variation in Price is explained by sqft.**

* 1. About how much is my 1441 square foot house worth according to part b)?

**= (1441)(289)+280650 = $697,099**

* 1. Force the y-intercept to 0. (Don’t need to print this one). What does the slope say now about the cost of a square foot?

**The price of a sqft has gone up to $391 in order to make up for the variation that used to be explained by the intercept. Thus, this changes the model from a fixed price and a unit price, and makes it one larger unit price.**

1. Using just the linear equation you found in 2b),
   1. Forecast the cost of each of the 100 homes.
   2. Calculate the absolute percentage error for each data point then compute the *mean* absolute percentage error (MAPE) and *median* absolute percentage error (Median APE – sounds like a terrible movie) for this method.

**MAPE = 20%**

**Med APE = 16%**

* 1. Go to Zillow ([What is a Zestimate? Zillow's Zestimate Accuracy | Zillow](https://www.zillow.com/z/zestimate/)) and find what the median error rate is for off-market nationwide and for the state of Colorado. Our model has a ways to go…

**Nat = 3.2%**

**CO = 1.5%**

1. Now, run multiple linear regression (available with the data analysis package in Excel, it’s an Add-In like Solver) with the zestimate as the y-variable and all four other columns (square footage, bedrooms, bathrooms and age) as the x-variables.
   1. What do the “Multiple R and R Square” values at the top of the output indicate about how well our model is preforming? How do they compare to your answers in 2c)?

|  |  |
| --- | --- |
| **Multiple R** | **0.74749994** |
| **R Square** | **0.55875616** |

**The multiple R is the correlation coefficient between three or more regressors, and it is the multiple regression equivalent of the correlation coefficient from 2c. It is higher in this model by .02, thus the positive correlation has increased.**

**The R^2 has also increased by about the same amount, and thus the variation in price explained by the model has increased. However, it is more reliable to use the adjusted R^2 because when more inputs are added, the R^2 will necessarily increase.**

* 1. The “Significance F” is really the overall p-value for the whole model – **if it’s close to 0% it means the model works; if it’s closer to 100%, the model doesn’t work**. What is this number – as a percentage – and what does it say about our model?

**The Significance F is 3.6 \* 10^-17, which is very near 0, in percentage form it is 0.000000000000036%. This value tells us how significant all of our regressors are at the same time.**

* 1. List the multiple linear regression equation.

**Y = 22360 + 252(sqft) + 47438(bed) + 42649 (bath) + 1298(age)**

* 1. About how much is my house worth with 1441 square feet, 3 bedrooms, 1.5 bathrooms, and built in 1979.

**$647,587**

* 1. Convert the p-values for the y-intercept and x-variables into percentages and list them. Which factor seems to be the least relevant to our model? Generally, p-values close to 100% represent *irrelevant* factors that should not be included in the model and are just noise; small p-values near 0% represent *relevant* factors.

|  |  |
| --- | --- |
|  | ***P-value*** |
| **Intercept** | **85%** |
| **Sq Ft** | **0.0000004%** |
| **Bed** | **20%** |
| **Bath** | **30%** |
| **Age** | **10%** |

**The intercept, bed and bath at least should be eliminated, with B0 being the worst. 333**

1. Eliminate the variable with the worst (highest) p-value, keep the others and rerun the multiple linear regression. If this is the intercept, click the “constant is zero” check box.
   1. What are the individual p-values now, listed as percentages?

|  |  |
| --- | --- |
|  | ***P-value*** |
| **Sq Ft** | **0%** |
| **Bed** | **11%** |
| **Bath** | **25%** |
| **Age** | **6%** |

* 1. How do the r value and r2 values compare to what you got in 4a) and 2c)?

|  |  |
| --- | --- |
| **Multiple R** | **0.97014582** |
| **R Square** | **0.9411829** |
| **Adjusted R Square** | **0.9289282** |

**These are higher, thus more variation is explained which makes the model a better fit, though the high p values of bath and bed are suspect.**

* 1. What is the new “Significance F” and what does it tell us?

|  |
| --- |
| ***Significance F*** |
| **1.0193E-57** |

**This model got more accurate because the f-sig level is even closer to 0. This stat tells you how likely the variation between the explaining variables and the responding variable is correlated just due to chance. Thus, a low level means a significant result is more likely.**

* 1. List the multiple linear regression equation **and interpret** all parts of it.

**Bo(0) + Sqft(252) + Age(1351) + Bed(50871) + Bath(44753) = Price**

**The intercept is 0 so the minimum amount a house can be worth is 0, each additional sqft adds $252 to the selling price, each year older a house is raises the price by $1,351, each bedroom increases the price by $50,871, and each bath increases it by $44,753.**

* 1. About how much is my house worth (sqft =1441, bed = 3, bath = 1.5, built in 1979)?

**$$613,948**

* 1. Use the multiple linear regression equation to forecast all 100 home values and then calculate the *mean* absolute percentage error (MAPE) and *median* absolute percentage error (Median APE) for this method. How do these compare to what you got in part 3b)?

|  |  |
| --- | --- |
| **MAPE** | **18.63%** |
| **Med APE** | **15.32%** |

1. Eliminate the x-variable with the worst (highest) p-value and rerun the regression. Repeat parts a) – f) of 5).
   1. What are the individual p-values now, listed as percentages?

|  |  |
| --- | --- |
|  | ***P-value*** |
| **Sq Ft** | **0%** |
| **Bed** | **1%** |
| **Age** | **10%** |

* 1. How do the r value and r2 values compare to what you got in 4a) and 2c)?

|  |  |
| --- | --- |
| ***Regression Statistics*** | |
| **Multiple R** | **0.9697277** |
| **R Square** | **0.94037181** |
| **Adjusted R Square** | **0.92883309** |

**They are slightly less than before, which means the model explains less variation, however, the p value is lower so the result is more significant.**

* 1. What is the new “Significance F” and what does it tell us?

|  |
| --- |
| ***Significance F*** |
| **7.9796E-59** |

**This model got more accurate because the f-sig level is even closer to 0. This stat tells you how likely the variation between the explaining variables and the responding variable is correlated just due to chance. Thus, a low level means a significant result is more likely.**

* 1. List the multiple linear regression equation **and interpret** all parts of it.

**Bed(69392) + Sqft(276) + Age(1170) = Price**

**The intercept is 0 so the minimum amount a house can be worth is 0, each additional sqft adds $251 to the selling price, and each year older a house is raises the price by $2,269.**

* 1. About how much is my house worth (sqft =1441, bed = 3, bath = 1.5, built in 1979)?

**$632,802**

* 1. Use the multiple linear regression equation to forecast all 100 home values and then calculate the *mean* absolute percentage error (MAPE) and *median* absolute percentage error (Median APE) for this method. How do these compare to what you got in part 3b)?

|  |  |
| --- | --- |
| **MAPE** | **18.52%** |
| **Med APE** | **15.34%** |

1. We now have three models for predicting the Zestimate. One (in parts 2) and 3)) uses the square footage as the only x-variable. One (in part 5)) uses everything but the intercept and one (in part 6)) that uses only three x-variables. Compare the MAPE and Median APE for each model. Which model do you think is best? There is typically a trade-off between simplicity and accuracy. In this case, is it worth adding the extra variables – does it increase the accuracy of our forecasts enough to justify the increased complexity?
2. **MAPE = 20%**
3. **Med APE = 16%**

|  |  |
| --- | --- |
| **2. MAPE** | **18.63%** |
| **2. Med APE** | **15.32%** |
|  |  |

**3: MAPE 18.52%, Med APE 15.34%**

**Model 1 is the simplest, and is within several decimal places of the median error of the more complex models. The mean error is about 2% greater, which can be significant when dealing with large values. However, for the average house, the simplest model is not significantly worse at estimating the selling price as the multiple regression. I do not think the increased accuracy is worth it to justify the complexity, however that said, I would still develop the most precise model possible if I was involved in the transaction.**

1. In part 5), the coefficient for Age is positive meaning that **the older a house is, the more it is worth**; this seems counter-intuitive or even paradoxical – generally newer houses (with newer everything) are worth more and older houses – with old electrical, plumbing, etc. that need repairs – are worth less. Explain this Durango paradox. Think like a real estate agent here.

**In Durango, many of the houses are old, and the location of the old houses is in prime real estate (the grid, nearby downtown). Moreover, the old houses have historic value and are generally in good condition. Thus, though in normal circumstances, age decreases the selling value, in Durango, it increases the price. When there are not many new homes built, and the existing homes are old but in the prime location with minimal supply and lots of demand, the age is not a deterrent to buyers.**