**BA 355: Business Analytics (30 points)**

**Case 3.1: Zillow.com**

Collect 10 in town Durango data points from Zillow.com and 10 from somewhere else (your hometown, somewhere you’d like to live someday, somewhere awful – your call). Include the address, the zestimate price estimate, the square footage, number of bedrooms, number of bathrooms and year built. Keep track of which properties are in your data set as we will revisit them later. Think about other factors, too, such as property tax, location, school district, previous sales information, etc.

1. Input the two data sets into two separate Excel sheets with the address in column A, the zestimate in column B, square footage in C, bedrooms in D, bathrooms in E and age of the house in F.
2. Examine both data sets for “outliers” – data points that seem very different from the rest of the data. (We’ll learn a method for this next week, but for now just eyeball it.) Are there one or two houses that are significantly more expensive (or less, I guess?) than the others? Decide whether to include or exclude any potential outliers from the analysis below. The benefit of including them is more data, the possible issue is they may throw off your analysis below. If you do include them, they will probably stand out on the graph you plot in part c). This is decision to include or exclude is yours – no real right or wrong answer here – but think carefully about it first.

**I have one outlier in each data set on the high end. I am going to include the outliers because even though least squares is only reliable without outliers (otherwise the errors will not produce coefficients that represent the population), the outliers ARE real population level data, and might not be outliers if I sampled the population many times over.**

1. On the Durango data, run linear regression letting the zestimate be the y-value and using square footage as the only x-value. ***Interpret*** the slope **and** y-intercept of this line. **Graph and include** the data points with the LR line fitted to them.

**B0 is $86,017, and B1 is $419. Thus, for every 1 sqft, the price of a home is expected to rise by $419, and the average house price with sqft held constant is $86,017. It doesn’t make sense to interpret B0 as a house with no sqft, so it makes more sense to interpret it as the lower bound for price in our data.**

1. Now, force the y-intercept to be zero and re-interpret the slope.

**With B0 held to 0, B1 becomes $454, so for every increase of 1 sqft, a house goes up on value by $454. Thus, a house of 1 sqft would be worth $454. The slope necessarily increased because it had to make up for the difference in starting points from the constant.**

1. For one more way to estimate the cost per square foot, simply divide the sum of column B by the sum of column C. How does this compare to the slopes from c) and d)?

|  |  |
| --- | --- |
| 9517862 | /20649 |
|  | **= 461** |
|  |  |

**$461 is similar to the slopes above, and it is the largest coefficient yet.**

1. Using the line from part c), determine the bestimate™ for each property and compare it to the zestimate. Calculate the **mean absolute percentage error** for how far our bestimate™ is from the zestimate.

|  |  |  |
| --- | --- | --- |
| **zestimate** | **bestimate** | **difference** |
| **621776** | **625629.138** | **0.6%** |
| **1250001** | **1581585.05** | **26.5%** |
| **949000** | **1352239.48** | **42.5%** |
| **604529** | **584539.805** | **3.3%** |
| **984712** | **1307376.64** | **32.8%** |
| **799907** | **791663.585** | **1.0%** |
| **580900** | **627725.533** | **8.1%** |
| **2412207** | **1381169.73** | **42.7%** |
| **989000** | **847846.959** | **14.3%** |
| **325830** | **418086.079** | **28.3%** |

**mean absolute percentage error = 20%, this is getting pulled up by two very high data points, but I am still including them.**

1. Run multiple linear regression on the 10 Durango data points (in Excel, I can show you how) using square footage, bedrooms, bathrooms and how old the house is as your x-variables (plural) and the zestimate as your y-variable. Determine the multiple linear regression equation that predicts price based on these variables.

**Y = 1,808 + sqft(319) + bed(5,784) + bath(163,602) – age(1,814)**

1. Interpret the p-value and slope for each x-variable and the y-intercept. Which factors seem relevant to the model and which don’t? (Factors with higher p-values are generally *irrelevant,* factors with lower p-values are generally *relevant.*)

|  |  |  |
| --- | --- | --- |
|  | *Coefficients* | *P-value* |
| Intercept | 1807.75695 | 0.99759786 |
| sqft | 319.705283 | 0.21991633 |
| bed | 5783.86817 | 0.98002632 |
| bath | 163,601.813 | 0.40054336 |
| year | -1814.3828 | 0.71694509 |
|  |  |  |

**The highlighted independent variables (B0, bed, and year) have P-vales above 50%, which imply that there is not a true relationship between them and the dependent variable (price).**

**The coefficients are interpreted the same as before where the lowest price our model will predict for a house is $1808, and for each increase in one sqft, price increases by $320, for each additional bedroom, price increases by $5784, for each additional bathroom price increases price by $163,602, and for each additional year, price decreases by $1,814 dollars. The direction of these variables make sense, but their magnitude is suspect.**

**I am going to rerun the regression with B0 set to 0, and bed and year eliminated because of their high p-values.**

1. Rerun the multiple regression, eliminating the factors that are not relevant. Using this model, determine the bestimate™ for each property and compare it to the zestimate. Calculate the **mean absolute percentage error** for how far our bestimate™ is from the zestimate.

|  |  |  |
| --- | --- | --- |
|  | ***Coefficients*** | ***P-value*** |
| **Intercept** | **0** | **#N/A** |
| **sqft** | **284.490468** | **0.04978078** |
| **bath** | **166,714.899** | **0.16652679** |

|  |  |  |
| --- | --- | --- |
| **zestimate** | **bestimate** | **difference** |
| **621776** | **866283.929** | **28%** |
| **1250001** | **1681637.1** | **26%** |
| **949000** | **1025876.11** | **7%** |
| **604529** | **504974.065** | **-20%** |
| **984712** | **1162150.53** | **15%** |
| **799907** | **812227.256** | **2%** |
| **580900** | **867706.382** | **33%** |
| **2412207** | **1545650.65** | **-56%** |
| **989000** | **850348.978** | **-16%** |
| **325830** | **392031.35** | **17%** |

**mean absolute percentage = 22%**

1. Repeat c) – f) for the other 10 data points from elsewhere. (Skip g) through l) – multiple regression on one small data set is enough.)

|  |  |  |
| --- | --- | --- |
|  | *Coefficients* | *P-value* |
| Intercept | -74526.613 | 0.86001957 |
| sqft | 296.871049 | 0.10447293 |
|  |  |
|  | *Coefficients* | *P-value* |
| Intercept | 0 | #N/A |
| sqft | 263.143329 | 0.00025274 |
|  |  |  |

**The coefficients change, but completely differently than in the Durango data. B0 starts off as negative here when it is not forced to 0, so the sqft coefficient increases in order to make up for this difference. Thus, when B0 INCREASES to 0, B1 must decrease in order to maintain consistent predictions.**

**The intercept loses its interpretation as a lower bound when it decreases less than 0, and thus it can be thought of as an indication of an unreliable model due to collinearity. B1 is still interpreted as for each 1 additional sqft, price increases by $298**

**Hence, the model where B0 is forced to 0 makes more sense, and B1 is still for each increase in sqft, price increases by $263.**

1. Compare the Durango data to the other data. What are the main differences between the two areas? What do the slopes tell you about the cost of a square foot of housing, etc.?

**When the B0 was forced to 0 in Durango, sqft moved up to $453 from $419, for an 8% change, and in Anchorage, it moved down from $296 to $263, for a 13% change. First, the higher value in Durango indicates that each sqft is worth more than in Anchorage, however it’s tough to tell whether this is because the houses are more expensive overall or if one sqft really is valued more. Thus, comparing the difference in percentage change would imply that sqft is a bigger determinant of the price because it nearly changes twice as much once B0 is set to zero. However, this could be a quirk of the data.**

**On a macro level, the market in Anchorage is shifted down by a few hundred thousand dollars. This in demonstrated by the difference in magnitude of coefficients, and price predictions.**

1. Estimate how much my house is worth in Durango and what it might be worth in your other location using whichever model from above that you think is best overall. 1441 square feet, 3 bedrooms, 1.5 bathrooms, built in 1979.

**$ 660,023.11 according to my model, 1.6B according to my heart**