**Ames, Iowa Home Price Modeling**

**Preeti Swaminathan & Patrick McDevitt**

**Homework 13**

**Analysis Question 1**

**Restatement of Problem:**

Century 21 in Ames Iowa wants to build a model to predict sales price of 3 neighbourhood (BrkSide, NAmes, EdWard) in Iowa based on living area. They have provided with historical data of sales done in these neighborhood so far.

To-do: Build a model which uses (independent variables) living area in sq.ft and the neighborhood and predicts sales price (Dependent variable)

**Specify the model:**

SalesPrice= 𝛽0+𝛽1\*GrLivArea+ 𝛽2\* neighborhood

**Checking Assumptions**

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | |
| 1. From scatter plot, QQ plot and histogram, data is normally distributed 2. Interactions: looking at analysis of covariance plot, High leverage mild departures: 3. Looking at scatter plot b/w Sales price and living area, there are a few outliers, skewness, we can consider doing log transformation on sales price. 4. Looking at the plots, data is clustered in one group with outliers. This data needs to be log transformed | | | |
|  |  | | **Categorical variable analysis:**  Data is concentrated around living area of < 30 sq foot and sales price for < 300k.  Edward neighborhood has a few bigger houses and higher price. |

**Model 1: Log transformed on Sales price and living area.**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | | |
| **Checking Assumptions**   1. Normality: From QQ plot, scatter and histogram, we can assume normality. Although histogram shows a little skewness it’s not very strong evidence against normality. 2. Linearity: Looking at the residual plots we can make an assumption of linearity. 3. Equal variance: After log transformation, since none of the plot looks too bad, we continue with our model. 4. CookDs shows couple of extrems. We will keep a check on it. | | | |
|  | |  | |
| **T test and pValue**  P value for Neighborhood BrkSide and logGrvLivA \* Edwards is > 0.05. We will make **BrkSide** as reference and recalculate the model | | | |
|  | |  | |
| Pvalue < 0.05 for all variables. Living Area and neighborhood are significant predictors of sales price. | | | |
| **Fit the model**  Log(SalesPrice)= 𝛽0+𝛽1\*Log(GrLivArea)+ 𝛽2\* neighborhood BrkSide+ 𝛽3 neighborhood Edwards + 𝛽4 log(GrLivA) \* neighborhood BrkSide + 𝛽5 log(GrLivA) \* neighborhood Edwards  Log(SalesPrice) = 9.69 + 0.820\*Log(GrLivArea)+ 0.712\* neighborhood Edwards +0.984 neighborhood NAmes - 0.30 log(GrLivA) \* neighborhood Edward- 0.347 log(GrLivA) \* neighborhood NAmes | | | |
| **Parameter Interpretation:**  Log(SalesPrice) = 9.69 + 0.820\*Log(GrLivArea)+ 0.712\* neighborhood Edwards +0.984 neighborhood NAmes - 0.30 log(GrLivA) \* neighborhood Edward- 0.347 log(GrLivA) \* neighborhood NAmes  **Edward:**  Log(salesPrice) = 9.69 + 0.820\*Log(GrLivArea)+ 0.712 - 0.30 log(GrLivArea)  Log(salesPrice) = 10. 402 + 0.52 \* Log(GrLivArea)  SalesPrice = e ^ 10. 402 + 0.52 \* Log(GrLivArea)  SalesPrice = e ^ 10.402 \* grLivArea ^ 0.52  **SalesPrice {EdWard} = 32925 \* grLivArea ^ 0.52**  For a 1sqr.ft house salesPrice of EdWArd = 32925$  Doubling it, GrLivingArea = 2, SalesPricec increases to 32926.43$  **Names**  Log(SalesPrice) = 9.69 + 0.820\*Log(GrLivArea) +0.984 - 0.347 log(GrLivArea)  Log(SalesPrice) = 10.674 + 0.473\*Log(GrLivArea)  SalesPrice = e^10.674 \* GrLivArea ^ 0.473  **SalesPrice{NAmes } = 43217 \* GrLivArea ^ 0.473**  For a 1sqr.ft house salesPrice of EdWArd = 32925$  Doubling it, GrLivingArea = 2, SalesPricec increases to 32926.43$  **BrkSide**  Log(SalesPrice) = 9.69 + 0.820\*Log(GrLivArea)  SalesPrice = e^ 9.69 \* GrLivArea ^ 0.82  **SalesPrice {BrkSide} = 16155 \* GrLivArea ^ 0.82** | | | |
| **R2 and Root MSE** | | |  |
| **Summary of anlysis:**   1. R2 = 0.512, 51.2 % of variation in salesprice is affected by living area and neighborhood. 2. It appears that higher squart ft the sales price increases for all 3 neighborhood. 3. BrkSide neighborhood increases salesprice significantly compared to NAmes and EdWard for this dataset. 4. Model without interactions was also developed. It did not produce reasonable results hence was discarded.  |  | | --- | |  | | | | |