Gulshankumar Bakle **Stat511 Report**

Introduction:

The prices of land properties or houses depends on various features. These features equally or independently affect the prices in any area. For houses some of common features are number of rooms and bedrooms, age of house, population of the area where house is located, average income of residents already living in that area, etc. These features are called as explanatory or predictor variables. The prices of houses which are affected due to increase or decrease of any of these variables is termed as response or predicted variable. Purpose of this report is twofold:

1. To compare the difference in population means of prices of houses between feature groups using Analysis of Variance (ANOVA);
2. To predict the mean of prices of houses using one of predictor variable using a simple regression model.

The results are derived and inferred from the analysis done on USA Housing dataset, collected from here [1]. The dataset contains 5000 different prices of houses, over various cities and towns in the US. There are 6 features which govern the prices of houses -average income of residents, average age of house, average number of bedrooms, average population of area, address of house and average number of rooms in house. The 7th column lists the average price of houses. Density plot for all the columns is shown in fig 1.

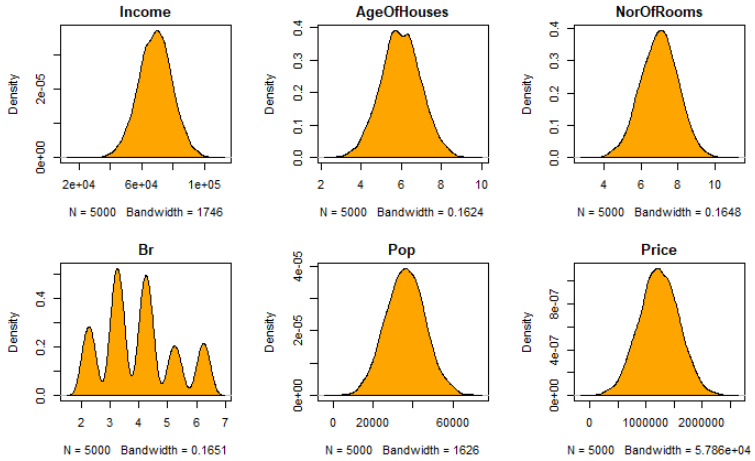

Fig 1

Methods:

For comparing the population means of prices of houses between different groups, ANOVA method is used. Following are the assumptions in order to perform ANOVA:

a. Independence within and between groups
b. Normality of populations
c. Equal variance

For simple linear regression, following assumptions are made:

a. The mean of Y variable (price of houses in this case) is a linear function of Age of house.
b. The pairs (prices, age of houses) are independent of each other.

c. Variance of each Y around mean of $\mu(Y|X) = \beta0+\beta1*X$ is same value σ^2.
d. Since we have large dataset, distribution of response variable is normal.

Results:

ANOVA:

- Null Hypothesis: All the population means of prices of houses are equal for all groups of age of houses. Mathematically, H0: $\mu1= \mu2= \mu3=\ldots \mu1$
- Alternative hypothesis Ha: At least two population means of prices are different for two groups

On performing aov() on linear model of Prices and age of houses, it was observed that p-value is close to 0, which is less than all the significance levels alpha. Hence null hypothesis is rejected in favor of alternative that at least for two groups could have unequal population means. Following were the key results observed:

- Sum of squares for Age groups is 1.245e+14
- F statistic value is 138.4
- P-value is <2e-16

Linear Regression:

- Null Hypothesis: There is no linear relationship between response variable mean of price of house and explanatory variable age of house
- Alternative hypothesis: There is linear relationship between mean of price and age of houses.

From fitting a linear regression model, we observe following results:

- The parameters for best fit lines are as intercept($\beta0$) =268677 and slope($\beta1$)=161178.
- 95% confidence interval for intercept is between 215315 to 322039.1 and for slope is 152370.6 to 169985.1
- P-values for both intercept and slope is <2e-16, which shows that null hypothesis is rejected, that is there could be linear relationship between mean of price of houses and age of houses.
- Estimated standard deviation is 314900 on 4998 degree if freedom.
- Multiple r-squared value is 0.2048

Assessment:

ANOVA: On plotting residual vs fitted graph it was observed that, though first three assumptions were met, but fourth assumption was not observed. Thus, we perform Kruskal Wallis test which is non-parametric and an alternative ANOVA. Fig 2 shows that there is equal spread across residual 0, hence population variance for prices of different groups of houses don't necessarily have to be the same.

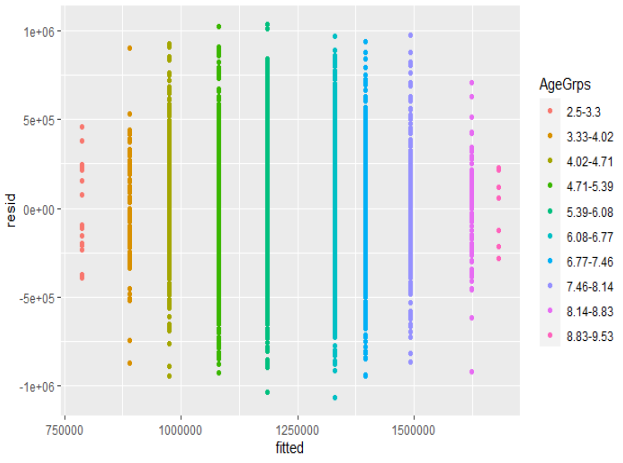

Fig 2

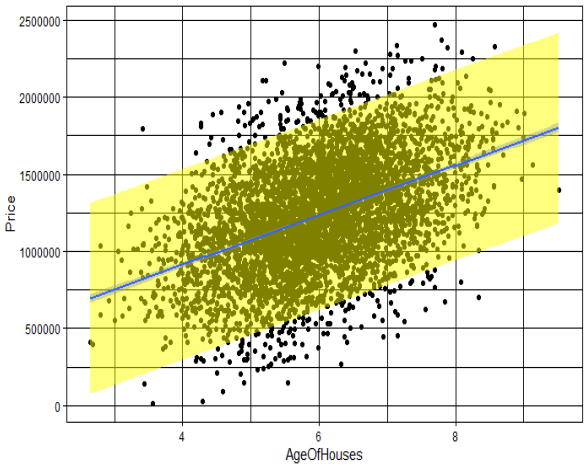

Fig 3

Simple Regression: All the assumptions were met w.r.t predicted and predictor variables. Fig 3 depicts the fitting of simple linear regression model for (Prices ~ Age of Houses). Yellow shade indicates the predicted interval for fitted values, blue line indicates the best fit line around the scattered points. Linear regression is also performed on other explanatory variables like Number of rooms, which are included in code.

Conclusion:

From ANOVA test it can be concluded that population mean of prices of houses could be different for different groups of age. There could be at least two groups that don't have same population mean. Other multiple comparison procedures like Tukey-Kramer or Fischer's test could also be considered, which provide deeper insight about group means by considering all possible combination of groups.

From the fitted model, mean of price of houses can be predicted depending upon any random age of house. The predicted prices of house could be made more generalized by fitting a multiple regression model using other features like number of rooms, population of locality etc. Using multiple regression model, can produce a better plane around mean of price of houses and other explanatory features, hence making our analysis even more robust and accurate.