Homework 2

Answer 2

1. **Import the data house_train.csv and house_test.csv in R.** The two files consist of 7 and 6 variables respectively. House_train consist of 7 variables namely id, zip, state, county, poverty, price2007 and price2013 while the house_test data consists of all the variables except price2013. Our objective in this assignment is to build a model to predict the prices of houses for year 2013 for the house_test data. After importing the data in R the data was attached to make our further commands run easily.

2. **To avoid any null records run a command in R to omit them.** It was observed that house_train consists of 8979 observations both after the na.omit command and before that command as well confirming the fact that there were no missing values in the data.

- **Now in order to predict prices of houses for the year 2013 we are making a linear regression model with categorical variable state.** Our next aim was to build a regression model to predict the house prices for year 2013.
- **Use the summary function to view the model built.** To view the values got from the model like the intercept value etc.
- **Intercept Value.** The value of the intercept is 281730 and it corresponds to average cost of house in state 'AK'. In order to calculate the average cost of a house in any other state this value is added along with the value of regression coefficient for that sate.
- **To get this information from the model** we keep the value of all the independent variable in the model = 0. The value then got for the dependent variable is equal to the value of the intercept which in this case is the average cost of the houses in state AK**.**
- **Based on the regression coefficients the state with most expensive average home is DC and the least expensive is WV.**
- **To get this information** we have to check for each of the coefficient values for each of the state and keep them in the equation (model equation) where the value of coefficient for a particular state is the independent variable and the value got after adding the value of intercept is the value of the dependent variable i.e. the value we need to compare.
- **The average price of homes in DC is 514288.9 and the average price of homes in WV is 98423.1.**
- **To get this information the value of intercept was added to the regression coefficients of DC and WV one by one.**

3. **The next step is to build a regression model including the dummy variables county and state i.e to do multiple linear regression.**

- The county Pitkin has the highest house prices while the house prices for calaveras is the lowest. These counties have the highest and lowest prices because after doing the multiple regression when we look at the values of the coefficients we see that Pitkin has the highest value which when added to the intercept value will become the highest value and Calaveras has the least value.

4. Here we have to predict the value for the house prices for the year 2013 in the house_test data. **To do this we do multiple regression using different dummy variables to get the best score on Kaggle.**

- While predicting the value the first error that came was to include the six counties to house_train data which were present in house_test data and I was unable to run the regression. I just included the counties with the value of other variables = 0. Then I started with state and gradually increased the variables each time and did multiple trails to get the best score. The way to predict that the score could be better than the previous one was to look at the Adjusted R squared value which if close to 1 will give the best result.
- My best Kaggle score was 44361.928.
- My Kaggle username is Yashika Bajaj

5. **P(White ball) = (P(white ball/Bag2)*P(Bag2))+(P(white ball/Bag1)*P(Bag1))**
= (3/4 * 1/2) + (2/3 * 1/2)
= 17/24

6. **A = Opposing team scores first**
B = Own team scores first
P(Win) = P(Win/A)*P(A) + P(Win/B)*P(B)
= 0.1*0.7 + 0.6*0.3
= 0.25
Therefore, if the team scores the first goal about 30% of the time, it can win 25% of the games.