**Assignment 04 (423: Data Analysis and Regression)**

**Saransh Thakur**

**Id: 2020070**

**Problem 1**

**Table

Description automatically generated**

1. I first looked for a scatter plot and summary of the model but could not find any relevant information from it. So, I tried to remove the variable one after another, but I find not so useful. I removed the city code, made and built-in and tried various second-order models and interaction terms., some of them worked well and some didn’t.

For the second-order model, I use room squares and it worked well, I have tried all other variables, but they didn't lie within the significance value of 0.05. The interaction term we have (stromprotector \* guest rooms) has a significant value of 0.05.

1. The significant second-order term I find is the number of rooms. I have tried all the variables in the dataset and tried to do squaring for them, but only 1 variable i.e., the number of rooms lies under the significant value of 0.05.

I have looked at scatter plots for finding second-order terms, but I did not find them relevant enough. To predict the second-order model from the scatter plot we need to see if the scatter plot is showing any curve, but I did not find it in any of the plots.

The benefit of trying all the second-order terms is that you will find which term works well for you and which does not, but the major drawback of this method is you need to check for all datasets. It will work well for small datasets but will have issues with a large dataset. It’s not feasible for a large dataset.

On the other hand, looking at the scatter plot will give you an idea of the relationship between the independent and dependent variables. But, for scatter plot the interpretation of the scatter plot should be right other it will give issues.

1. Yes, I find one interaction term: (Strom protector \* guestroom). I did not try all the combination terms, but I tried most of them.

No, I did not find this appropriate strategy because there are too many terms, and checking combination for each term is time-consuming, hence this is not a good method.

As the number of independent terms increases, it will directly affect the interaction term. An increase in independent terms will increase interaction terms., so there is the chance that data will be in different formats, hence it will make the model more complex.

1. F-Test is 1.53e+09 and the p-value is 2.2e-16 which means we reject the null hypothesis that all the Betas are zero and accept the alternative hypothesis i.e., at least one of them is equal to zero. and adj-R2 is 1 which means 100 % of the variability is explained by our model. The t-test isA screenshot of a computer

   Description automatically generated with medium confidence

We can see that square meter, has a yard, has pool, floor, city range, and the guest room has passed the t-test. The second-order term(numberofRoomSQ) and the interaction term (strom\_guest) have also passed the t-test. As, the interaction term passes the t-test the children also pass the t-test, which means that Strom protector and guest room also pass the test. So, some variables have significance, hence from this, I can conclude that this is a good model.

1. It is a mistake to treat the zip code as a number because the zip code will affect our data set and the zip code will not add or reduce the value of the house. For example, 60613, If we try to find the second-order model from this, squaring that will make it a larger number and make our model complicated.

I have completed this work independently. The solutions given are entirely my work.