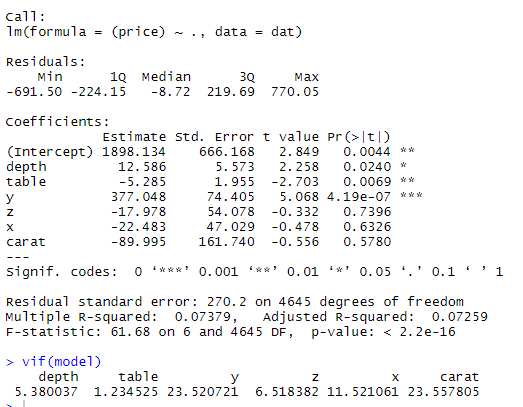
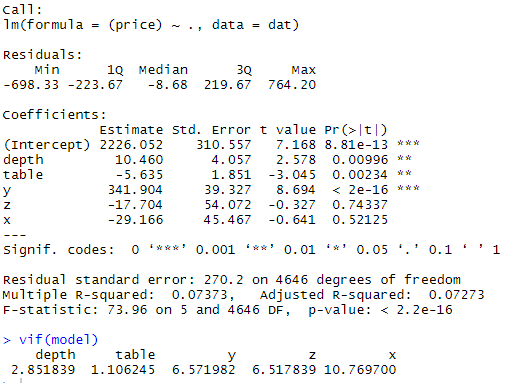
We fit the model using just the continuous variables and try identifying multicollinearity using VIF



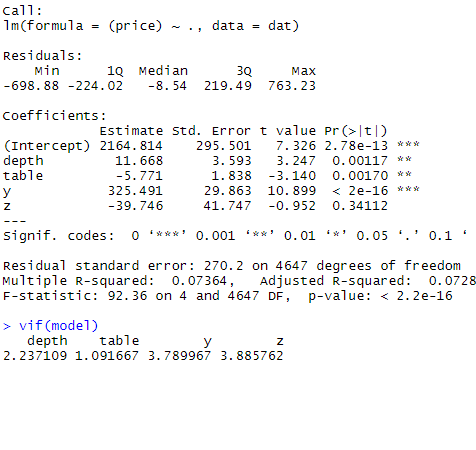
We see that the variable **carat** has the highest VIF and is above the generally used cut-off of 5.

Hence it is removed from the dataset. The VIF is checked again.



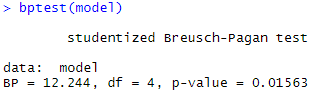
We see that the variable **x** has the highest VIF and is above 5. So we remove it from the dataset

We repeat the process is again.



The VIFs do not indicate any multicollinearity now.

We now check for the presence of heteroscedasticity using the breusch-pagan test

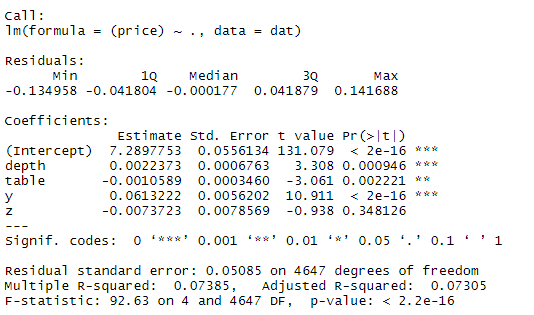


We get a p-value of less than 0.05 which is the level of significance we consider. Therefore the null hypothesis of homoscedasticity is rejected and we have heteroscedasticity issue now.

To overcome this issue, we use the Box-Cox transformation for the **price** variable.

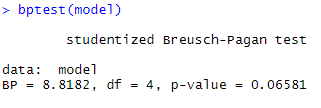
We get a lambda value of -0.0202 for the transformation and the process is done accordingly.

We run the regression again using the transformed **price** variable as the target.



We already see a slight improvement in the adjusted R squared.

We now perform the breusch-pagan test once again.



This time, our null hypothesis of homoscedasticity is accepted at the 5% level of significance.



We see there is constant variance in the residual plot.

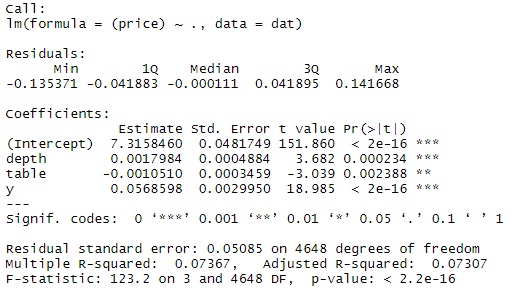
We now check for the normality of errors.



The qq-plot reveals a strong non-normal distribution of errors. Since this happens even after the box-cox transformation, we decide to use the adjusted R squared for model building (variable selection) and not look into the partial F statistics and p-values to decide on the significance of the independent variables. Therefore our model reduces to a purely predictive one.

We now perform backward elimination using the adjusted R squared to arrive at our final model.

Only the **z** variable is removed from the model as it leads to a slight increase in the adjusted R squared. No other elimination proved to be beneficial.



We define the performance of the model using the observed mean squared error.



Which corresponds to approximately an average error of 1 price unit by taking the inverse of the box-cox transformation.

We also conduct a train-test evaluation by taking a 75-25 split and fitting the same model on the train data and predicting on the test set. We obtain a mean squared error better than even the train MSE. This shows the model generalizes well.



Conclusion

We realise that the dataset we worked with posed several issues such as multicollinearity, heteroscedasticity and non-normality of errors. Standardizing variables, using random sampling to balance the distribution of the **price** variable, weighted least squares were attempted but failed to produce improved results.

Since we are restricted to only model the data for predictive purpose, the next step would be using ML algorithms with better predictive power despite being black-box. Tree-based models could be a way ahead as they do not require distributional assumptions. Another way could be to perform classification based on a target variable that has 5 or more levels according to the **price**.

The adjusted R squared is on the lower side but we conclude that it is up to the users to depend on the predictive and generalizing power of the model based on the train and test mean squared errors we have obtained.

Note:

We use only subset of the data that contains price values between 4000 and 5000 and exclude the dummy categorical variables as not doing so leads to severe heteroscedasticity issues that could not be addressed even through further transformations and weighted least squares.

Personal notes:

* Please include information from online about box-cox transform, bp test etc and basically every concept we have used
* I didn’t use variance decomposition because it shows high problems and leaves us with just one independent variable eventually. So I skipped that.
* The residual plot is a clear straight line which means errors have correlation issue. So I’ll think of a way to respond to ma’am if she asks us about it, just in case. (Correlation issue leads to bad predictions so if you want, you can remove that plot. Lets take a vote in the group for this maybe.)