UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS



Máster Data Science

HOMEWORK 2: MULTIPLE REGRESSION ANALYSIS

EL ABASSI WIDAD

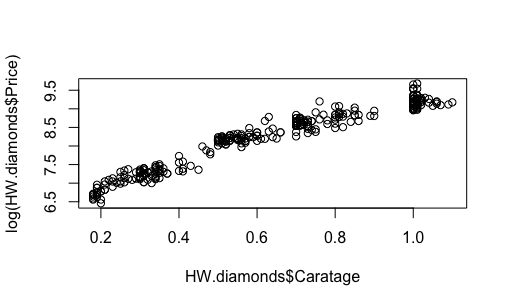
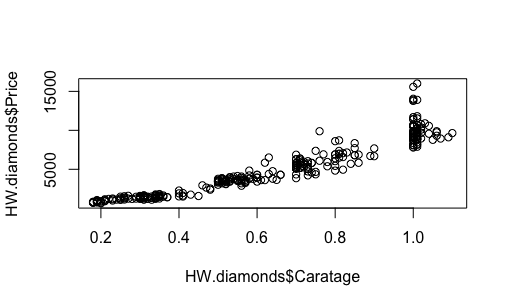
JIAYI LIN

XIAO LUO

AZIZ NABIL

Madrid, December 6th , 2019

1. **Plot price vs. caratage and log(price) vs. caratage. Decide on which response variable is better to use.**

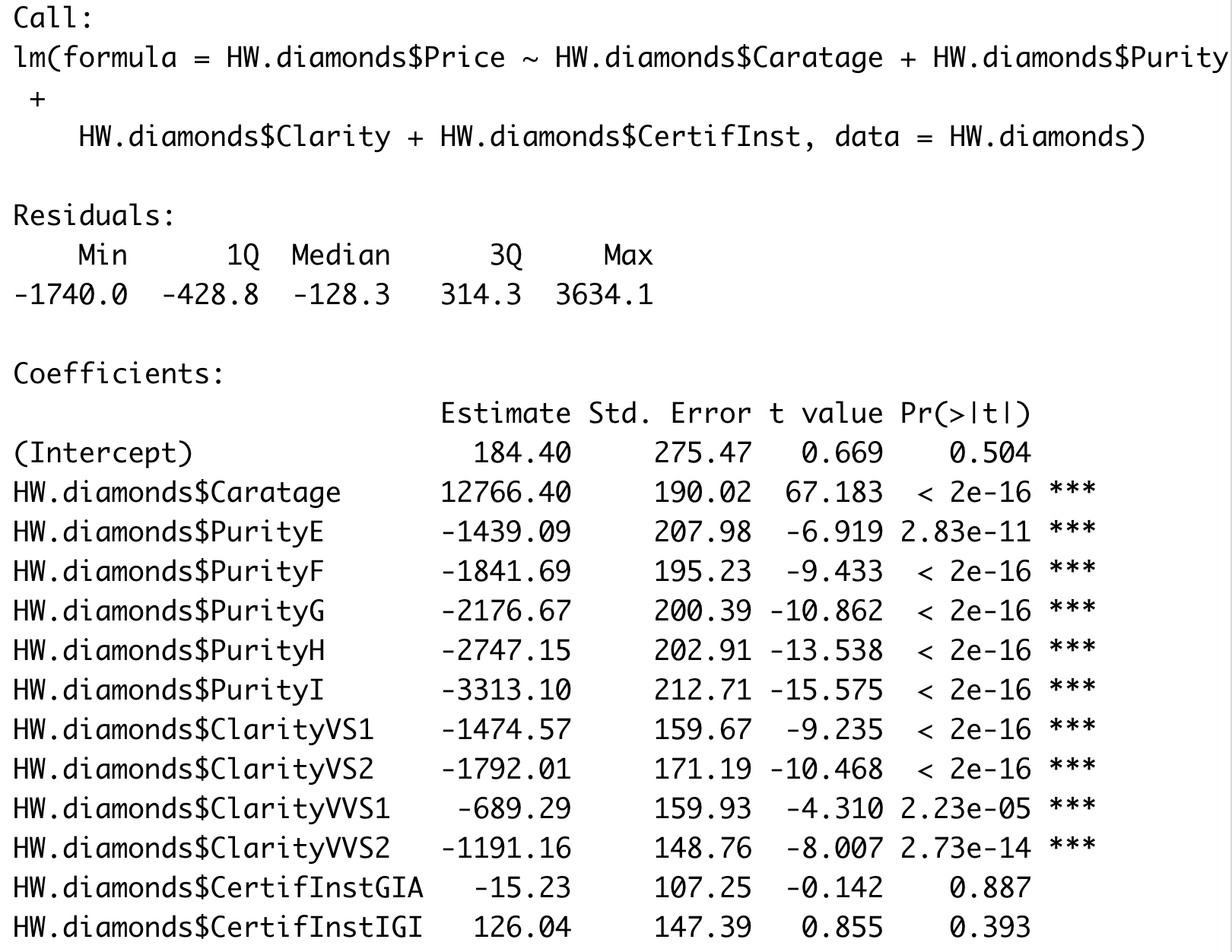


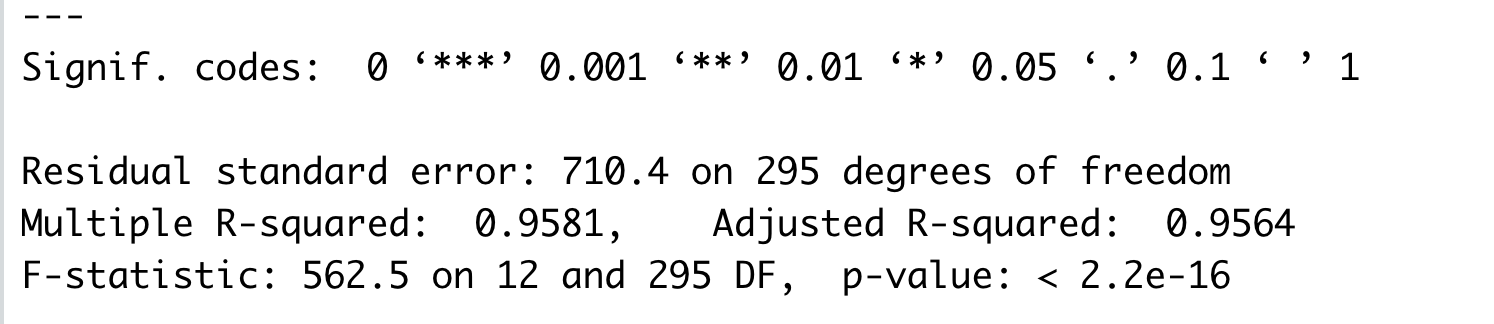
The linearity of our variable is not the best formerly ,this take us to do some transformation on some of our original variables we are going to use the logarithmic since it’s the Famous Transformation,and it allow us to induce symmetry and linearity when the variance is not constant , the two plots up allow us to compare cartage plotted against the original price and the cartage against the log of the price , in the first plot the variance seems to be concentrated in 0.4 which is smaller than 1,and the trend does not seem perfectly linear. On the same time, if we consider the log(price), the plot looks linearly nicer and the variance is reduced.

The best option is to consider the log(price) as a response variable ,but we must give attention to the interpretation because we are not considering the original measures.

The plot log(price) vs. caratage is better

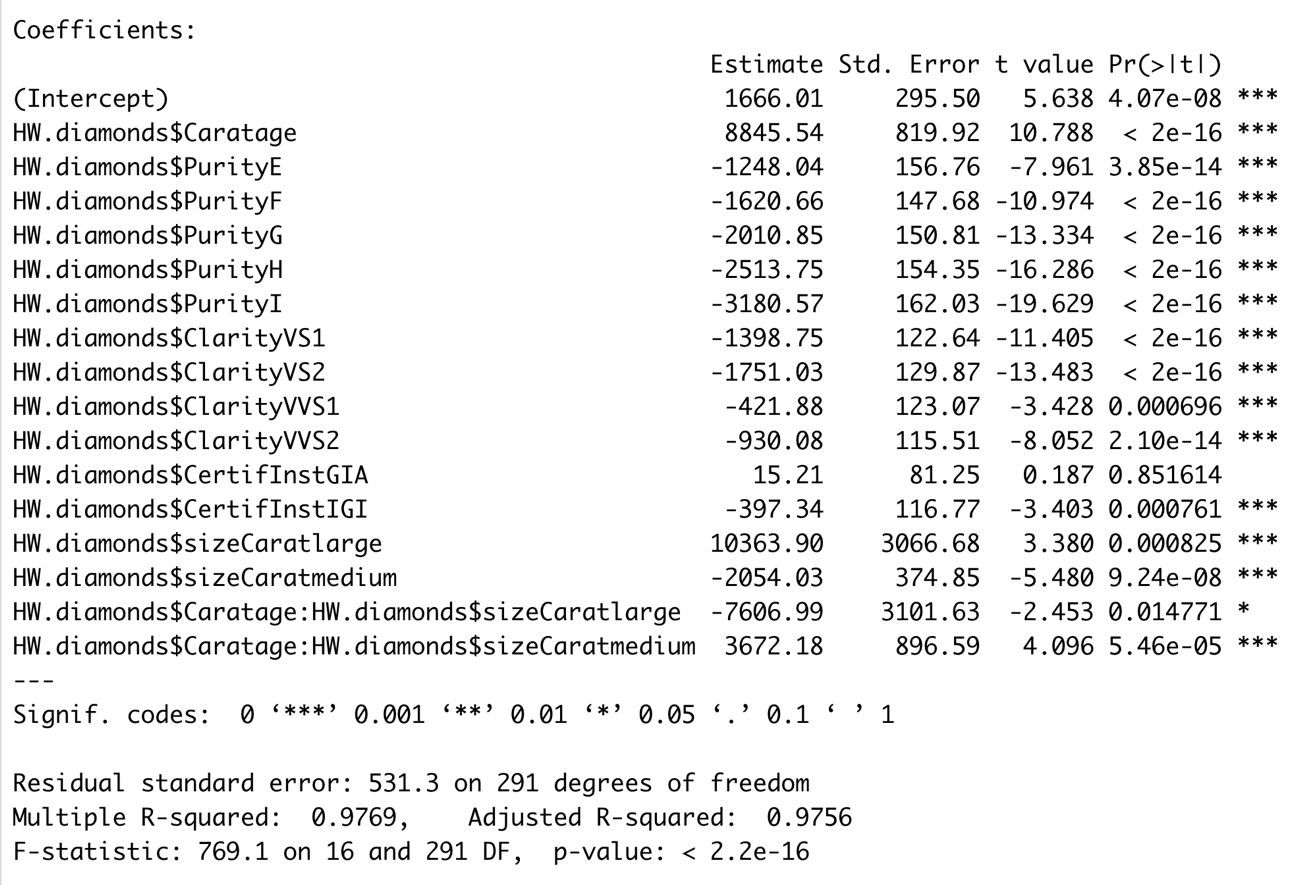
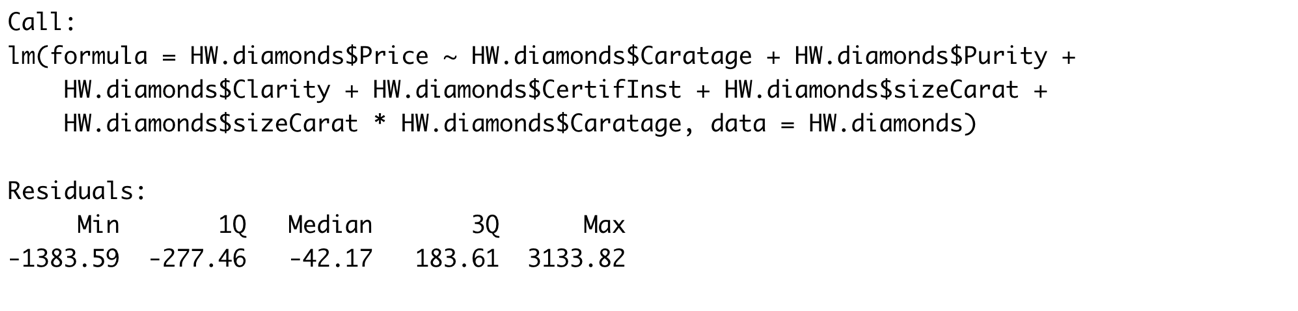
1. **Find a suitable way to include, besides caratage, the other categorical information available: clarity, color and certificate. Use the worst level of each categorical variable as the reference category and HRD for certi􏰁cation institution. Comment on the model fitted, and perform a basic analysis of the residuals (normality, constant variance, independence, you may also want to use the function outlierTest or residualPlot).**





**3.**

**3a.  Create a new categorical variable to segregate the stones according to caratage: let's say less than 0.5 carats small, 0.5 to less than 1 carat (medium) and 1 carat and over (large). Make small as the reference category. Add this new variable to the existing model as well as an interaction term between this new variable and caratage.**

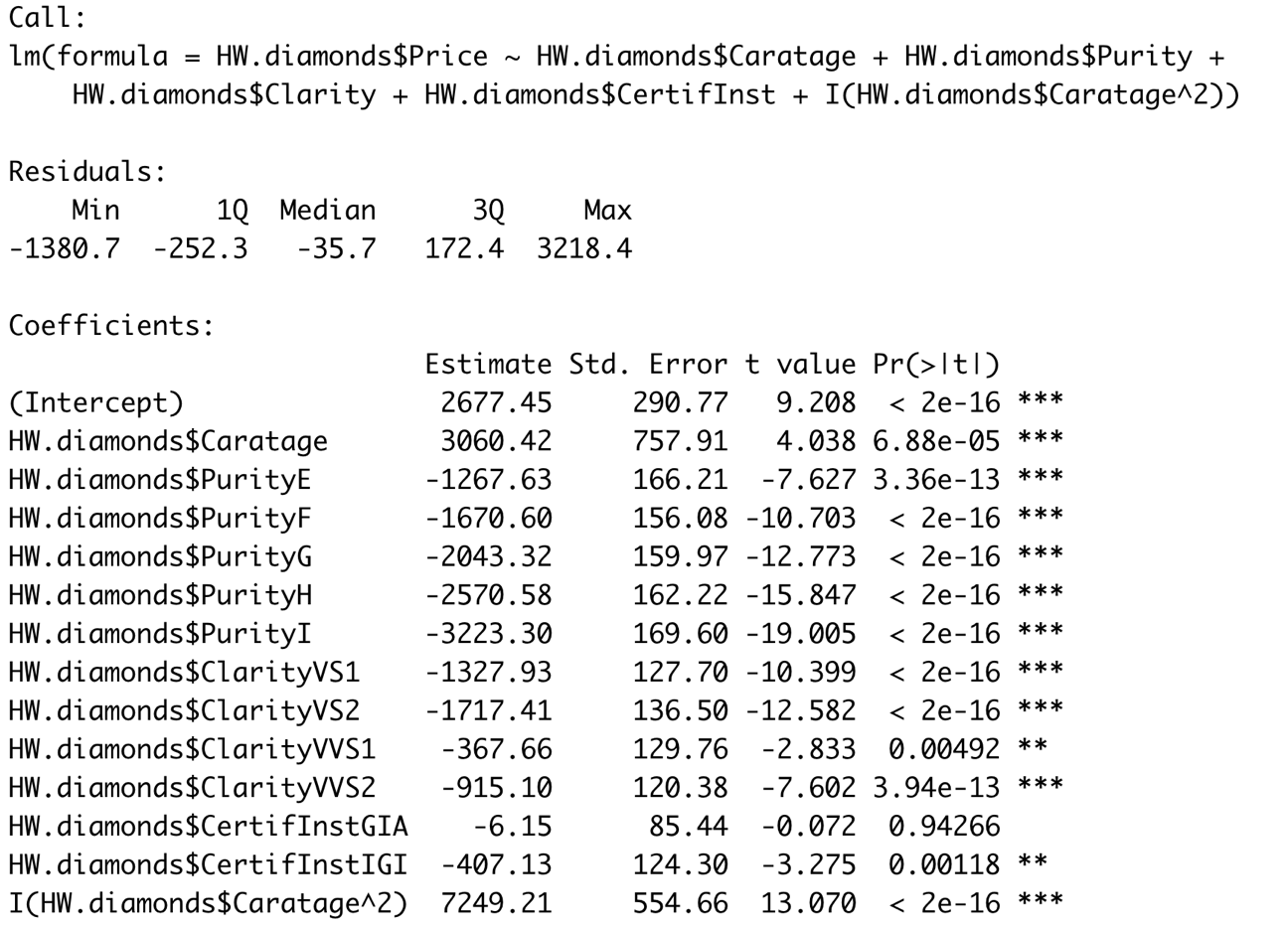


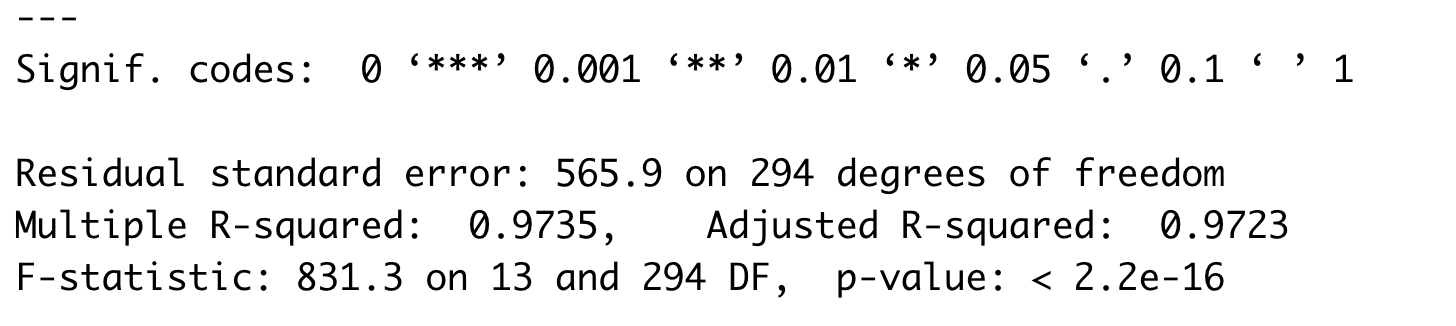
As the results shows, the multiple R-squared is very high, 0.9769, so the regression model is satisfactory.

Taking sizeCaratsmall as reference category, the incremental pricing of caratage in the 3 clusters has different mean values since the p-value of both is very low.

Colour purity and clarity are both highly valued because of very low p-value.

**3b. Include the square of carat as a new explanatory variable. It avoids the subjectivity of clusters definition.**





4. Which of the two remedial actions do you prefer and why? Think on terms of inter- pretability and validity of the assumptions.