PROJECT WORK: Shine Bright Like A Diamond

A regression model is required to support the pricing of individual diamonds. The file ShineBright.xlsx contains the individual selling price and various characteristics of 10,000 diamonds. Here is a description of the variables in the dataset: Price Carat Cut Colour Clarity Table Length Width Depth DepthPercentage

Price in US dollars
Weight in carats
1 (worst) to 5 (best), representing AGS grades fair, good, very good, premium and ideal
1 (worst) to 7 (best), representing GIA grades J to D
1 (worst) to 8 (best), representing GIA grades I1, SI2, SI1, VS2, VS1, VVS2, VVS1 and IF
Width of the top of the diamond as a percentage of its widest diameter
Length in mm
Width in mm
Depth in mm
Depth divided by average of Length and Width

(a) Split the data (80-20) for the 10,000 diamonds into a training dataset and a testing dataset.

(b) Using only the training dataset, estimate a regression model. Make sure to describe the steps of your modeling approach, and to interpret your model. Please include the following:

1. Create scatterplots of between Price and each of the independent variables.
2. If Price has an exponential relationship to Carat, Length, Width, and Depth. Taking this into consideration, transform Price and get its natural logarithm.
3. Additionally, add the logarithmic value of Carat1 to keep the linear relationship with the transformed dependent variable
4. Create a multivariable regression model and input all the variables. Provide Equation showing the initial model.
5. Provide the Table that shows the results of the model outlined in Equation
6. Describe any variables that are not significant

7. Develop new model/equation with only significant variables.

8. Provide new equation and summary table

9. Next develop a correlation matrix. Provide this as Table. Describe the variables that are highly correlated amongst each other, suggesting potential multicollinearity problems. Describe the issue. **10.** Obtain the final model shown as equation and describe the model.

11. Lastly, investigate the residuals of the model and provide graphs and description.

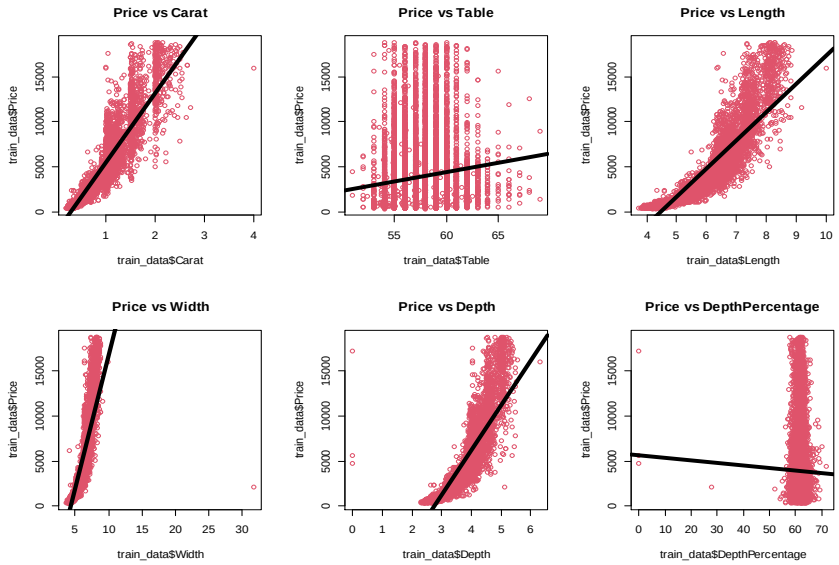
(c) Evaluate the accuracy of your model from part (b) in terms of its ability to predict the price of the diamonds in the testing dataset. Compare this with the accuracy of at least one other regression model estimated using the training dataset.

SOLUTIONS

(a) Splitting 10,000 diamonds into 80% training dataset and 20% testing dataset using 'sample()' in r. Check the R script

(b) Using training data

1. Creating Scatterplot of price with each independent variables



The graphs show that price has an exponential relationship with the dependent variables

2. Since Price has exponential relationship with the independent variable. Taking natural logarithm of price to make the samples linear
3. Also taking natural logarithm of carat. Check the R script

4. Multivariable regression model with all variables

Multivariable regression model is given as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots.. + \beta_k x_k + E$$

Hypothesis Testing is:

$H_0 = \beta_1 = \beta_2 = \beta_3$ implies; all coefficient in the model is equal to zero (no relationship between independent and dependent variables)

$H_1 = \beta_1 \neq 0$ implies; at least one coefficient is non-zero (there is a relationship)

The regression model for the price of the diamond prediction when including all the variables is:

Log(Price) = β_0 + βlog(Carat) + βCut2 + βCut3 + βCut4 + βCut5 + βColour2 + βColour3 + βColour4 + βColour5 + βColour6 + βColour7 + βClarity2 + βClarity3 + βClarity4 + βClarity5+ βClarity6 + βClarity7 + βClarity8 + βClarity + βTable + βLength + βWidth + βdepth + βDepthPercentage

Where, β_0 is the intercept, β is the coefficient of the individual variable in the model and the first level (1) of the categorical variables (Cut, Color and Clarity) is the reference for the variables and the categorical variables are recorded as dummy variable (0 and 1) for each of the levels of the categorical variables (Cut, Color and Clarity).

Thus, the initial model will be:

Log(Price) = 5.79 + 1.66*log(Carat) + 0.09*Cut + 0.11*Cut3 + 0.13*Cut4 + 0.15*Cut5 + 0.15*Colour2 + 0.28*Colour3 + 0.38*Colour4 + 0.44*Colour5 +0.48*Colour6 + 0.55*Colour7 + 0.42*Clarity2 + 0.59*Clarity3 + 0.74*Clarity4 + 0.82*Clarity5 +0.94*Clarity6 + 1.01*Clarity7 + 1.10*Clarity8 – 0.00*Table + 0.18*Length + 0.03*Width – 0.16*Depth + 0.04*DepthPercentage

5. Result of the initial model

Variable	P-value	Variable	P-value	Variable	P-value
Carat	0.001	Colour6	0.0002	Table	0.14
Cut2	0.0001	Colour7	0.0001	Length	0.0001
Cut3	0.0001	Clarity2	0.003	Width	0.0001
Cut4	0.0002	Clarity3	0.001	Depth	0.0002
Cut5	0.0002	Clarity4	0.002	DepthPercentage	0.0001
Colour2	0.0003	Clarity5	0.0001	Overall p-value	0.00001
Colour3	0.0003	Clarity6	0.0002	R^2	98%
Colour4	0.001	Clarity7	0.001		
Colour5	0.001	Clarity8	0.0001		

Initial model p-value table

The variability of the price of the diamond can be explained by weight of the carat adjusting for all other variable in the model by 98%, and since the significant value (p-value) of the overall model is less than the level of significant (0.05), ($p<0.05$), the null hypothesis for the model will be rejected and it can be concluded that the independent variables in the model have effect on the price of diamond significantly.

6. Only Table variable is statistically insignificant because its p-value (0.14) is greater than level of significance (0.05)
7. New model with significant variables will be;

$$\text{Log(Price)} = \beta_0 + \beta\text{log(Carat)} + \beta\text{Cut2} + \beta\text{Cut3} + \beta\text{Cut4} + \beta\text{Cut5} + \beta\text{Colour2} + \beta\text{Colour3} + \beta\text{Colour4} + \beta\text{Colour5} + \beta\text{Colour6} + \beta\text{Colour7} + \beta\text{Clarity2} + \beta\text{Clarity3} + \beta\text{Clarity4} + \beta\text{Clarity5} + \beta\text{Clarity6} + \beta\text{Clarity7} + \beta\text{Clarity8} + \beta\text{Clarity} + \beta\text{Length} + \beta\text{Width} + \beta\text{depth} + \beta\text{DepthPercentage}$$

8. The new model equation is;

Log(Price) = 5.61 + 1.65*log(Carat) + 0.09*Cut + 0.11*Cut3 + 0.13*Cut4 + 0.15*Cut5 + 0.15*Colour2 + 0.28*Colour3 + 0.38*Colour4 + 0.44*Colour5 +0.48*Colour6 + 0.55*Colour7 + 0.42*Clarity2 + 0.59*Clarity3 + 0.74*Clarity4 + 0.82*Clarity5

+0.94*Clarity6 + 1.01*Clarity7 + 1.10*Clarity8 + 0.19*Length + 0.03*Width – 0.17*Depth + 0.02*DepthPercentage

Variable	P-value	Variable	P-value	Variable	P-value
Cut2	0.0001	Colour7	0.0001	Length	0.0001
Cut3	0.0001	Clarity2	0.003	Width	0.0001
Cut4	0.0002	Clarity3	0.001	Depth	0.0002
Cut5	0.0002	Clarity4	0.002	DepthPercentage	0.0001
Colour2	0.0003	Clarity5	0.0001	Overall p-value	0.00001
Colour3	0.0003	Clarity6	0.0002	R^2	98%
Colour4	0.001	Clarity7	0.001		
Colour5	0.001	Clarity8	0.0001		

The coefficient of all the independent variables in the model has effect on the price of the diamond since their p-values are less than the level of significance, which accounts for statistically significant for all predictor variables in the model. The variability of the price of the diamond can be explained by weight of the carat adjusting for all other variable in the model by 98%, and since the significant value (p-value) of the overall model is less than the level of significant (0.05), ($p<0.05$), the null hypothesis for the model will be rejected and it can be concluded that the independent variables in the model have effect on the price of diamond significantly. Removing the Table variable from the model has no effect on the price of the diamond.

9. Correlation matrix table

Predictor Variables	Carat	Length	Width	Depth	DepthPercentage
Carat	1	0.98	0.94	0.97	0.01
Length	0.98	1	0.96	0.98	-0.03
Width	0.94	0.96	1	0.96	-0.08
Depth	0.97	0.98	0.95	1	0.14
DepthPercentage	0.01	-0.03	-0.08	0.14	1

The table above indicates that there is strong correlation between Carat, Length, Width and Depth which account for multicollinarity, and results to unstable and unreliable estimates of the coefficients of the correlated variable and challenge the effect of the individual independent

variables on the price of the diamond. It decrease the predictive accuracy of the model by allowing the highly correlated independent variables to duplicate information, which also challenge to their effectiveness on the model.

In order to deal with multicollinarity, there will be need for variable elimination, in such that, the highly correlated variables will be removed from the model, leaving just one among the variable.

Conceptually, let remove Length, Width and Depth from the model, leaving only Carat, since the weight of the carat has more effect on the price of the diamond than the others and being the variable of interest.

10. The final model is;

$$\text{Log(Price)} = \beta_0 + \beta\text{log(Carat)} + \beta\text{Cut2} + \beta\text{Cut3} + \beta\text{Cut4} + \beta\text{Cut5} + \beta\text{Colour2} + \beta\text{Colour3} + \beta\text{Colour4} + \beta\text{Colour5} + \beta\text{Colour6} + \beta\text{Colour7} + \beta\text{Clarity2} + \beta\text{Clarity3} + \beta\text{Clarity4} + \beta\text{Clarity5} + \beta\text{Clarity6} + \beta\text{Clarity7} + \beta\text{Clarity8} + \beta\text{Clarity} + \beta\text{DepthPercentage}$$

Log(Price) = 7.30 + 1.88*log(Carat) + 0.09*Cut + 0.11*Cut3 + 0.14*Cut4 + 0.16*Cut5 + 0.15*Colour2 + 0.28*Colour3 + 0.37*Colour4 + 0.43*Colour5 +0.47*Colour6 + 0.54*Colour7 + 0.43*Clarity2 + 0.54*Clarity3 + 0.74*Clarity4 + 0.81*Clarity5 +0.94*Clarity6 + 1.01*Clarity7 + 1.10*Clarity8 + 0.00*DepthPercentage

Variable	P-value	Variable	P-value	Variable	P-value
Cut5	0.0002	Clarity4	0.002	DepthPercentage	0.0001
Colour2	0.0003	Clarity5	0.0001	Overall p-value	0.00001
Colour3	0.0003	Clarity6	0.0002	R^2	98%
Colour4	0.001	Clarity7	0.001		
Colour5	0.001	Clarity8	0.0001		

The variability of the price of the diamond can be explained by weight of the carat adjusting for all other variable in the model by 98%, and the depthpercentage in the model is not statistically significant. Let remove it and have a model call new final model.

New final model is;

Log(Price) = 7.30 + 1.88*log(Carat) + 0.09*Cut + 0.11*Cut3 + 0.14*Cut4 + 0.16*Cut5 + 0.15*Colour2 + 0.28*Colour3 + 0.37*Colour4 + 0.43*Colour5 +0.47*Colour6 +

0.54*Colour7 + 0.43*Clarity2 + 0.54*Clarity3 + 0.74*Clarity4 + 0.81*Clarity5 +0.94*Clarity6 + 1.01*Clarity7 + 1.10*Clarity8

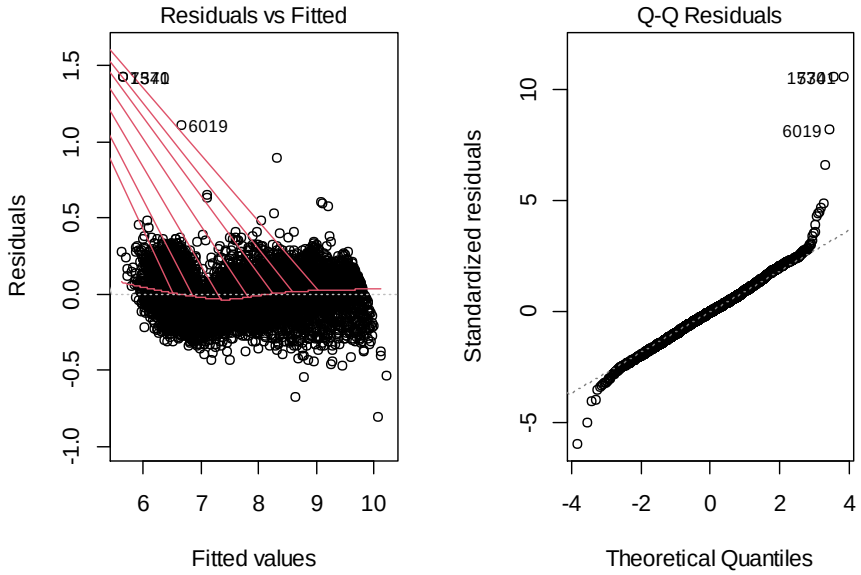
From the results, all the predictor variables is statistically significant and still explains 98% of variability in the price of the diamond. Check R Script.

Let remove cut, colour and clarity from the model to have reduced model and to check if it will fit better than the new final model.

Log(Price) = β_0 + βlog(Carat)

From the results, Carat variables is statistically significant but only explains 93% of variability in the price of the diamond and putting cut, colour and clarity has significant effect in the price of diamond with a variation of 98%. Check R Script. **The model that fit well is New Final Model.**

11. Investigating the residual plot of the new final model



From the residual plot, the assumption of regression model is met because the spread of the residuals relatively remain constant along the predicted value in the plot and the points on the normal plot form a relative straight line, which validates the assumption of regression analysis. Hence, the new final model fit better.

(c) Evaluating the accuracy of new final model and comparing with two other models (Final and Reduced model) using mean absolute error (MAE)

$$\text{MAE} = \frac{1}{n}\,\boldsymbol{\Sigma}(\boldsymbol{w} - \boldsymbol{\varpi})^{\mathbf{2}}$$

Where, w is the actual value and ϖ is the predicted value of the price

	New final model	Final model	Reduced model
MAE	0.10	0.21	0.21

The model with the lowest mean square error (MAE) fit better, therefore, the new final model is better in predicting the price of the diamond, which has only Carat, Cut, Colour and Clarity variables in the model.