A pricing model for diamond stones: The data refers to 308 stones listed in an advertisement, which educates the layerperson on the relative pricing of caratage and the different grades of clarity and colour. The goal is to come up with a multiple linear regression model for pricing the diamonds based on different variables, being caratage (1 carat = 0.2 grams) the most expected to be related to price (in Singapore dollars). But we want also to consider other information:

a) Colour purity, ordered from top to worse degrees of purity:

   a) D=top colour purity.
   b) E
   c) F
   d) G
   e) H
   f) I

b) Clarity, labeled from better to worse:

   a) IF=internal flawless.
   b) VVS1=very, very slightly imperfect 1.
   c) VVS2=very, very slightly imperfect 2.
   d) VS1=very sligthly imperfect 1.
   e) VS2=very sligthly imperfect 2.

In addition, each stone comes with a certificate by an independent institution: New York based Gemmological Institute of America (GIA), Antwerp based International Gemmological Institute (IGI) or Hoge Raad Voor Diamant (HRD). Their reputations could be a factor in the pricing of the diamond stones.

Answer the following items:

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

2. 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 certification 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. Try two different remedial actions:

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.

- Is this regression model satisfactory? Are the standard assumptions of linear regression validated? Are the numerical estimates sensible?
- Interpret the interaction parameter *med*carat*. What can we infer on the incremental pricing of caratage in the 3 clusters?
- Which is more highly valued: colour or clarity?
- All other things being equal, what is the average price difference between a grade D diamond and another one graded (a) I (b) E?
- All other things being equal, are there price differences amongst the stones appraised by the GIA, IGI and HRD?

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 interpretability and validity of the assumptions.