Stat 6021: Guided Question Set 6

For this question we will use the data set "cereals.txt". The data set contains nutritional information on 77 brands of cereal. For this question, we will focus on the variables $y = calories$ (number of calories per serving), $x_1 = sugars$ (grams of sugars per serving), and mfr indicating the manufacturer of the cereal brand. The manufacturers are coded $G =$ General Mills, $K =$ Kellogg's, and $other =$ other manufacturers.

1. Create a scatterplot of the number of calories per serving against the grams of sugars per serving for the 77 brands. Overlay separate regression lines for each of the three manufacturers. Based only on the scatterplot and the regression lines, describe the relationship between the variables of interest.

2. Create a regression model with interactions, i.e.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_1 I_1 + \beta_5 x_1 I_2 + \epsilon,$$

   where I_1 and I_2 are indicator variables where $I_1 = 1$ for a General Mills cereal and 0 otherwise, and $I_2 = 1$ for a Kellogg's cereal and 0 otherwise. Write down the estimated regression equation for this model.

3. Carry out the relevant hypothesis test to see if the interaction terms can be dropped. If they can be dropped, re-fit the regression model, and write down the estimated regression equation for the model without interactions.

4. Assess if the regression assumptions are met, for the model you will recommend to use. Also, be sure to carry out Levene's test of equality of variances since we have a categorical predictor.

5. Conduct pairwise comparisons for the difference in mean calories among all pairs of manufacturers for given values of grams of sugars, i.e.,

   (a) General Mills and other manufacturers,

   (b) Kellogg's and other manufacturers,

   (c) General Mills and Kellogg's.

Be sure to contextually interpret the results of these hypothesis tests to someone who doesn't know statistics.