Lab 6

STA 216 F19

Let’s have another look at the dataset from the first couple of labs, “lab 1 - cereal.xlsx”. Please look back at lab 1 if you need a refresher on the definitions of the variables. (It is not necessary to use any of the labels or formats we created in those labs.)

Of particular interest in this lab is the variable rating, which rates the cereal from 0 to 100. At first glance, we weren’t sure what this rating measures, but it probably isn’t taste. In lab 2, we noted that rating has a negative relationship with amount of sugar per serving and a positive relationship with the amount of fiber. Let’s also consider its relationship to other variables describing the cereals.

Part 1

We will consider how the rating is influenced by the variables sugars, fiber, fat, and sodium.

1. We discovered earlier that Quaker Oatmeal has sugars = -1 which makes no sense. So this does not adversely influence our analysis, remove this cereal from the dataset by using the line   
      
   within a DATA step.   
   Afterwards, copy and paste the last 3 rows from the output of PROC PRINT that displays only the variable name to verify that your dataset contains 76 cereals instead of the 77 in the original dataset.  
   1. Copy and paste a scatterplot matrix of the variables rating, sugars, fiber, fat, and sodium.
   2. For each of the four explanatory variables, state whether it has a positive or negative linear relationship with rating.
   3. You should notice that the scatterplots in the rows or columns containing the variable fat have points arranged in lines separated by empty space. Explain why this is.
   4. Copy and paste a correlation matrix of the same 5 variables.
   5. Report the correlation values between rating and each of the 4 explanatory variables. Which explanatory variable has the strongest correlation with rating?
   6. What about of the correlation matrix supports your answer to 2b? Explain why.
   7. Are there any strong linear relationships between the explanatory variables? Justify your answer using the correlation matrix.
2. Fit a multiple linear regression model using these 4 variables as explanatory variables.
   1. Write the equation of the least square regression line, using the variable names and (hat) where appropriate.
   2. Copy and paste the portion of the SAS output you used to obtain the equation of the line.
3. Check the assumptions for this regression model.
   1. State the 4 assumptions.
      1. Copy and paste the plot used to check 2 of these assumptions.
      2. For each of these 2 assumptions, explain what about the plot indicates that the assumption is satisfied or violated.
   2. Copy and paste the plots used to check the other assumption and explain what about the plots indicates that the assumption is satisfied or violated.
4. In checking the assumptions, you have added new variables to the dataset representing the predicted ratings and the residuals. Cinnamon Toast Crunch is yummy.
   1. What is the predicted rating for Cinnamon Toast Crunch?
   2. Show how the answer to (a) is calculated by plugging actual variable values from the dataset.
   3. What is the residual for Cinnamon Toast Crunch?
   4. Show how the answer to (c) is calculated by plugging in actual variable values from the dataset.
5. Let’s consider some healthy cereals. Blegh.
   1. What is the predicted rating for Puffed Rice? Show how this is calculated by plugging in actual variable values.
   2. Explain the connection between your answer to (a) and the estimated intercept .
   3. What is the predicted rating for Puffed Wheat? Show how this is calculated by plugging in actual variable values.
   4. Interpret the slope of fiber (from the regression line) in the context of the problem.
   5. Explain the connection between the predicted ratings of Puffed Rice and Puffed Wheat and the slope of fiber.
      1. What is the predicted rating for Shredded Wheat? Show how this is calculated by plugging in actual variable values.
      2. Explain the connection between the predicted ratings of Puffed Wheat and Shredded Wheat and the slope of fiber.
   6. What is the predicted rating for Frosted Mini-Wheats? Show how this is calculated by plugging in actual variable values.
   7. Interpret the slope of sugars in the context of the problem.
   8. Explain the connection between the predicted ratings of Shredded Wheat and Frosted Mini-Wheats and the slope of sugars.
   9. Copy and paste the portion of the SAS output that provides the (regular, not adjusted) R-squared and root MSE values.
   10. Interpret the R-squared value within the context of the problem.
   11. Interpret the root MSE value within the context of the problem.
   12. Would you say the regression model with fat, fiber, sodium, and sugars predicts rating well? Explain why or why not?

Part 2

Let’s also consider calories, carbo, potass, and protein as potential explanatory variables.

1. Use a MODEL statement in PROC REG containing all 8 explanatory variables (the four from part 1 and the four discussed above. At the end of the MODEL statement, include the option SELECTION=rsquare. That is, using pseudocode, your MODEL statement will be  
   
   1. This will fit all possible models based on including/excluding the 8 explanatory variables and sort the results by number of variables and R-squared values.  
      Copy and paste the line from the output table that gives the R-squared value and explanatory variables for our Part 1 model. (It happens to have the highest R-squared of all models with 4 x’s.)
   2. Copy and paste the line showing the variables included in the model that produce the highest R-squared value of all 5 variable models.
   3. The R-squared value in (b) is higher than the R-squared value in (a). Does this fact reliably indicate that we should add protein to the Part 1 model? Explain why or why not.
2. Changing the code in #10 to have SELECTION=adjrsq ranks all of the models according to adjusted R-squared. Which model is chosen as best according to this criterion? Copy and paste supporting SAS output.