Multiple Regression and Model Building

Open the *New_York.dat* data file using the r command "read.delim()". The data set contains demographic information about a set of towns in New York state. The response "MALE_FEM" is the number of males in the town for every 100 females. The predictors are the percentage under the age of 18, the percentage between 18 and 65, and the percentage over 65 living in the town (all expressed in percents such as "57.0"), along with the town's total population:

1. Build a regression with "MALE_FEM" as the response and all other variables as the predictors. Explain why the predictor "PCT_o65" is excluded in the model.

2. What is your conclusion regarding the significance of the overall regression?

3. What is the typical error in prediction in the model?

4. How many towns are included in the sample?

5. Which of the predictors probably do/does not belong to the model? Explain how you know this. What might be your next step after viewing this results?

6. Suppose you omit the predictor "TOT_POP" from the model and rerun the regression. Explain what will happen to the value of R-square.

7. Discuss the presence of multicollinearity. Evaluate the strength of evidence for the presence of multicollinearity. On the basis of this, should you turn to Principal Component Analysis?

8. Clearly and completely express the interpretation for the coefficients for PCT_u18. Discuss whether this makes sense.