Statistics Assignment: Health Care Cost Analysis

One key issue in health care reform is to determine what exactly is driving up the costs of health care. The `ischemic.xlsx` file provides a data set from one health insurance company regarding the total costs for services provided to 788 subscribers with ischemic (coronary) heart disease during 1998 and 1999. It is almost certain that more than one predictor is needed to explain total cost. For this problem, you will be required to examine which ones are important in helping determine the total cost. Read the entire question and write a program that will perform the required tasks. (See the ischemic data dictionary posted on the course learning management system for descriptions of the file and the variables.)

1. Read the file using an appropriate procedure.
2. Create appropriate labels for all variables.
3. Create appropriate formats for the categorical variables.
4. Check to make sure the dataset was created correctly.
5. Produce an appropriate scatter plot **MATRIX** of the predictors vs. the response variable.
   (a) Does there appear to be a linear relationship between each of the predictor variables and the response?
   (b) Does the plot suggest that some of the independent variables might be linearly related to each other? Which ones?
   (c) Report the results above as: variable so, so, and so have statistical significance, etc. **Be as brief and compact as possible. Points will be deducted for over-verbalization!**
6. Compute the pairwise correlation coefficients.
   (a) For which of the predictor variables is the correlation coefficient with the dependent variable statistically significant? **(Justify your answer by providing the appropriate p-values.)**
   (b) Does it appear that you may have a problem of (multi)collinearity? If yes, which predictor variables seem to be involved? **Justify your answer.**
   (c) Report the results above as: variable so, so, and so have significant linear relationship, etc. **Be as brief and compact as possible. Points will be deducted for over-verbalization!**
7. Fit a regression model using **ALL** predictor variables.
8. At the 0.05 level of significance, is there evidence that some of the predictors **may not be important** in determining the total cost after taking into consideration the rest of them? If yes, which ones? **Justify your answer.**
9. Using one of the methods described in class, weed out all the variables that are not contributing to the prediction of the total cost (you may need to apply multiple models until you reach your final choice).
   (a) Provide the list of variables that were not eliminated from the model. **Justify your decision to keep them in the model.**
   (b) Provide the final regression equation model in the context of the problem.

10. Provide some graphs that will help you determine if the regression line is **STATISTICALLY APPROPRIATE** for these data.

    (a) Which graphs do you need to provide?

    (b) Which statistical assumption does each of them check?

    (c) Does each graph imply the assumption was met? Yes? No? **Clearly justify your answer.**

11. Is the selected model PRACTICALLY useful for predicting the cost by the selected predictor variables? **Justify your answer.**