Stat 6021: Guided Question Set 4

For this question, we will use the data set "nfl.txt", which contains data on NFL team performance from the 1976 season. The variables are:

- y: Games won (14-game season)
- x_1: Rushing yards (season)
- x_2: Passing yards (season)
- x_3: Punting average (yards/punt)
- x_4: Field goal percentage (FGs made/FGs attempted)
- x_5: Turnover differential (turnovers acquired - turnovers lost)
- x_6: Penalty yards (season)
- x_7: Percent rushing (rushing plays/total plays)
- x_8: Opponents' rushing yards (season)
- x_9: Opponents' passing yards (season)

Instead of creating separate scatterplots for each pair of variables, you can create a scatterplot matrix. The function to use is `pairs()`. The arguments to add is a dataframe containing the variables you wish to create scatterplots for. I like to add an optional argument `lower.panel = NULL`. Try using the `pairs()` function with and without this optional argument to see which you prefer.

You can also use the function `cor` to find the correlation between all pairs of variables. The argument to add is a dataframe containing the variables you wish to find the correlation between.

1. Create a scatterplot matrix and find the correlation between all pairs of variables for this data set. Answer the following questions based on the output:

   (a) Which predictors appear to be linearly related to the number of wins? Which predictors do not appear to have a linear relationhsip with the number of wins?

(b) Do you notice if any of the predictors are highly correlated with one another? If so, which ones?

(c) What predictors would you first consider to use in a multiplie linear regression? Briefly explain your choices.

2. Regardless of your answer to the previous question, fit a multiple regression model for the number of games won against the following three predictors: the team's passing yardage, the percentage of rushing plays, and the opponents' yards rushing. Write the estimated regression equation.

3. Interpret the estimated coefficient for the predictor x_7 in context.

4. A team with $x_2 = 2000$ yards, $x_7 = 48$ percent, and $x_8 = 2350$ yards would like to estimate the number of games it would win. Also provide a relevant interval for this estimate with 95% confidence.

5. Using the output for the multiple linear regression model from part 2, answer the following question from a client: "Is this regression model useful in predicting the number of wins during the 1976 season?" Be sure to write the null and alternative hypotheses, state the value of the test statistic, state the p-value, and state a relevant conclusion. What is the critical value associated with this hypothesis test? Perform the test at 0.05 significance level.

6. Report the value of the t statistic for the predictor x_7. What is the relevant conclusion from this t statistic? Also report the critical value for this hypothesis test. Perform the test at 0.05 significance level.

7. Check the regression assumptions by creating a residual plot, an ACF plot of the residuals, and a QQ plot of the residuals. Comment on these plots.

8. Consider adding another predictor, x_1, the team's rushing yards for the season, to the model. Interpret the results of the t test for the coefficient of this predictor. A classmate says: "Since the result of the t test is insignificant, the team's rushing yards for the season is not linearly related to the number of wins." Do you agree with your classmate's statement?