Stat 6021: Guided Question Set 7

We will continue to use the "nfl.txt" data set from the tutorial for module 7.

1. From the tutorial, the model with the "best" BIC has x_2, x_7, x_8 as predictors. We will continue to work with this regression model for the rest of this question.

2. The PRESS statistic can be used in model validation as well as a criteria for model selection. Unfortunately, the `regsubsets()` function from the `leaps` package does not compute the PRESS statistic. The PRESS statistic can be written as

$$\begin{aligned} PRESS &= \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2 \\ &= \sum_{i=1}^{n} (\frac{e_i}{1 - h_{ii}})^2 \end{aligned}$$

   where h_{ii} denotes the ith diagonal element from the hat matrix.

   Write a function that computes the PRESS statistic for a regression model. **Hint**: the diagonal elements from the hat matrix can be found using the `lm.influence()` function.

3. Using the function you wrote in part 2, calculate the PRESS statistic for your regression model from part 1. Calculate the $R^2_{Prediction}$ for this model, and compare this value with its R^2. What comments can you make about the likely predictive performance of this model?

4. Delete half the observations (chosen at random), and refit the regression model. Calculate the $R^2_{Prediction}$ for this model. How well does this model predict the number of games won? **Hint**: the `sample()` function will be useful here.

5. Based on the models you fitted from the previous 2 parts, compare the standard errors of the regression coefficients for both models.

6. Using both models, calculate the predicted number of wins for all the NFL teams. Compare these models in terms of how well they predict the number of wins for all the NFL teams by computing the SS_{res} for both models.