# Guided Question Set 9

For this question, we will use the "nfl.txt" data set that we used in Guided Question Set 6. As a reminder, the data are 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 minus 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)

1. Use the `regsubsets()` function from the `leaps` package to run all possible regressions. Set `nbest=2`.
2. Identify the model (the predictors and the corresponding estimated coefficients) that is best in terms of
   (a) Adjusted $R^2$
   (b) Mallow's $C_p$
   (c) $BIC$
3. Run forward selection, starting with an intercept-only model. Report the predictors and the estimated coefficients of the model selected.

4. Run backward elimination, starting with the model with all predictors. Report the predictors and the estimated coefficients of the model selected.

5. Run stepwise regression, starting with an intercept-only model. Report the predictors and the estimated coefficients of the model selected.

6. 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 $i$th 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.

7. Using the function you wrote in part 6, calculate the PRESS statistic for your regression model with $x_2, x_7, x_8$ as predictors. 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?