Stat 6021 R Tutorial: All Possible Regressions and Automated Search Procedures

In this tutorial, we will learn how to use the `regsubsets()` function from the `leaps` package to carry out all possible regressions as well as the `step()` function for automatic search procedures. We will use the “nfl.txt” data set that we used in Guided Question Set 4. 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. All possible regressions. The `regsubsets()` function from the `leaps` package will run all possible regressions, and calculate the values of R^2, adjusted R^2, SS_{res}, Mallows C_p, and BIC. It does not calculate the PRESS statistic and the AIC. Type

   ```
   data<-read.table("nfl.txt", header=TRUE, sep="")
   attach(data)
   library(leaps)
   allreg <- regsubsets(y ~., data=data, nbest=9)
   ```

Note that you need to specify the data frame to be used. The argument `nbest` specifies the number of models you would like to consider for each number of predictors.

2. Getting the output from `regsubsets()`. In my opinion, the output from `regsubsets()` is not very visually pleasing (try typing `summary(allreg)`). I suggest creating a data frame that clearly lists the predictors being considered, and append the values of the various criteria in the data frame, i.e.

```
best <- as.data.frame(summary(allreg)$outmat)
best$p <- as.numeric(substr(rownames(best),1,1))+1
best$r2 <- summary(allreg)$rsq
best$adjr2 <- summary(allreg)$adjr2
best$mse <- (summary(allreg)$rss)/(dim(data)[1]-best$p)
best$cp <- summary(allreg)$cp
best$bic <- summary(allreg)$bic
best
```

The first line removes the quotes, making it a lot clearer to see which predictors are being considered. The second line appends a new vector containing the number of coefficients for each model. The remaining lines append vectors containing the values of R^2, adjusted R^2, SS_{res}, Mallows C_p, and BIC.

3. Finding the model that is "best" according to various criteria. We can sort, or order, our data frame to find the model with the best value for each of the criteria. For example, type

```
best[order(best$adjr2),]
```

to sort based on R^2. You can change the sorting based on other criteria accordingly. What is the best model based on each of the criteria: R^2, adjusted R^2, SS_{res}, Mallows C_p, and BIC?

4. Automated search procedures. Next we explore forward selection, backward elimination, and stepwise regression. These are performed using the `step()` function. We usually start with an intercept only model and consider the full model with all predictors as the start and end points for these automatic procedures.

```
regnull <- lm(y~1, data=data)
regfull <- lm(y~., data=data)
```

A starting and ending model have to be specified so the algorithm knows where to start and how far to go in the search procedure. To carry out the procedures, type

```
step(regnull, scope=list(lower=regnull, upper=regfull), direction="forward")
step(regfull, scope=list(lower=regnull, upper=regfull), direction="backward")
step(regnull, scope=list(lower=regnull, upper=regfull), direction="both")
```

What model(s) is chosen with these procedures? It is important to bear in mind that the same model is not always going to be selected by all of these procedures, and the starting point may influence the result.