DS 303 Homework 3
Due: Sept. 18, 2023 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Best subset selection

The data for this problem comes from a study by Stamey et al. (1989). They examined the relationship between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (`lcavol`), log prostate weight (`lweight`), `age`, log of the amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), log of capsular penetration (`lcp`), Gleason score (`gleason`), and percent of Gleason scores 4 or 5 (`ppp45`). The last column corresponds to which observations were used in the training set and which were used in the test set (`train`).

Read in the prostate data set using the following code:

```
prostate = read.table('.../prostate.data',header=TRUE)
```

In place of `'...'`, specify the pathway where you saved the dataset.

Our response of interest here is the log prostate-specific antigen (`lpsa`). We will use this data set to practice 3 common subset selection approaches.

a. **Approach 1**: Perform best subset selection on the entire data set with `lpsa` as the response. For each model size, you will obtain a 'best' model (size here is just the number of predictors in the model): M_1 is the best model with 1 predictor (size 1), M_2 is the best model with 2 predictors (size 2), and so on. Create a table of the AIC, BIC, adjusted R^2 and Mallow's C_p for each model size. Report the model with the smallest AIC, smallest BIC, largest adjusted R^2 and smallest Mallow's C_p. Do they lead to different results? Using your own judgement, choose a final model.

b. **Approach 2**: The dataset has already been split into a training and test set. Construct your training and test set based on this split. You may use the following code for convenience:

```
train = subset(prostate,train==TRUE)[,1:9]
test = subset(prostate,train==FALSE)[,1:9]
```

For each model size, you will obtain a ‘best’ model. Fit each of those models on the training set. Then evaluate the model performance on the test set by computing their test MSE. Choose a final model based on prediction accuracy. Fit that model to the full dataset and report your final model here.

c. **Approach 3**: This approach is used to select the optimal **size**, not which predictors will end up in our model. Split the dataset into k folds (you decide what k should be). We will perform best subset selection within each of the k training sets. Here are more detailed instructions:

   i. For each fold $k = 1, \dots, K$:

      1. Perform best subset selection using all the data except for those in fold k (training set). For each model size, you will obtain a ‘best’ model.

      2. For each ‘best’ model, evaluate the test MSE on the data in fold k (test set).

      3. Store the test MSE for each model.

      Once you have completed this for all k folds, take the average of your test MSEs for each model size. In other words, for all k models of size 1, you will compute their k-fold cross-validated error. For all the k models of size 2, you will compute their k-fold cross-validated errors, and so on. Report your 8 CV errors here.

   ii. Choose the model size that gives you the smallest CV error. Now perform best subset selection on the full data set again in order to obtain this final model. Report that model here. (For example, suppose cross-validation selected a 5-predictor model. I would perform best subset selection on the full data set again in order to obtain the 5-predictor model.)

Problem 2: Simulation Studies

We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

(a) Generate a data set with $p = 20$ features, $n = 1000$ observations, and an associated response vector Y generated according to the model:

$$Y = X\beta + \epsilon,$$

where β is a vector of population parameters that has some elements that are exactly equal to zero. Note you decide what β is and which are equal to 0.

(b) Split your data set into a training set containing 100 observations and a test set containing 900 observations.

(c) Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size.

(d) Plot the test set MSE associated with the best model of each size.

(e) For which model size does the test set MSE take on its minimum values? Comment on your results. If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you are generating the data in (a) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.

(f) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the regression coefficient values.

Problem 3: Cross-validation

(a) Explain how k-fold cross-validation is implemented.

(b) What are the advantages and disadvantages of k-fold cross-validation relative to:

i. The validation set approach?

ii. LOOCV?

(c) For the following questions, we will perform cross-validation on a simulated data set. Generate a simulated data set such that $Y = X - 2X^2 + \epsilon$, with $\epsilon \sim N(0, 1^2)$. Fill in the following code:

```
set.seed(1)
x = rnorm(100)
error = ??
y = ??
```

(d) Fit a linear model to the data set you simulated (`y ~ x`) and check whether or not the linearity assumption holds. Present the corresponding diagnostic plot and interpret what you observe.

(e) Set a random seed, and then compute the LOOCV errors that result from fitting the following 4 models using `lm` and `poly`:

$$\begin{aligned} M1 &: \text{a linear model with X} \\ M2 &: \text{a polynomial regression model with degree 2} \\ M3 &: \text{a polynomial regression model with degree 3} \\ M4 &: \text{a polynomial regression model with degree 4} \end{aligned}$$

You may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y.

(f) Repeat the above step using another random seed, and report your results. Are your results the same as what you got in (d). Why?

(g) Which of the models in (d) had the smallest LOOCV error? Is this what you expected? Explain your answer.

(h) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (d) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

Problem 4: Concept Review

a. For the following statement, state whether or not it is True or False. **Briefly justify your answer**.
Suppose I have 3 models to pick from:

$$\begin{aligned} M_A &: Y \sim X_1 + X_2 + X_3 + X_4 + X_5 \\ M_B &: Y \sim X_6 + X_7 + X_8 + X_9 + X_{10} \\ M_C &: Y \sim X_1 + X_2 + X_7 + X_9 + X_{10} \end{aligned}$$

Using AIC, BIC, Mallow's C_p, adjusted R^2 could lead us to pick different final models.

b. For the following statement, state whether or not it is True or False. **Briefly justify your answer**.
Suppose I have two models:

$$\begin{aligned} M_3 &: Y \sim X_1 + X_2 + X_4 \\ M_4 &: Y \sim X_4 + X_5 + X_6 + X_7 \end{aligned}$$

It must be that $\text{RSS}_3 \geq \text{RSS}_4$.

c. For the following statement, state whether or not it is True or False. **Briefly justify your answer**.

Suppose I have two models:

$$\begin{aligned} M_2 :& Y \sim X_4 + X_5 \\ M_4 :& Y \sim X_4 + X_5 + X_6 + X_7 \end{aligned}$$

It must be that $\text{RSS}_2 \geq \text{RSS}_4$.

(a) Subset selection will produce a collection of p models $M_1, M_2, \ldots, M_p$. These represent the 'best' model of each size (where 'best' here is defined as the model with the smallest RSS). Is it true that the model identified as M_{k+1} must contain a subset of the predictors found in M_k? In other words, is it true that if $M_1 : Y \sim X_1$, then M_2 must also contain X_1. And if M_2 contains X_1 and X_2, then M_3 must also contain X_1 and X_2? Explain your answer.

(b) What advantages are there to using AIC/BIC instead of using the test MSE as our model selection criteria? Explain.

(c) Suppose your colleague fit a multiple linear regression model on a dataset with response Y and p predictors $X_1, X_2, \ldots, X_p$. Their p-value for one of the predictors is 0.0647. It is not significant at $\alpha = 0.05$ so your colleague claims that the predictor is not meaningful and suggests fitting a model without this predictor. Do you agree or disagree with their claim? Justify your answer.