DS 301: Homework 5
Due: March 9, 2022 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` code or raw `R` output** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, an R file, text 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: Bias-variance tradeoff

Here we will use simulated data to better understand the bias-variance tradeoff.

a. Use the `rnorm()` function to generate a predictor X_1 of length $n = 100$.

b. Use the `rnorm()` function to generate the random error term ϵ of length $n = 100$ with mean 0 and $\sigma = 2$.

c. Generate a response Y of length $n = 100$ according to the true model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \epsilon$$

where $\beta_0, \beta_1, \beta_2$, and β_3 are parameters of your choice. Report what you decided to use for $\beta_0, \beta_1, \beta_2$, and β_3.

d. Create a data frame with your response Y and predictor X_1. You will need to use the `data.frame()` function to create a single data set. Now split your data set into a training and test set (50/50 split is fine).

e. We are going to fit 10 different models on our training set:

$$\begin{aligned}
&Y \sim X_1 \\
&Y \sim X_1 + X_1^2 \\
&Y \sim X_1 + X_1^2 + X_1^3 \\
&Y \sim X_1 + X_1^2 + X_1^3 + X_1^4 \\
&Y \sim X_1 + X_1^2 + X_1^3 + X_1^4 + X_1^5 \\
&\vdots \\
&Y \sim X_1 + X_1^2 + X_1^3 + X_1^4 + X_1^5 + \ldots + X_1^{10}
\end{aligned}$$

For each model obtain its training MSE and test MSE. You can do this in a for-loop.

f. In a single plot, plot the training MSE for each of the 10 models. Include your plot and explain what you observe.

g. In a single plot, plot the test MSE for each of the 10 models. Include your plot. Which model has the smallest test MSE? Explain what you observe in terms of the bias-variance tradeoff.

h. Generate a new response Y_{new} of length $n = 100$ according to the true model

$$Y_{\text{new}} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \beta_4 X_1^4 + \beta_5 X_1^5 + \beta_6 X_1^6 + \beta_7 X_1^7 + \epsilon$$

where $\beta_0, \beta_1, \beta_2$, and β_3 are the same as what you proposed in part (c) and $\beta_4, \ldots, \beta_7$ are constants of your choice. Repeat steps (d) - (g) using Y_{new}. How does the test MSE curve behave now that have we changed our true model? Is this what you expected? Explain your answer.

Problem 2: 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 $k = 1, \ldots, K$:

      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 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 $Y = X - 2X^2 + \epsilon, \epsilon \sim N(0, 1^2)$. Fill in the following code:

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

d. Set a random seed, and then compute the LOOCV errors that result from fitting the following 4 models using least squares:

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

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

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

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

g. 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

For the following statements, state whether or not they are True or False. **Briefly justify your answer**.

a. Suppose I have 3 models to pick from:

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

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

b. Suppose I have two models:

$$M_3 : Y \sim X_1 + X_2 + X_4$$
$$M_4 : Y \sim X_4 + X_5 + X_6 + X_7$$

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

c. Suppose I have two models:

$$M_2 : Y \sim X_4 + X_5$$
$$M_4 : Y \sim X_4 + X_5 + X_6 + X_7$$

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