DS 303 Homework 5
Due: Oct. 02, 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: Multiple Linear Regression

Suppose you work for a consulting firm. Your manager gives you the `Boston` dataset (part of `library(ISLR2)`) and asks you to build a multiple linear regression model to predict median home prices (`medv`). In typical form, he gives you no instructions or input. The final deliverable should be a model that can accurately predict `medv` from a set of predictors.

Create a report (no more than 2 pages) that includes the following items/discussion:

(a) Report your final model and its regression coefficients in a nicely summarized table.

(b) How did you justify which predictors to include in the model and which to omit? Explain your workflow and justify how you made your final selection of predictors.

(c) How did you empirically justify that your final model is good at accurately prediction of `medv`? How does it compare to other candidate models?

(d) What assumptions were needed to obtain this model? Are there any diagnostic checks you made to ensure those assumptions were valid? Did you update your model to accommodate any assumption violations?

(e) Were there any potential issues related to your model that you explored? Did you address these issues once identified?

Since you are the only data scientist on the team, your manager reminds you to make every effort to produce a report that is **rigorous** but **understandable to non-statisticians.**

Problem 2: Forward and backward selection

(a) Suppose we perform subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we can obtain p models containing $k = 1, 2, \ldots, p$ predictors. For a given k, best subset will give us a best model with k predictors. Call this $M_{k,subset}$. Forward stepwise selection will give us a best model with k predictors. Call this $M_{k,forward}$. Backward stepwise selection will give us a best model with k predictors. Call this $M_{k,backward}$. For a given k, which of these three models has the smallest training MSE? Explain your answer.

(b) Same setup as part (a). For a given k, which of these three models has the smallest test MSE? Explain your answer.

(c) We will use the `College` data set in the `ISLR2` library to predict the number of applications (`Apps`) each university received. Randomly split the data set so that 90% of the data belong to the training set and the remaining 10% belong to the test set. Implement forward and backward selection on the training set only. Do they lead you to the same model? For each approach, report the best model based on AIC. From these 2 models, pick a final model based on their performance on the test set. Report both model's test MSE and summarize your final model.

Problem 3: A Puzzling Problem

When fitting a linear regression model on a data set, you encounter the following `R` output. You notice there is something strange about the results. Point out what is strange in this output and **explain clearly** how this could happen.

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3700 -1.6364 -0.1208  1.4261  5.2558

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2936     0.5217   4.396 2.82e-05 ***
x1            1.2600     2.3006   0.548    0.585
x2            1.8968     2.5509   0.744    0.459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.376 on 97 degrees of freedom
Multiple R-squared:  0.09896,Adjusted R-squared:  0.08038
F-statistic: 5.326 on 2 and 97 DF,  p-value: 0.006385
```

Problem 4: Interaction Terms

We will use the `Credit` dataset for this problem. It is part of the library `ISLR2`.

a. This data set contains a few categorical predictors. As we already discussed in lecture, these predictors should be stored as `factors` so that `R` can handle them properly. Using the `str` function, check that all the qualitative predictors in our dataset are stored correctly in `R` as factors. Copy and paste your output.

b. Fit a model with the response (Y) as credit card balance and X_1= `Income` and X_2 = `Student` as the predictors. Call this model `fit`. Summarize your output.

c. Based on our results from part (b), write out the fitted model for students and write out the fitted model for non-students.

d. Interpret the regression coefficient related to `Income` for both models.

e. Notice that our model says that regardless of student status, the effect of `Income` on average `Balance` is the same. Do you think this is a reasonable constraint on our model? Construct some plots to back up your answer.

f. One way we could relax this assumption is by incorporating *interaction terms* into our model. Specifically:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i,$$

where X_1= `Income`, X_2 = `Student`, and X_3 = `Income` $\times$ `Student`. Fit a model with an interaction term using the following code:

```
lm(Balance ~ Income + Student + Income:Student, data=Credit)
```

Based on this model, write out the fitted model for students and write out the fitted model for non-students.

g. Interpret the regression coefficient related to `Income` for the fitted models obtained in part (f).