Stat 6021: Homework 3

1. For this question, we will use the “nfl.txt” data set. 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)

   (a) Use the `regsubsets()` function from the `leaps` package to run all possible regressions. Set `nbest=1`. Identify the model (the predictors and the corresponding estimated coefficients) that is best in terms of

      i. Adjusted R^2

      ii. Mallow’s C_p

      iii. BIC

   (b) For the models found in part 1a, use residual plots to assess if the regression assumptions are met, and address if any variables need to be transformed. If needed, transform the appropriate variable, and re-do part 1a using the transformed variables.

   (c) Run forward selection, starting with an intercept-only model. Report the predictors and the estimated coefficients of the model selected.

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

(e) 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

$$\begin{aligned} PRESS &= \sum_{i=1}^{n} [y_i - \hat{y}_{i(i)}]^2 \\ &= \sum_{i=1}^{n} (\frac{e_i}{1 - h_{ii}})^2 \end{aligned}$$

where h_{ii} denotes the ith 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.

(f) Using the function you wrote in part 1e, calculate the PRESS statistic for your regression model with x_2, x_7, x_8, x_9 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?

For the rest of the parts, we regress the number of games won against three predictors: passing yards, x_2, percent rushing, x_7, opponents' rushing yards in the season, x_8, and opponents' passing yards in the season, x_9.

(g) Create partial regression plots for this model. What are these plots telling us?

(h) Using externally studentized residuals, do we have any outliers? What teams are these?

(i) Do we have any high leverage data points for this multiple linear regression? What teams are these?

(j) Use $DFFITS_i$, $DFBETAS_{j,i}$, and Cook's distance to check for influential observations. What teams are influential?

2. The Western Collaborative Group Study (WCGS) is one of the earliest studies regarding heart disease. Data were collected from 3154 males aged 39 to 59 in the San Francisco area in 1960. They all did not have coronary heart disease at the beginning of the study. The data set comes from the `faraway` package and is called `wcgs`. We will focus on predicting the likelihood of developing coronary heart disease based on the following predictors:

- `age`: age in years
- `sdp`: systolic blood pressure in mm Hg
- `dbp`: diastolic blood pressure in mm Hg
- `cigs`: number of cigarettes smoked per day

- `dibep`: behavior type, labeled A and B for aggressive and passive respectively.

The response variable is `chd`, whether the person developed coronary heart disease during annual follow ups in the study. Read the data in. We will also randomly split the data into two: half the data will be the training data set, and the remaining half will be the test data set. We will explore the training-test split in more detail in the next module. For this exercise, perform all analysis on the training data. The code below will randomly split the data into two halves.

```
library(faraway)
Data<-faraway::wcgs
set.seed(6021) ##for reproducibility to get the same split
sample<-sample.int(nrow(Data), floor(.50*nrow(Data)), replace = F)
train<-Data[sample, ] ##training data frame
test<-Data[-sample, ] ##test data frame
```

(a) Before fitting a model, create some data visualizations to explore the relationship between these predictors and whether a middle-aged male develops coronary heart disease.

(b) Use R to fit the logistic regression model using all the predictors listed above, and write the estimated logistic regression equation.

(c) Interpret the estimated coefficient for `cigs` in context.

(d) Interpret the estimated coefficient for `dibep` in context.

(e) What are the estimated odds of developing heart disease for an adult male who is 45 years old, has a systolic blood pressure of 110 mm Hg, diastolic blood pressure of 70 mm Hg, does not smoke, and has type B personality? What is this person's corresponding probability of developing heart disease?

(f) Carry out the relevant hypothesis test to check if this logistic regression model with five predictors is useful in estimating the odds of heart disease. Clearly state the null and alternative hypotheses, test statistic, and conclusion in context.

(g) Suppose a co-worker of yours suggests fitting a logistic regression model without the two blood pressure variables. Carry out the relevant hypothesis test to check if this model without the blood pressure variables should be chosen over the previous model with all five predictors.

(h) Based on the Wald test, is diastolic blood pressure a significant predictor of heart disease, when the other predictors are already in the model?

(i) Based on all the analysis performed, which of these predictors would you use in your logistic regression model?

(j) Fit a logistic regression model based on your answer in part 2i. Based on the estimated coefficients of your logistic regression, briefly comment on the relationship between the predictors and the (log) odds of developing heart disease.

(k) Validate your logistic regression model using an ROC curve. What does your ROC curve tell you?

(l) Find the AUC associated with your ROC curve. What does your AUC tell you?

(m) Create a confusion matrix using a cutoff of 0.5. Report the accuracy, true positive rate (TPR), and false positive rate (FPR) at this cutoff.

(n) Based on the confusion matrix in part 2m, a classmate says the logistic regression at this cutoff is as good as a "no information classifier". Do you agree with your classmate's statement? Briefly explain.

(o) Discuss if the threshold should be adjusted. Will it be better to raise or lower the threshold? Briefly explain.

(p) Based on your answer in part 2o, adjust the threshold accordingly, and create the corresponding confusion matrix. Report the accuracy, TPR, and FPR for this threshold.

(q) Comment on the results from the confusions matrices in parts 2m and 2p. What do you think is happening?

3. For this question, we will revisit the `penguins` data set from the `palmerpenguins` package. The data set contains information regarding measurements of adult penguins near Palmer station, Antarctica. We will focus on using the four measurement variables (bill length, bill depth, flipper length, body mass) to model the gender of the penguins. Since there are three species involved, we also want to control for species in the logistic regression. We will not consider the island and year in this logistic regression.

   When you read the data in, notice that there are a number of penguins with missing values for gender. Remove these observations from the data frame. Also remove columns 2 and 8 since we are not considering island and year in this logistic regression. You can run the following block of code to carry out the needed steps.

```
library(palmerpenguins)

Data<-palmerpenguins::penguins
##remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]
##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
```

(a) Create some data visualizations to explore the relationship between the various body measurements and the gender of penguins. Be sure to briefly comment on your data visualizations.

(b) Use R to fit the logistic regression model. Based on the results of the Wald tests for the individual coefficients, which predictor(s) appears to be insignificant in the model?

(c) Based on your answer in part 3b, drop the predictor(s) and refit the logistic regression. Write out the estimated logistic regression equation. If you did not drop any predictor, write out the logistic regression equation from part 3b.

(d) Based on your estimated logistic regression equation in part 3c, how would you generalize the relationship between some of the body measurement predictors and the (log) odds of a penguin being male?

(e) Based on your estimated logistic regression equation in part 3c, interpret the estimated coefficient for bill length contextually.

(f) Consider a Gentoo penguin with bill length of 49mm, bill depth of 15mm, flipper length of 220mm, and body mass of 5700g. Based on your estimated logistic regression equation in part 3c, what are the log odds, odds, and probability that this penguin is male?

(g) Conduct a relevant hypothesis test to assess if the logistic regression in part 3c is a useful model. Be sure to write out the null and alternative hypotheses, report the value of the test statistic, and write a relevant conclusion.

(h) Validate your model from part 3c on the test data by creating an ROC curve. What does your ROC curve tell you?

(i) Find the AUC associated with your ROC curve. What does your AUC tell you?

(j) Create a confusion matrix using a threshold of 0.5. What is the false positive rate? What is the false negative rate? What is error rate?

(k) Discuss if the threshold should be changed. If it should be changed, explain why, and create another confusion matrix with a different threshold.

4. (You may only use R as a simple calculator or to find p-values or critical values) The data for this question are 36 monthly observations on variables affecting sales of a product. The objective is to determine an efficient model for predicting and explaining market share sales, *Share*, which is the average monthly market share for the product, in percent. The predictors are average monthly price in dollars, *price*, amount of advertising exposure based on gross Nielson rating, *nielsen*, whether a discount price was in effect, *discount* (1 if discount, 0 otherwise), whether a package promotion was in effect, *promo* (1 if promotion, 0 otherwise), and time in months, *time*.

(a) The output below is obtained after using the `step()` function using forward selection, starting with a model with just the intercept term. What predictors are selected based on forward selection?

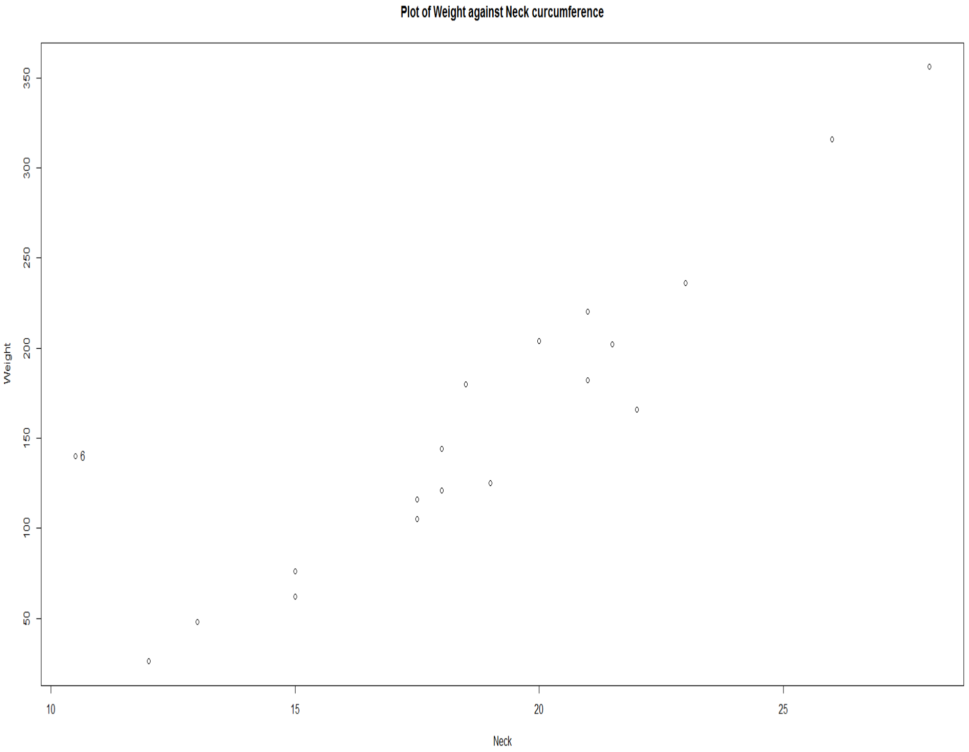
```
> start<-lm(Share˜1, data=data)
> end<-lm(Share˜.,data=data)
> result.f<-step(start, scope=list(lower=start,
+ upper=end), direction="forward")
Start:  AIC=-94.8
Share ~ 1

           Df Sum of Sq     RSS      AIC
+ discount  1   1.52953 0.91672 -128.137
+ promo     1   0.22756 2.21870  -96.318
<none>                  2.44626  -94.803
+ price     1   0.08693 2.35933  -94.105
+ nielsen   1   0.01288 2.43337  -92.993
+ time      1   0.00469 2.44156  -92.872

Step:  AIC=-128.14
Share ~ discount

          Df Sum of Sq     RSS     AIC
+ promo    1  0.086097 0.83063 -129.69
+ price    1  0.080864 0.83586 -129.46
+ time     1  0.058506 0.85822 -128.51
<none>                 0.91672 -128.14
+ nielsen  1  0.041559 0.87516 -127.81

Step:  AIC=-129.69
Share ~ discount + promo

          Df Sum of Sq     RSS     AIC
+ price    1  0.112673 0.71795 -132.94
+ time     1  0.075200 0.75543 -131.10
<none>                 0.83063 -129.69
+ nielsen  1  0.025277 0.80535 -128.80

Step:  AIC=-132.94
Share ~ discount + promo + price

          Df Sum of Sq     RSS     AIC
<none>                 0.71795 -132.94
+ time     1 0.0110210 0.70693 -131.49
+ nielsen  1 0.0003132 0.71764 -130.95
```

(b) Your client asks you to explain what each step in the output shown above means. Explain the forward selection procedure to your client, for this output.

(c) Your client asks if he should go ahead and use the models selected in part 4a. What advice do you have for your client?

5. (You may only use R as a simple calculator or to find p-values or critical values)

Data from $n = 19$ bears of varying ages are used to develop an equation for estimating *Weight* from *Neck* circumference. From a visual inspection of the scatterplot, it appears observation 6 may be an outlier.



The output below comes from fitting the linear regression model on the data.

```
##with all 19 bears
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -158.78      40.46  -3.924  0.00109 **
Neck           16.95       2.10   8.071 3.24e-07 ***

Residual standard error: 40.13 on 17 degrees of freedom
Multiple R-squared:  0.793,     Adjusted R-squared:  0.7809
F-statistic: 65.14 on 1 and 17 DF,  p-value: 3.235e-07
```

The output below comes from fitting the linear regression model on the data, with the outlier removed.

```
##with outlier removed, so 18 bears
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -234.60      25.93  -9.049 1.08e-07 ***
Neck           20.54       1.32  15.562 4.39e-11 ***
```

```
Residual standard error: 22.6 on 16 degrees of freedom
Multiple R-squared:  0.938,     Adjusted R-squared:  0.9342
F-statistic: 242.2 on 1 and 16 DF,  p-value: 4.394e-11
```

The output below displays the values of the predictor and response for the 6th observation.

```
> data[6,]
   Neck Weight
6  10.5    140
```

Some additional information from R, regarding ordinary residuals, e_i, and leverages, h_{ii} shown below, from the full data.

```
> result$residuals ##ordinary residuals
         1          2          3          4          5          6          7
-25.276933 -48.066801  22.880666  23.828133  -2.276933 120.829070 -32.803200
         8          9         10         11         12         13         14
-18.592131 -38.224400  25.249333 -21.803200 -15.119334  40.248397  34.143331
        15         16         17         18         19
 -3.593068 -33.434532   4.985732 -19.434532 -13.539598
```

```
> tmp$hat ##leverages
         1          2          3          4          5          6          7
0.05422642 0.08132161 0.06633278 0.05682064 0.05422642 0.23960510 0.05700079
         8          9         10         11         12         13         14
0.17788427 0.05278518 0.05282121 0.05700079 0.06633278 0.28626504 0.19604381
        15         16         17         18         19
0.07314261 0.09141025 0.10178713 0.09141025 0.14358291
```

(a) Calculate the externally studentized residual, t_i, for observation 6. Will this be considered outlying?

(b) What is the leverage for observation 6? Based on the criterion that leverages greater than $\frac{2p}{n}$ are considered outlying in the predictor(s), is this observation high leverage?

(c) Calculate the DFFITS for observation 6.

(d) Calculate Cook's distance for observation 6.

(e) Would you say that observation 6 is influential, based on DFFITS and Cook's distance?

(f) Briefly describe the difference in what DFFITS and Cook's distance are measuring.