Homework 8: Model Selection With Categorical Variables

FIRSTNAME LASTNAME

October 25, 2023

Instructions:

The purpose of this homework assignment is to practice doing model selection for multiple linear regression models that contain combinations of categorical and numerical predictor variables. Make sure you read each question carefully. In each question, I will give you a task to do, and I will tell you what I want you to output. You can write as much code as you want in each code chunk, but make sure you complete the task and only print the output I asked you to print. Don't sort the data unless you are told to sort the data. You should remove the "#" sign in each code chunk before writing your code. Also, if you see the comment "#DO NOT CHANGE", then I don't want you to make any modifications to that code. **You should knit your RMD file to a PDF after you answer every question.**

After you are done, knit the RMarkdown file to PDF and submit the PDF to Gradescope under HW8.

For this assignment, you will be working with a dataset found on Kaggle at this link. You need to go to the link provided to read about what each of the variables represent. This will be important for understanding the dataset. The source of the data is also cited at this link.

This data was collected for the purpose of examining the impact that student alcohol consumption has on math grades. The nice thing about this data is there are many other variables measured on these students that could potentially impact the performance in a math course. Each observation is a student and there are 395 different students listed. In the code below, I read the data into R and print out the data so you can see the content in the dataset that I named **Math**.

```
Math = read.csv("student-mat.csv")
str(Math)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

Questions

Q1 (2 Points)

The first thing I want you to do is to create a response variable in the dataset **Math** called *Grade* which is the average of the three period grades *G1*, *G2*, *G3*. After you do this, you should also remove the three original period grade variables from the dataset. Use the `str()` function on the dataset **Math** to prove that your code accomplished these tasks.

Q2 (2 Points)

Next we want to combine the alcohol related variables into a new variable called *Oalc* representing overall alcohol consumption. I want you to identify the two variables that measure workday alcohol consumption and weekend alcohol consumption. The variable *Oalc* should be the sum of these two variables. Then, remove the two original alcohol variables from the dataset and use the `str()` function to prove that this was done correctly.

Q3 (4 Points)

We want to find the best model to predict the new variable `Grade`. Use the appropriate function to fit all subset models up to $k = 10$ where k represents the number of predictors in the model and find the best model for each of those choices for k. Then, I want you to create model objects named **mod.adjRsq** and **mod.Cp*`** that are the linear models based off either maximizing adjusted R-squared or minimizing Mallow's Cp, respectively. Use the`summary()‘ function to print out the output from these two models. **For the categorical variables, if there are ANY binary indicator variables associated with those categorical variables that are significant, include the full categorical variable in the model objects. This is primarily important for variables like Mjob and Fjob where there will be multiple binary indicator variables to handle the multiple categories for Mjob and Fjob.**

I only want to see the output from the `summary()` function applied to the 2 model objects.

Q4 (2 Points)

Now, I want you to find the best model that minimizes AIC using the forward selection algorithm seen in the code from class. Create a model object called **mod.forward** from the model identified in this algorithm and output the `summary()` function applied to this model object. In your output, I want you to suppress the step-by-step process of the forward selection algorithm. You should look at the step-by-step process to understand the process the algorithm goes through to get to the final model.

The only output should be from the `summary()` function.

Q5 (2 Points)

Now, I want you to find the best model that minimizes AIC using the backward elimination algorithm seen in the code from class. Create a model object called **mod.backward** from the model identified in this algorithm and output the `summary()` function applied to this model object. In your output, I want you to suppress the step-by-step process of the backward elimination algorithm. You should look at the step-by-step process to understand the process the algorithm goes through to get to the final model.

The only output should be from the `summary()` function.

Q6 (2 Points)

Now, I want you to find the best model that minimizes AIC using the stepwise regression algorithm seen in the code from class. Create a model object called **mod.stepwise** from the model identified in this algorithm and output the `summary()` function applied to this model object. In your output, I want you to suppress the step-by-step process of the stepwise regression algorithm. You should look at the step-by-step process to understand the process the algorithm goes through to get to the final model.

The only output should be from the `summary()` function.

Q7 (2 Points)

Now, I want you to look at the output from the full model and create a model object called **mod.sig** which is the multiple linear regression model that only contains the significant predictor variables based on the individual t-tests for each coefficient. Then, use the `summary()` function to print out the output from this model. This should be your only output.

Q8 (3 Points)

Now, we want to present a table that summarizes the Adjusted R-squared ($AdjRSQ$) and AIC (AIC) for 7 models that we fit in the process of doing this assignment. You can get the adjusted r-squared from the models from the output from the summary function and you can get the AIC from the `AIC()` function applied to the model objects that you created. Replace all the NA's in the table below with the adjusted r-squared and AIC measured for each of the models listed in the table.

```
Results = data.frame(Model = c("Full","Significant","Forward","Backwards","Stepwise","Adjusted R-Squared
                     AdjRSQ = c(NA,NA,NA,NA,NA,NA,NA),
                     AIC=c(NA,NA,NA,NA,NA,NA,NA)
                     )
Results
```

```
##                Model AdjRSQ AIC
## 1               Full     NA  NA
## 2        Significant     NA  NA
## 3            Forward     NA  NA
## 4          Backwards     NA  NA
## 5           Stepwise     NA  NA
## 6 Adjusted R-Squared     NA  NA
## 7        Mallow's Cp     NA  NA
```