Homework 9: Nested F-Test and Cross Validation

Mario Giacomazzo

November 07, 2023

Instructions:

The purpose of this homework assignment is to practice the Nested F-Test and Cross Validation 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 HW9.

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 split the data into two datasets **TRAIN** and **TEST**. Typically, you would do this based off random sampling, but we will use the *school* variable to split the data. In the dataset, there are two schools represented. I want you to create **TRAIN** so that it contains all students in school "GP" and create **TEST** so that it contains all students in school "MS". After splitting up the data, I want you to remove the *school* variable from both datasets and use the `head()` function twice to show the first 6 rows of **TRAIN** and **TEST**. This should be the only output.

Q2 (2 Points)

Next we want to create a new variable called *deltaG* which is the equal to *G3* minus *G1*. This variable can be used to measure improvement. This will be our response variable for this homework assignment. After doing this, remove *G1*, *G2* and *G3* variables. You need to do this for both **TRAIN** and **TEST** sets. Finally, use the code `str(TRAIN[,29:30])` and `str(TEST[,29:30])` to show that this worked. This should be the only output.

Q3 (4 Points)

Using only the data in **TRAIN**, I want you to build a linear regression model and save the model to your global environment as an object named **mod.full**. The response variable for this model is *deltaG*. All the predictor variables I want you to include in the model, which are listed in subgroups with names reflecting what they have in common, are below:

- Demographic: (*sex*, *age*, *address*, *famsize*, *health*)
- Parental: (*Pstatus*, *Medu*, *Fedu*, *Mjob*, *Fjob*, *famsup*, *famrel*)
- Education: (*studytime*, *failures*, *higher*, *absences*)
- Freetime: (*activities*, *romantic*, *goout*)
- Alcohol: (*Dalc*, *Walc*)

The names of the subgroups will be important later in this assignment. After fitting the model, I only want to see the output from the `summary()` function applied to the model object named **mod.full**. This should be your only output.

Q4 (2 Points)

Next, I want you to fit the empty model for predicting *deltaG* to the data in **TRAIN**. Save this model to an object called **mod.empty** and use the `summary()` function to print out the output from this model. This should be the only output.

Q5 (5 Points)

In Q3, I split up the predictor variables into different subgroups and I gave these subgroups names. For each of these subgroups, I want you to create models using the data in **TRAIN** that contain all of the predictors considered EXCEPT for the variables in the subgroup. I want a model object named **mod.nodem** to contain all of the variables except the variables in the group *Demographic*. Similarly, I want you to create model objects, **mod.nopar**, **mod.noedu**, **mod.nofree**, and **mod.noalc**.

I DON'T want to see any output, but the grader should see you created each of the 5 model objects correctly

Q6 (5 Points)

Now I want you to perform Nested F-tests for each of the 5 subsets of the variables. The reduced models are the models you created in the previous question. Write code to get the p-values, and then round your p-values to four decimal places and write them in the appropriate space provided below the code chunk.

Your output, from the `anova()` function applied 5 times, should show the p-values of the 5 separate Nested F-tests and these p-values should match your rounded answers in the appropriate space provided below the code chunk

- Demographic: (REPLACE ALL CAPS WITH P-VALUE)
- Parental: (REPLACE ALL CAPS WITH P-VALUE)
- Education: (REPLACE ALL CAPS WITH P-VALUE)
- Freetime: (REPLACE ALL CAPS WITH P-VALUE)
- Alcohol: (REPLACE ALL CAPS WITH P-VALUE)

Q7 (2 Points)

Now, only using the data in **TRAIN**, I want you to create a linear regression model containing only subgroups that showed evidence that at least one of the variables in the subgroup is significant based off the results of the Nested F-tests. Save this model into an object called **mod.small** and use the `summary()` function to print out the output from this model. This should be your only output.

Q8 (3 Points)

Now, for each of the following models, I want you to calculate the out-of-sample R-squared.

- mod.empty
- mod.full
- mod.small

To do this, you need to get predictions of each of the models on the **TEST** data, calculate the correlation between the actual *deltaG* in **TEST** and the predictions of *deltaG* for each of the models, and then square the correlation for each of the models.

After you calculate the out-of-sample R-squared (R-squared on **TEST**), round the values to 4 decimal places, and write your answers in the appropriate space below the code chunk. Your code should show your calculations of the numbers that you write to verify that you calculated the out-of-sample R-squared correctly

- Out-of-Sample R-squared for **mod.empty**: (REPLACE ALL CAPS WITH R-SQUARED)
- Out-of-Sample R-squared for **mod.full**: (REPLACE ALL CAPS WITH R-SQUARED)
- Out-of-Sample R-squared for **mod.small**: (REPLACE ALL CAPS WITH R-SQUARED)

Q9 (3 Points)

Now, for each of the following models, I want you to calculate the out-of-sample RMSE.

- mod.empty
- mod.full
- mod.small

To do this, you need to get predictions of each of the models on the **TEST** data, calculate the average squared error between the predictions and actual values for *deltaG*, and then take the square root of that average.

After you calculate the out-of-sample RMSE (RMSE on **TEST**), round the values to 4 decimal places, and write your answers in the appropriate space below the code chunk. Your code should show your calculations of the numbers that you write to verify that you calculated the out-of-sample RMSE correctly

- Out-of-Sample RMSE for **mod.empty**: (REPLACE ALL CAPS WITH RMSE)
- Out-of-Sample RMSE for **mod.full**: (REPLACE ALL CAPS WITH RMSE)
- Out-of-Sample RMSE for **mod.small**: (REPLACE ALL CAPS WITH RMSE)

Q10 (2 Points)

After looking at all the models when fitted to **TRAIN** and all the models when predicted on **TEST**, which model would you recommend to researchers as the "best" model? What did we learn about the use of linear regression models to predict *deltaG* based off all of the predictor variables we considered?

Write a paragraph with 5 to 6 sentences citing numeric information from any output in the entire homework assignment to make your case for which model you would recommend and defend your opinions about what we learned about the use of linear regression models to predict *deltaG*. Remember to write in the context of the data. We are trying to see what information is useful in predicting the improvement or lack of improvement of a student's grade between the first period grade and final grade in math.

REPLACE ALL CAPS WITH A PROFESSIONALLY WRITTEN PARAGRAPH TO AN AUDIENCE WITH BASIC UNDERSTANDING OF LINEAR REGRESSION