Data Science

Homework 10

1. 한동 honor code 에 맞게 과제를 진행하여 주세요.
2. 과제의 경우 팀당 1 개의 결과물을 제출하면 됩니다.
3. 과제 제출 기한은 ~5/30 23:59 입니다. (1 분당 0.1 감점)
4. 제출은 LMS>과제 및 평가>Homework10 로 하시면 됩니다. (팀 내 1 명이 제출)
5. LMS 제출이 안되는 경우는 TA 이메일로 제출하시기 바랍니다. (22100733@handong.ac.kr)

1. Please proceed with the assignment following the Handong honor code.
2. For assignments, one submission per team is sufficient.
3. The deadline for assignment submission is until 5/30, 23:59. (0.1 points deducted per minute late)
4. Submissions should be made to LMS>Assignments>Homework10. (One member of the team should submit)
5. If you cannot submit via LMS, please submit to the TA email. (22100733@handong.ac.kr)

모든 학생들은 아래의 링크에 접속하여, 본 과제물에 대한 Peer Evaluation을 진행하여 주시기 바랍니다.
제출시간 마감 이전에 응답하지 않으면 불이익이 있을 수 있습니다.
All students are requested to access the link below and conduct a Peer Evaluation for this assignment.

01분반(KOR) - https://forms.gle/gr15nQ5b8pKgBEsd8
2rd Class(ENG) - https://forms.gle/VU2GVSdmRKgM5cyn7

Practice 10: Linear Regression

Loading data into R

```
student<- read.csv("regression_student.csv")
dim(student)
```

```
## [1] 861  32
```

prediction for student grade

You will be working with dataset containing student information who took a class of Math and Portuguese. The dataset contains following variables:

(Korean Trans1ation) 수학과 포루투갈어 수업을 수강한 학생들의 정보를 담고 있는 데이터를 사용할 것입니다.
데이터에 포함된 변수는 아래와 같습니다.

1. school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2. sex - student's sex (binary: "F" - female or "M" - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: “U" - urban or "R" - rural)
5. famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6. Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9. Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at _ home" or "other")
10. Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
11. reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12. guardian - student's guardian (nominal: “mother”, “father” or “other”)
13. traveltime - home to school travel time (numeric: 1 - < 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - > 1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)
31. class – course subject: Math or Portuguese
32. **G3 - final grade (numeric: from 0 to 20, outcome variable)**

* student 데이터를 train데이터셋(80%)과 test데이터셋(20%)으로 분류하여 진행하시기 바랍니다.
(Split the student data into training dataset (80%) and test dataset (20%) for your analysis.)

Question 1

Build a linear regression model with **lm** function to predict the final grade of student (G3) using all variables in the dataset. Describe the process, including data preparation, if necessary, to obtain your model.

학생의 최종 성적을 예측하는 선형 회귀 모델을 만들어라. (주어진 모든 변수를 사용). 모델을 만드는 과정을 설명하고, 필요하다면 전처리도 수행하고 전처리 과정도 설명하여라.

Question 2

What is the RMSE and R^2 of your model for both training and test dataset?
Attach a captured image of the evaluation web_site for your prediction on the test dataset.

1번 문제에서 학습한 모델의 RMSE와 R^2를 측정하시오.

Question 3

Interpret the linear model you got in **Q1.**
Explain what are the variables that affect the Final Grade G3, positively or negatively.
You do not have to explain every variable's influence, but only variables that you think is significant for the model.

1번 문제에서 얻은 선형 회귀모델을 해석해 보시오.
최종성적에 긍정적인 영향을 주는 변수와 부정적인 변수를 주는 변수는 무엇인가요?
모든 변수의 영향력을 다 설명할 필요는 없고, 모델에서 성적에 상당한 영향을 끼친다고 생각되는 변수만 설명하면 됩니다.

Question 4

In order to improve the model's performance, try to add new features (input variables) to the linear model or remove some variables that you might think unnecessary or irrelevant to the final grade from the model.
Try at least 3 different models, and compare their performance in terms of RMSE and R^2
Explain how changing input variables influence the model's performance overall.

모델의 성능을 개선하기 위해서, 새로운 변수를 추가하거나 필요 없거나 성적과 관련 없는 변수들을 제외해 보시오.
최소한 3개의 다른 모델을 시도해보고, 성능을 비교해 보시오 (RMSE, R^2)
입력변수를 변경하는 것이 모델의 성능에 어떤 영향을 주는지 설명해 보시오.

Question 5

Describe your best linear regression model that you have found including how you obtained the model and improved it.
What are the RMSE and R^2 of the best model?

- You may take different way to improve your linear regression model other than adding or removing variables.

위 문제에서 여러분이 얻은 best linear regression model을 설명하고,
(어떤 변수를 어떻게 사용하여 만들었는지, 어떻게 개선했는지)
성능을 기록하시오.

- best model을 얻기 위해 변수를 추가하거나 제거하는 방법 외에 다른 방법을 사용할 수도 있다.