Practice 10: Linear Regression

Loading data into R

```
load('regression_student.Rdata')
dim(student.train)
```

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

```
dim(student.test.nolabel)
```

```
## [1] 183  31
```

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 Translation) 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)**

The target variable we try to predict is **G3** which is not available in test dataset. You need to use following website to evaluate the prediction result of your regression model. The site calcuates RMSE and R^2.

예측하고자하는 목적 변수는 **G3**이며, test dataset에는 목적변수가 누락되어있다. test dataset에 대한 목적변수에 대한 RMSE와 R^2를 측정하려고하면 아래 사이트를 이용하시오.

(Click) Test Dataset Evaluation

Put your prediction result on test dataset in a RData file, and upload the file to the site.

The prediction result should be in a numeric vector named **pred_grade_test** with length of 183 which equals to the number of samples in the test dataset.

Use **save** command to include the vector of prediction result in the RData file.

The name of RData file should be named as **st[Student ID].RData**.

For example, **st21100123.RData**. otherwise the site would decline the evaluation of your work.

여러분의 test data에 대한 예측 결과를 RData 파일에 넣어서 위 사이트에 제출하면, 성능을 평가해 줄 것입니다.

결과는 **pred_grade_test**라는 이름의 numeric vector 이어야 합니다. vector의 길이는 183이어야합니다. (test data sample 수)

결과파일은 save 명령을 사용하여 RData 파일에 저장한 후 사이트에 업로드 해야하며,

RData 파일은 **st본인학번.RData** 의 형식을 지켜줘야 합니다.

예를 들어 **st21100123.RData**

그렇지 않으면 웹사이트가 여러분의 제출물을 평가할 수 없습니다.

Note: The evaluation site is only accessible in the Handong campus.

주의: 웹사이트는 한동대학교 내부 네트워크에서만 접근할 수 있습니다.

문제 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.

문제 2

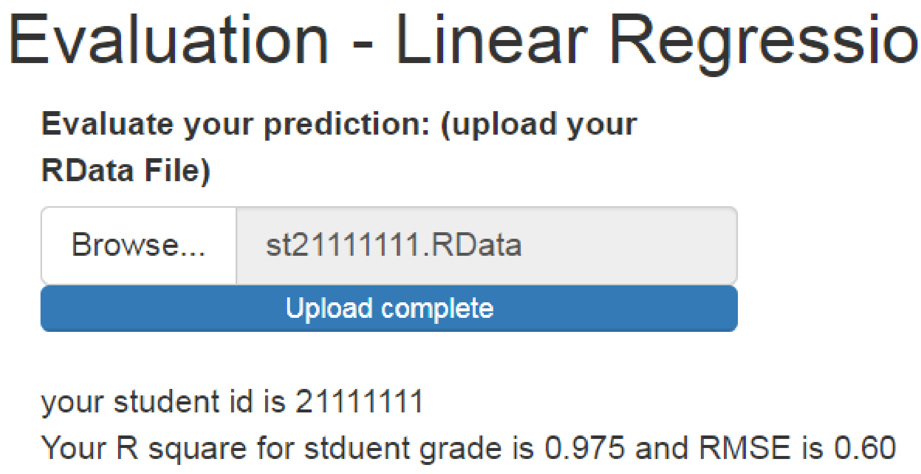
What is the RMSE and R2 of your model for both training and test dataset?

Attach a captured image of the evaluation web-site for your prediction on the test dataset.

1번 문제에서 학습한 모델의 RMSE와 R^2를 측정하시오.

test 데이터에 대한 성능은 아래와 같이 사이트의 캡처를 첨부하시오.

(How to attach a image to RMarkdown(마크다운에 이미지 첨부하기): http://a.to/21iRUrS)



Evaluation - Linear Regression Practice

Evaluate your prediction: (upload your RData File)

your student id is 21111111

Your R square for stduent grade is 0.975 and RMSE is 0.60

Figure 1: sample result

문제 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 해석해보시오.

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.

문제 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.

- For the test dataset, attach a captured image of the evaluation result.

모델의 성능을 개선하기 위해서, 새로운 변수를 추가하거나 필요없거나 성적과 관련 없는 변수들을 제외해보시오.

최소한 3개의 다른 모델을 시도해보고, 성능을 비교해보시오 (RMSE, R^2)

입력변수를 변경하는 것이 모델의 성능에 어떤 영향을 주는지 설명해보시오.

- test 데이터에 대한 성능은 사이트의 캡처를 첨부하시오.

문제 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?

- For the test dataset, attach a captured image of the evaluation result.
- 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.

- test 데이터에 대한 성능은 사이트의 캡처를 첨부하시오.
- best model을 얻기위해 변수를 추가하거나 제거하는 방법외에 다른 방법을 사용할 수도 있다.