Example: General Linear Model. AMS 580

Name:_____SBU ID:_____

Dear all, this example will teach you the necessary R procedures that you will need for Quiz #4.

General Linear Model (* multiple linear regression with at least one categorical predictor) with the Math Performance Data

The accompanying math.csv file contains data including student math scores in three school periods, demographic, social and school related features. Each case is a student, and the variables are:

1 school - student's school (binary: each student is from one of the two schools, 'GP' or 'MS')
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)

each student's Math exam scores in three periods:
31 G1 - first period grade (numeric: from 0 to 20)
32 G2 - second period grade (numeric: from 0 to 20)
33 G3 - third period grade (numeric: from 0 to 20)

Our goal is to use the first 30 variables as predictors for the last variable, **G3** (3^{rd} period math performance). Please note that we will **not** include G1 and G2 in our analysis.

You can study the following related websites for additional examples and procedures:

http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/
https://www.r-bloggers.com/2018/04/y-is-for-ys-y-hats-and-residuals/
https://www.statology.org/sst-ssr-sse-in-r/
https://www.rdocumentation.org/packages/MASS/versions/7.3-60.0.1/topics/stepAIC
https://www.rdocumentation.org/packages/leaps/versions/3.1/topics/regsubsets
http://www.science.smith.edu/~jcrouser/SDS293/labs/lab8-r.html
https://bookdown.org/egarpor/PM-UC3M/lm-ii-modsel.html

1. Please use the random seed 123 to divide the data into 75% training and 25% testing.

2. Please find the best model using the stepwise variable selection method (based on the BIC criterion) using the training data. Please (a) display the coefficients of the fitted model; (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination R^2.

3. Please find the best model using the best subset variable selection method (based on the SSE criterion) using the training data. Please (a) display the coefficients of the fitted model; (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination R^2.

4. Which model selection method among the 2 we have used above is the best? (a) Please compare the BIC of these models using the training data, as well as display these two models so we can see the parameter estimators and model goodness of fit measures. (b) Furthermore, please compare the RMSE and R^2 of these models using the test data. (c) Please discuss any modifications you can do to further improve your model(s).