# Homework 3

Your Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Using Case Study 1 data, Case1\_Student Grades\_Large.csv

Using student demographic info and learning behaviors (weekly hrs in different categories) to predict “Grade”

Note:

* Exclude nominal variables, and student performance variables (e.g., Exam) from the list of x variables
* Using hold-out evaluation only, let’s say, 75% as training set
* Use feature selection to build multiple models, and compare the models based on RMSE
* Backward method using p-value as metric
* Backward method using AIC as metric
* Forward method using AIC as metric
* Stepwise method using ACI as metric
* Best subset method using Adj-R2 as metric

Q1 Show the R coding and outputs/snapshots, as well as your explanations for each step in linear regression.

Q2 Particularly, you should answer the following questions (at the end of your homework submission)

1. Do all x variables have linear relationship with y?
2. Write down the null and alternative hypothesis of F-test, use your outputs to draw conclusions of F-test
3. Which model is the best in terms of Adj-R2?
4. Interpret the Adj-R2 in the best model above
5. Any issues in the residual analysis for the model above?
6. Which model is the best in terms of RMSE?
7. Write down the best model (by RMSE), and explain the intercepts and coefficients in the model