STAT GU4205/GR5205 Assignment 5 [20 pts]
Due 11:59pm Monday, December 9th

Please scan the written part (if you do not want to use Word or LaTeX), and submit your assignment on **Gradescope**. For coding problems, your `R` code is also required.

Note: for coding problems, even the code is submitted through Canvas, please make sure in your Gradescope submission, you include all output, plots, and necessary explanations for the output.

Problem 1: Multicollinearity, 10 points

An assistant in the district sales office of a national cosmetics firm obtained data on advertising expenditures and sales last year in the district's 44 territories. X_1 denotes expenditures for point-of-sale displays in beauty salons and department stores (in thousand dollars), and X_2 and X_3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Let Y denote sales (in thousand cases). The assistant was instructed to study the influence variables X_1 and X_2 have on sales Y. The data set `CosmeticsSales.txt` is posted on Canvas.

Use `R` to perform the following tasks:

i. Run the simple linear regression $Y \sim X_1$. Test if expenditures for point-of-sale displays in beauty salons and department stores (X_1) statistically influences sales (Y).

ii. Run the simple linear regression $Y \sim X_2$. Test if expenditures for local media advertising (X_2) statistically influences sales (Y).

iii. Now run the the full regression $Y \sim X_1 + X_2 + X_3$. Perform *marginal* t-tests to see if X_1 statistically influences sales (Y) and if X_2 statistically influences sales (Y), after controlling for the variance of X_3. Briefly compare the results to Parts i. and ii. and comment on any discrepancies. **Note: No need for Bonferroni.**

iv. You should have noticed some discrepancies in Part iii. Explain why these discrepancies are occurring and provide graphical or exploratory evidence to complement your argument.

Problem 2: Automated Variable Selection Procedures, 10 points

In this problem, we compare different variable selection methods. We study the `Credit` data set, which can be downloaded from CourseWorks. The data set records `balance` (average credit card debt) as well as several quantitative predictors: `age`, `cards` (number of credit cards), `education` (years of education), `income` (in thousands of dollars), `limit` (credit limit), and `rating` (credit rating). There are also four qualitative variables: `gender`,

`student` (student status), `status` (marital status), and `ethnicity` (Caucasion, African American or Asian). We want to fit a regression model of `balance` on the rest of the variables.

- *(Best subset selection)* The `regsubsets()` function in R (part of the `leaps` library) performs the best subsec selection by indentifying the best model that contains a given number of predictors, where *best* is defined to be the one which minimizes the residual sum-of-squares (RSS).

  Here we need to represent the qualitative predictors by dummy variables. `gender`, `student` and `ethnicity` are all two-level categorical variables, and each of them is coded by one dummy variable. `ethnicity` takes on tree values and is coded by two dummy variables. Therefore, we have 11 predictors in total.

- *(Forward stepwise selection)* We can also use the `regsubsets()` function to perform forward stepwise selection, using the argument `method=''forward''`.

- *(Backward stepwise selection)* The `regsubsets()` function can be used to perform backward stepwise selection as well (`method=''backward''`). Here we start from the full model and at each step remove a predictor which leaves a model having smallest RSS.

- *(Choosing the optimal model)* After obtaining a sequence of models by using the subsect selection approaches, we will choose a single best model which minimizes the prediction error. For this problem, we use C_p or BIC statistc as estimates of the prediction error. C_p statistic is defined by

$$C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2),$$

  where $\hat{\sigma}^2$ is an estimate of the variance of the error term. BIC is defined by

$$\text{BIC} = \frac{1}{n}(\text{RSS} + \log(n)p\hat{\sigma}^2).$$

  The summary() fuction returns RSS, C_p and BIC. You DO NOT need to compute them by yourselves.

Homework problems.

i. Apply the three subset selection methods mentioned above to Credit data set. Plot the RSS as a function of the number of variables for these three methods in the same figure.

ii. Each subsect selection method selects a sequence of models. For each approach, choose a single optimal model by using C_p and BIC statistics respectively. Report the optimal models for each approach (i.e. specify the predictors in the optimal model).

Remark. From this problem, you may notice that BIC tends to select a model with less predictors when compared to C_p.