# DSC 241 - Final Exam

Spring 2023

**Directions:**

- Read carefully and answer ALL parts of every question.
- Please write answers clearly. To receive full credit, all answers need to be properly justified. Include all relevant mathematical derivations and `R` code.
- You may consult the material from the course and other sources.
- Please use the Piazza forum to communicate privately with the instruction team if there are questions.
- *Honor code:* The work is to be done individually. Consultation with others is forbidden.

**Data:** The dataset `stackloss.csv` describes 21 days of operation of a chemical plant for oxidation of ammonia ($NH_3$) to nitric acid ($HNO_3$), a key reaction in the production of agricultural fertilizers. The variable of interest `Stack.Loss` is the percentage of the ingoing ammonia that is lost in the process. The explanatory variables are `Air.Flow` (flow of cooling air), `Water.temp` (cooling water temperature) and `Acid.Conc` (percentage concentration of circulating acid).

**Problem 1:** [25 points] Fit an ordinary least squares regression of `Stack.Loss` as a function of `Acid.Conc`.

a. Comment on the estimated value of the intercept and slope and their interpretation.

b. Let `Acid.Conc.c` be the centered variable `Acid.Conc`. Refit the model as a function of `Acid.Conc.c` instead. We shall call this model OLS1. Compare the results to those of Part (a).

c. In model OLS1, make an assessment of which model assumptions (zero mean, homoscedasticity, normality) may be violated, and their consequences for interpreting the model fit summary output. Is it possible to assess normality of the residuals?

d. Find the point with highest leverage and assess its influence over the slope parameter.

e. Obtain a bootstrap standard error and 95% bootstrap confidence interval for the slope parameter. Compare to the standard error and confidence interval estimates given by standard OLS. Explain the differences.

**Problem 2:** [20 pts] Consider a multiple regression of `Stack.Loss` as a function of `Air.Flow`, `Water.Temp` and `Acid.Conc`.

a. Evaluate the basic assumptions about the errors (zero mean, homoscedasticity, normality). Do you notice any differences with respect to Problem 1c?

b. Evaluate the degree of multicollinearity between the predictors. What effect does this have on the model fit? (Consider variance inflation factors)

c. Use $C_p$ to choose the best subset regression model out of all possible predictor subsets and the best subset LASSO fit. Compare the results from the two methods and explain.

**Problem 3:** [30 pts] Consider a generic linear model with outcome $y$ and continuous predictors $x = (x_1, \ldots, x_p)$.

a. Write a function `coeffPath = BackwardPath(x,y)` that performs backward variable selection in the following way:

   1. Start with the entire set of predictors. Set $k$ equal to the number of predictors.
   2. For a given set of $k$ predictors, fit all possible models with $k-1$ predictors. Store their coefficients and AIC values.
   3. Among the models in Step 2, choose the model with the smallest AIC.
   4. Reduce $k$ by 1 and go back to Step 2.

b. Write a function `BackwardVisualize(coeffPath)` that takes as input the output of `BackwardPath` and produces two plots: a plot of the estimated coefficients for each of the predictors, and a plot of the corresponding AIC, both as a function of the number of predictors in the model.

c. Apply your functions to the `stackloss` dataset and determine the best subset regression model according to this method. Compare your results to those of Problem 2c.

**Problem 4:** [25 pts]

a. Generate random predictors $x_1, \ldots, x_5$ independent uniformly distributed in $[-1, 1]$. Fit a multiple regression of `Stack.Loss` as a function of `Air.Flow`, `Water.Temp`, `Acid.Conc` and $x_1, \ldots, x_5$.

b. Repeat Problems 2c and 3c applied to this new dataset (including the coefficient plot) and compare your results. Does adding junk variables change the variable selection?

c. Repeat the above simulation 100 times and evaluate the proportion of times that each one of the 8 predictors is included in the final model. Does adding junk variables change the variable selection? Explain.