Statistical Modelling & Machine Learning Midterm Exam
(9:00 - 10:15AM 10/21/2021, Thursday)

- **Instruction:**
  - Solve all 3 problems.
  - Upload a file with your R code and solutions on I-campus.

1.[15pts] Consider the dataset `Q1.csv` file. The dataset has an output variable `Y` and 3 input variables `X1,X2,X3`. Suppose that we want to predict `Y` using a parametric regression model.

(1) Investigate whether there are irrelevant input variables for the prediction of Y. If they exist, find them and justify why they are irrelevant.

(2) Construct the best parametric regression model excluding irrelevant input variables found in part (1) and estimate the model parameters.

(3) Show the residual plot for the best model obtained from part (2). Based on the residual plot, justify that the model from part (2) is best.

2.[15pts] Consider the dataset `Q2.csv` file. Let X and Y be the oxygen level and the density of bacteria, respectively. Both X and Y were measured at every hour (from 1 hour to 80 hours). It is known that the density of bacteria exponentially increases as the oxygen level increases. Our goal is to predict the density of bacteria based on the oxygen level.

(1) Build a regression model and estimate the model parameters under the assumption that the errors have iid $N(0, \sigma^2)$.

(2) From part (1), obtain residuals and investigate the iid normal assumptions of errors. What assumption was violated? [**NOTE:** `dwtest()` function might not work in this problem. In that case, compute the Durbin-Watson statistic as follows:

$$dw = \frac{\sum_{t=2}^{80}(r_t - r_{t-1})^2}{\sum_{t=1}^{80} r_t^2},$$

re r_t is residual. dw has a value between 0 and 4. If dw is far from 2, it means that there is correlations between errors.]

(3) Estimate the parameters of the regression model to remedy the violated assumption found in part (2).

(4) From part (3), find the estimated covariance matrix of the error terms.

3.[10pts] Consider the dataset `Q3.csv` file. There are `Y` and 5 `X` variables (`X1,X2,...,X5`) in the dataset. `Y` variable is the number of pine trees in $1Km^2$ area and `X1,...X5` variables describe the environment of the area. Suppose that we want to predict the number of pine trees based on the 5 environmental variables. Find the best **linear model** excluding irrelevant variables [**NOTE:** For variable selection, you can use `stepAIC()` built-in function in `MASS` package].