# Midterm Examination, Data-mining, 2019

**Student Id. Number and Name:**

Visit 'e-campus' and read the samples introduced in the class. Let $y$ be the output variable (sensitivity), $x_1$ and $x_2$ be the first two categorical input variables (type and pressure), and $v_1, \ldots, v_{18}$ be the other 18 continuous input variables (V1, ..., V18) in the samples. Answer the followings.

A. (Usual forward selection) Consider the linear regression model,

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \sum_{k=1}^{18} \beta_j v_j + \varepsilon.$$

1) Write a function that returns the best model by the forward selection with respect to the AIC, keeping $x_1$ and $x_2$ in the model.

B. (Forward selection by the marginal ranking) Consider the linear regression models,

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_j v_j + \varepsilon,\ j \leq 18.$$

1) Find $|\hat{\beta}_1|$.
2) Find $|\hat{\beta}_2|$.
3) Find the order of $|\hat{\beta}_j|, j \leq 18$, in descent manner.
4) Let $z_k, k \leq 18$, be the input variable with the $k$th largest order in 3). Calculate the AIC from the model,
$$y = \alpha_0 1 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 z_1 + \varepsilon.$$
5) Calculate the AIC from the model,
$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 z_1 + \beta_2 z_2 + \varepsilon.$$
6) Calculate the order of the AIC from the models,
$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \sum_{s=1}^{k} \beta_s z_s + \varepsilon,\ k = 1, \ldots, 18$$
in descent manner.
7) Write a function that returns the best model with respect to the AIC among the models in 6).

C. (Forward selection by the joint ranking) Consider the linear regression model,

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \sum_{k=1}^{18} \beta_j x_j + \varepsilon.$$

1) Obtain $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_{18})$ from the model by using $m = n$ bootstrap samples.
2) Repeat 1) 50 times and calculate the average of each regression coefficient, and then find the orders of the absolute values of the averages.
3) Let $z_k, k \leq 18$, be the input variable with the $k$th largest order in 2). Calculate the AIC from the model,
$$y = \alpha_0 1 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 z_1 + \varepsilon.$$
4) Calculate the AIC from the model,
$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 z_1 + \beta_2 z_2 + \varepsilon.$$

5) Calculate the order of the AIC from the models,

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \sum_{s=1}^{k} \beta_s z_s + \varepsilon,\ k = 1, \ldots, 18$$

in descent manner.

6) Write a function that returns the best model with respect to the AIC among the models in 5).

D. (Model comparison) Use your own functions in A, B and C.

1) Calculate prediction errors of the three methods A, B and C by randomly splitting the whole samples into 800 training and the other test samples. The prediction error is the averages of the squared differences between the true input values and predicted input values.

2) Repeat 1) 50 times and draw the boxplots of the 50 prediction errors of the methods A, B and C.

E. (Transformation checking) Assume that we consider three transformations for four input variables $v_5$ to $v_8$.

$$x = v^2,\ x = \log(|v|),\ x = \sqrt{v}.$$

To obtain the best transformation for each variables, you should compare $4^{3+1} = 256$ models by D, applying the three methods A, B and C. If you submit the results of comparison by next Wednsday, you will have extra points in this exam.