Instructions:

1. Formula sheet will be provided to you in exam. You are not allowed to bring, with you, formula sheet or any other printed/written paper.

2. Mobiles are not allowed in exam. If you have your mobile with you, turn it off and put it under your seat so that it is visible to proctor.

3. Make sure you have 12 unique pages of exam paper (including this title page).

4. Show all the calculation steps. There are points for the steps so if your miss them, you would lose points.
Q.No.1: (3+2+2+2+5+5 = 19 points) A study is conducted to determine the effects of company size and the presence or absence of a safety program on the number of hours lost due to work-related accidents. A total of 40 companies are selected for the study. The variables are as follows:

\[ y = \text{lost work hours} \]
\[ x_1 = \text{number of employees} \]
\[ x_2 = \begin{cases} 
1 & \text{safety program used} \\
0 & \text{no safety program used} 
\end{cases} \]

a) Fit a linear regression model that regresses lost work hours \((y)\) on the number of employees \((x_1)\). This model should incorporate two lines i.e. first when safety program is used and the second when no safety program is used. Moreover, both the lines should have different intercepts and different slopes.

b) Referring to the full model from part a), should we force both the lines to have equal slopes? Why?

c) Referring to the full model from part a), write down the intercept of the line for predicting lost work hours \((y)\) based on the number of employees \((x_1)\) when safety program is used.

d) Referring to the full model from part a), write down the slope of the line for predicting lost work hours \((y)\) based on the number of employees \((x_1)\) when safety program is used.
e) Referring to the full model from part a), construct a 99% confidence interval for slope of the line for predicting lost work hours \( (y) \) based on the number of employees \( (x_1) \) when safety program is used.

f) Referring to the full model from part a), check using the Box-Cox method if there is transformation needed on the response variable. Use \( \alpha = 0.05 \) for creating an interval estimate.
Q.No.2: - (3+4+4 = 11 points) Using the Hald cement data for this question.

a) Find the variance inflation factors for all the 4 predictors. Comment on the presence of multicollinearity among the predictors.

b) Referring to part a), how will you get rid of the multicollinearity by deleting the responsible variable(s). Write down your final model with no multicollinearity.

c) Assuming that all the 4 variables should be retained, run the ridge regression using $k \in (0.01, 0.1, 0.005)$. Create a ridge trace and write down the optimal value of $k$. Also, write down the final regression equation.
Q.No.3: - (4+3+5+8 = 18 points) Data on the thrust of a jet turbine engine and six candidate regressors are given in the Excel sheet.

a) Fit a full model and calculate $R^2$, $R_0^2$ and $R_{adj}^2$. Check if there are any influential value(s) using Cook’s D statistic.

b) Using the backward elimination technique with $\alpha_{out} = 0.1$, find the final model. Write down all the details of each step.
c) Using the forward selection technique with $\alpha_{IN} = 0.1$, find the final model. Write down all the details of each step. Calculate $R^2$, $C_p$, PRESS and BIC for the final model.
d) Using the stepwise regression technique with $\alpha_{IN} = \alpha_{OUT} = 0.1$, find the final model. Write down all the details of each step. Calculate $R^2$, $C_p$, PRESS and $BIC$ for the final model. Perform the residuals analysis (residual plot, normal probability plot and time series plot) and check for the possible violation of assumptions.