MATH7606 109.12.19

**Note for the data analysis exam**

1. According to the pairwise scatterplot, you may suspect that x2 and x7 are nonlinearly related to Y, x5 and x6 are co-linear, and x3 and x4 interact with each other.

2. Also, you may find some outliers or influential points from the scatterplots. If you fit a multiple linear regression model, you can check the residuals, calculate the leverages, Cook’s distances, etc., to identify these points. You may compare the models with and without these outlying points to see their differences from various aspects of a regression model. These may include but are not limited to the regression coefficient estimates and their significance, the model fits, and R-squares changes.

3. With the observations above, you may want to check if transforming the original variables can better explain the relationship between the response and the predictor variables and improve the regression model. You may use the partial residual plots or make some trials for possible transformations. The Box-Cox method to find a proper transformation on Y is also possible. It turns out that x2^2 and log(x7) are suggested, and the interaction term x3x4 is a plus.

4. You may then do model diagnosis through the residual analysis using various techniques repeatedly to verify the model assumptions and improve your regression model. The leave-one-point-out cross-validation method can also help you to check and to compare the models.

5. Finally, the test data set provide an opportunity to validate your final model and to compare the prediction performance among the models considered.