FALL 2021 MATH 484/564 PROJECT

Due: December 5, 11:59PM, submit in Blackboard.

I. PROJECT DESCRIPTION

In this project, you will analyze the Linthurst data and identify the important physicochemical properties of the substrate influencing the aerial biomass production in the Cape Fear Estuary of North Carolina.

The response variable Y is BIO (the biomass production), and there are 14 predictor variables characterizing the soil. For instance, SAL is the percentage of salinity and pH is the acidity in the water, etc.

There are 45 observations. The first column is the index of the observation, the second column "Loc" and the third column "Type" are not used in this project.

The full multiple linear regression model is

$$Y \sim X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+X13+X14$$

- Y: BIO
- X1: H2S
- X2: SAL
- X3: Eh7
- X4: pH
- X5: BUF
- X6: P
- X7: K
- X8: Ca
- X9: Mg
- X10: Na
- X11: Mn
- X12: Zn
- X13: Cu
- X14: NH4

The project includes three parts.

A. *Part I*

Consider the 14-predictor data set (LINTHALL.txt). Use the ordinary least square estimation to estimate the regression coefficients. Run the collinearity diagnostics and identify if there is any collinearity.

B. *Part II*

Consider the 14-predictor data set (LINTHALL.txt). Use the Principle Components Regression method with collinearity reduction to decide which principle components will be included in the model. From the results of Principle Component Regression, compute the regression coefficients in the original multiple linear regression model.

C. *Part III*

In Part III, we consider a smaller data set (LINTH-5.txt) for convenience. The full multiple linear regression model is:

$$Y \sim X2 + X4 + X7 + X10 + X12$$

- Y: BIO
- X2: SAL
- X4: pH
- X7: K
- X10: Na
- X12: Zn

The data set only has 5 predictor variables, and yet it preserved some of the collinearity problem. We will use the 5-predictor data set (LINTH-5.txt) to perform a variable selection procedure.

1) Use the stepwise regression method to decide the best model. Use significance level $\alpha_E = \alpha_R = 0.15$. At each step, report the result of regression, indicate which predictor variable enters or leaves the model, and how the decision is made.

2) Use the subset selection method to decide the best two-variable model on the basis of C_p. If there is a tie, use VIF to break the tie.

II. Submission

What to submit: 1) your source code, and 2) your project report.

- Your code must be able to run directly, so submit the source file such as R file or Python file.
- Your project report will not only have numerical results, but also have adequate explanation and analysis of the results.