FALL 2023 MATH 484/564 PROJECT

- Due: Tuesday, December 5, 11:59PM.
- How to submit: via Blackboard. If you have multiple files, upload a zipped file
- Submission link will disappear after 48 hours

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. Try at least two collinearity diagnostics methods. What is the consistent conclusion you can draw from the two methods?

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 on the reduced model, compute the regression coefficients $\hat{\beta}_j$ in the original multiple linear regression model. Compare the standard error sum $\sum_j \text{s.e.}(\hat{\beta}_j)$ and SSE with their counterparts in Part I.

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 $\alpha_E = \alpha_R = 0.10$. At each step, report the result of regression, indicate which predictor variable enters or leaves the model, and how the decision is made. In the end, run the collinearity diagnostics again to verify that collinearity has disappeared.
2) Use ridge regression on the 5-predictor model, and use ridge trace to do variable selection. Refit the model that includes the remaining variables and then run the collinearity diagnostics again to verify that collinearity has disappeared.
3) Use the subset selection method to decide the best two-variable model on the basis of BIC. If there is a tie, use VIF to break the tie.

II. Submission

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

- Your code must be able to run directly.
- Your report needs to have substantial content developed beyond the source code to be considered as a report, describing the methods and the results to address each question, in particular including not only numerical results, but also adequate explanation and analysis of the results.