STAT 441/541 Statistical Methods II

Homework Assignment 4 Multiple Regression Part 2

Submit a single pdf document to the Dropbox folder *Homework Assignment 4 Multiple Regression Part 2*.

NOTE: Part 2 extends our analysis in Part 1 using the same scenario and dataset

Dataset 1

Scenario

Mercury contamination in freshwater fish has been a recognized problem in North America for over four decades. High concentrations of mercury in fish can pose a serious health threat to humans and birds. The purpose of this assignment is to investigate the relationships between mercury concentrations and selected physical and chemical lake characteristics. Data has been collected to determine if these characteristics strongly influenced the bioaccumulation of mercury in largemouth bass. The study included 53 lakes of various sizes which are hydrologically diverse.

The data set contains the following variables:

Lake ID numbers of the 53 lakes

EHg Estimated mercury concentration () for a three-year-old fish

Alk Alkalinity level of lake ()

pH Degree of acidity () or alkalinity ()

Ca Calcium level ()

Chlo Chlorophyll ()

The R code file: Mercury in Fish Parts 1 & 2 R Code.R

The dataset is Excel file: Mercury in Fish.xlsx

Multiple regression will be used with EHg as the dependent variable and Alk, pH, Ca, and Chlo as independent variables. For consistency, we will give the model in terms of the mnemonic variable names. That is, we will use the variable names EHg, Alk, pH, Ca, and Chlo in the model equation.

**(a)** Assess the goodness of fit of the regression model. Paste the summary table (below the coefficients table) from R output. Looking at the summary table:

(i) What is the estimate of the model standard deviation?

(ii) Give the value and interpretation for Adjusted R-squared:

**(b)** Assess multicollinearity of the independent variables. Paste the VIF values from R output:

(i) Give the VIF value for each independent variable and its interpretation:

(ii) Should we be concerned about multicollinearity? Justify your answer.

**(c)** Predict new *y* values using the estimated multiple regression model when Alk=80, pH=7, Ca=40, and Chlo=25:

(i) Paste and interpret the confidence interval from R output:

(ii) Paste and interpret the prediction interval from R output:

**(d)** Check assumptions for regression analysis. The assumptions are:

The model has been properly specified

The variance of the errors is for all observations

The errors are independent

The errors are normally distributed and there are no outliers

(i) Paste and interpret the Residuals vs Fitted plot from R output:

(ii) Paste and interpret the Normal Q-Q plot from R output:

(iii) Paste and interpret the Scale-Location plot from R output:

(iv) Paste and interpret the Residuals vs Leverage plot from R output:

(v) Paste and interpret the histogram of residuals from R output:

(vi) Paste and interpret the boxplot of residuals from R output:

(vii) Paste the Shapiro-Wilk test from R output and perform a hypothesis test at a significance level of :

Hypotheses:

Test Statistic:

P-value:

Decision about the null hypothesis:

Conclusion:

(viii) Paste the Breusch-Pagan test from R output and perform a hypothesis test at a significance level of :

Hypotheses:

Test Statistic:

P-value:

Decision about the null hypothesis:

Conclusion:

(ix) Are there any potential outliers identified by the influence.measures function? If so, which observations are flagged with an asterisk and why has it been flagged?

(x) Have the assumptions been met? Justify your answer.

**(e)** What is your overall conclusion about this multiple regression model?

**(f)** How would you proceed with the analysis?