ASSIGNMENT 4

MULTIPLE LINEAR REGRESSION

In this assignment you will use multiple regression tools in SPSS to examine the evidence of an association between nitrogen concentration in the rivers around the world and human population density along the rivers. You will consider different models to choose the one that adequately approximates the mean of the response as a function of the explanatory variables, and conveniently allows for the questions of interest to be investigated. Moreover, you will test the regression model assumptions with the appropriate diagnostic tools in SPSS.

River Nitrogen

The production and use of nitrogen fertilizers, the burning of fossil fuels in automobiles, power generation plants, and industries have greatly increased the amount of nitrogen cycling between the living world and the soil, water, and atmosphere. One of the major factors affecting the nitrogen cycle is greatly increased transport of nitrogen by rivers into estuaries and coastal waters where it is a major pollutant. Human populations can affect nitrogen inputs to river through industrial and automobile emissions to the atmosphere (causing the nitrogen to enter the river through rainfall), through fertilizer runoff, through sewage discharge, and through watershed disturbance.

The assignment is based on the data provided by J.L. Cole in the paper *"Nitrogen Loading of Rivers as a Human-driven Process"* in the book *"Humans as Components of Ecosystems"* (1993).

The data are available in the SPSS file *lab4.sav* located on the *STAT 337* Laboratories web site at http://www.stat.ualberta.ca/statslabs/index.htm (click *Stat 337* link). In order to download the data for the lab, click on the link *Data* for lab 4 and follow the instructions. The data are not to be printed in your submission.

The following is a description of the variables used in the data file:

Variable Name	**Description of Variable**
NO3	Nitrate concentration (in μM/l),
DISCHARG	The estimated annual average discharge of the river into an ocean (m^3/sec),
RUNOFF	The estimated annual average runoff from the watershed (in liters/(sec x km^2),
AREA	Area of watershed (in km^2),
DENSITY	Population density along the river (people/km^2),
DEP	Nitrate deposition, the flux of nitrogen from the atmosphere to the river,
NPREC	Nitrate precipitation, the concentration of nitrate in wet precipitation at sites located near the watersheds (in μmol NO3/(sec x km^2),
PREC	Mean annual precipitation (in cm/year).

The response variables is nitrate concentration (NO3). The explanatory variables include discharge, runoff, area of watershed, population density along the rivers, deposition, nitrate precipitation, and precipitation. You will study whether the response variable (NO3) is associated with the seven explanatory variables.

Answer the following questions using the data:

1. Comment on the study design. What kind of inferences can be made given the study design? Can you demonstrate using a linear regression model alone that high nitrogen concentrations in the rivers are caused by human activities? Explain briefly.

2. In this part you will examine the bivariate relationships between the variables with a matrix of scatterplots and a correlation matrix.

   (a) Obtain a matrix of scatterplots for the eight variables. Paste the scatterplot into your report. Identify the pairs of variables exhibiting strong and moderate linear relationships.

   (b) Comment on the relationship between nitrate concentration (NO3) and each of the seven predictors. If you wished to use only one explanatory variable to predict NO3, which variable would you choose? Does it look that a linear model on the original scales is appropriate for describing the relationship between NO3 and the seven predictors? Explain.

3. Now you will apply the natural logarithm transformation to each of the eight variables to make the relationship between the response variable (NO3) and each explanatory variable approximately linear, reduce spread, and neutralize outliers.

   (a) Obtain a matrix of scatterplots for the eight log-transformed variables. Check the effectiveness of the transformations by comparing the scatterplots with the corresponding scatterplots obtained in Question 2 and comment. Paste the matrix of scatterplots into your report.

   (b) Obtain the correlation matrix for the eight log-transformed variables. Paste the matrix into your report. Comment briefly. Does it look that collinearity may be a problem in this case?

4. Define a multiple regression model with log-NO3 as the response variable and the seven log-transformed explanatory variables. State the model assumptions.

5. Use the backward elimination procedure to obtain the least squares fit to the linear regression model defined in Question 4.

   (a) How many of the seven explanatory variables got eliminated by the backward elimination procedure? What is the order in which they were deleted?

   (b) What is the estimated regression equation for the final model determined by the backward elimination procedure? What percentage of the variation in log-NO3 is explained by the explanatory variables in the model?

   (c) What is the p-value of the test for overall significance of the regression model? Moreover, comment on how significantly each explanatory variable contributes individually, given the other variables in the model and using $\alpha = 0.05$.

6. In this part you will obtain some diagnostic plots to verify the assumptions for the regression model obtained in the previous question.

   (a) Obtain a plot of residuals versus the fitted values for the model obtained in Question 5. Paste the plot into your report. Is there evidence that the variance of the residuals increases with increasing fitted values or that there are any outliers? Explain briefly.

   (b) Obtain a normal probability plot of standardized residuals for the model obtained in Question 5. Paste the plot into your report. Is there any evidence that the assumption of normality is violated? Comment briefly.

In Questions 7 and 8 you will attempt to determine the extent to which the human effect is from pollutants discharged into the river directly as opposed to those discharged indirectly through atmospheric pollution. Moreover, you will use some case-influence statistics to identify the influential cases.

7. Now you will study, among others, the effect of acid rain on the levels of NO3 in the rivers. Is there convincing evidence of an association of nitrate concentration (NO3) with deposition or nitrate precipitation after accounting for discharge, runoff, precipitation, and area (all variables on the log-scale)?

   (a) As the first step to answer the above question, use the forward selection procedure to obtain a regression model with NO3 as the response variable and discharge, runoff, area, and precipitation as the initial explanatory variables (all variables on the log-scale). Report the R^2 value of the model.

   (b) Obtain the case-statistics (studentized residuals, Cook's distance, and leverages) to identify influential case(s) (if any) for the model in part (a). Provide the case statistics for the influential case(s). Paste the plot of residuals versus fitted values (standardized) into your report and mark clearly the influential observation(s) in the plot. Rerun the forward regression without the case(s). Report the estimated regression equation and the R^2 value for the model. Comment briefly.

   (c) Define a new model by adding deposition and nitrate precipitation (both variables on the log-scale) to the regression model obtained in part (b). Run regression to estimate its regression coefficients. State the above question about the association of NO3 with deposition or nitrate precipitation as null and alternative hypotheses about the regression coefficients (accounting *only* for the variables remaining after the forward selection as per part (b)). Use appropriate SPSS computer outputs and hand calculations to obtain the value of a test statistic and the corresponding p-value. State your conclusions.

8. Is there a further effect of population density on nitrogen concentration, after accounting for the effects of nitrogen deposition and nitrogen precipitation? In order to answer the question, use the model obtained in part (c) of Question 7 to see whether adding density (on log scale) is significant. In particular, state the hypotheses, report the value of the test statistic, the p-value of the test, and give your conclusion.

LAB 4 ASSIGNMENT: MARKING SCHEMA

Question 1 (5)

Study design: 3 points
Cause-and-effect conclusions: 2 points

Question 2 (15)

(a) Matrix of scatterplots: 3 points
Pairs of variables exhibiting strong and moderate linear relationship: 2 points each (4 points total)
(b) Relationship between NO3 and each of the seven predictors: 4 points
(c) The explanatory variable with the highest correlation with the response: 2 points
The need for a log-transformation: 2 points

Question 3 (11)

(a) Matrix of scatterplots: 3 points
Comments about the effectiveness of the log-transformation: 2 points
(b) Correlation matrix: 3 points
Comments about the bivariate correlations: 2 points
Collinearity: 1 point

Question 4 (4)

Multiple regression model: 2 points
Assumptions: 2 points

Question 5 (14)

(a) Number of explanatory variables eliminated: 2 points
The order of elimination: 2 points
(b) The estimated regression surface: 3 points
Percentage of variation in NO3: 2 points
(c) The p-value of the test for the overall significance: 2 points
t-tests: 3 points

Question 6 (10)

(a) Plot of residuals versus the fitted values: 3 points
Comments about the pattern in the residual plot: 2 points
(b) Normal probability plot: 3 points
Comments about the pattern in the plot: 2 points

Question 7 (31)

(a) R^2 value: 3 points
(b) Influential case and its case statistics: 3 points
Scatterplot of residuals versus fitted values (influential circled): 3 points
Estimated regression equation: 3 points
R^2 value: 3 points
(c) Stating the new model with deposition and nitrate precipitation added: 2 points
Estimated regression coefficients: 2 points
Null and alternative hypotheses: 3 points

The value of the F-statistic: 5 points
The estimated p-value: 2 points
Conclusions: 2 points

Question 8 (10)

Null and alternative hypotheses: 3 points
The value of the t or F test statistic: 3 points
P-value: 2 points
Conclusions: 2 points

TOTAL = 5 + 15 + 11 + 4 + 14 + 10 + 31 + 10 = 100