Harsha Somaya and Jane Maguire
SC321
4/26/2024

Data Analysis Report #2

Introduction:

This data comes from a 1993 study on how human population density relates to river nitrates. This dataset uses a random sample of 42 rivers around the world. Seven different variables were recorded for each river. The variables recorded are nitrate concentration, density of people, annual average discharge of the river into an ocean, annual average runoff from the watershed, watershed area, nitrate precipitation, and precipitation. The objective of this research is to identify a model that can be used to predict nitrate concentration in a given river.

Methods:

This dataset consists of seven variables: NO3, Density, Discharge, Runoff, Area, NPrec, and Prec as defined in Table 1. To conduct our statistical analysis, we transformed existing variables in order to create an effective model; defined in Table 2. The objective of predicting nitrate concentration in a given river will be carried out using a third order polynomial model. The model was found using the interaction terms indicated by Regression Trees and t tests. A logistic transformation was performed on the response variable of NO3 to satisfy the conditions of multiple linear regression. We also conducted an anova test to compare this model to our simpler models. Analysis will be carried out in R-studio.

Table 1: Shows the variable name and definition for each variable in the dataset

Variable Name:	**Definition**:
NO3	Nitrate concentration (mcmol/L)
Density	Density of people (people/km^2)
Discharge	Annual average discharge of the river into an ocean (m^3/sec)
Runoff	Annual average runoff from the watershed (L/(sec*km^2))
Area	Watershed area (km^2)

NPrec	Nitrate Precipitation (mcmol/(sec*km^2))
Prec	Precipitation (cm/year)

Table 2: Shows the variable name and definition for each of the modified variables used for the model

Variable Name:	**Definition**:
ln.NO3	**The natural logarithm of** Nitrate concentration (mcmol/L). Calculated as ln.NO3 = ln(NO3)
Density ^2	Density of people (people/km^2) squared
Density ^3	Density of people (people/km^2) Cubed

Results:

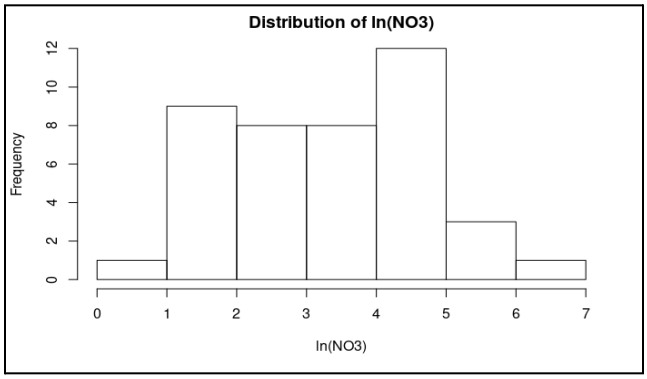
Table 3 shows descriptive summary statistics for the variables in the model, and Figure 1 shows the distribution of ln.NO3, the response variable. After running the regression trees, we found a model that included the main effect of density, NPrec and runoff along with the interaction terms of density with itself, with runoff, and with NPrec. However, we wanted to see if we can simplify this complicated model. Hence, we fitted a model with just the main effects of the form NO3 ~ Density + Discharge + Runoff + Area + NPrec + Prec and found that density was the only significant variable with a p-value of 8.8e-07 from the t test. Hence, we kept density in our model. We plotted NO3 versus density in figure 3. Next, we wanted to see if either density^2, runoff, runoff:density, NPrec or Nprec:density was needed. We ran an anova to compare two models with the following explanatory variables: one with only density and the other with the main effects of runoff, NPrec and density and the interaction terms of density^2, density:runoff, and NPrec:density. The latter is defined as the complicated model. We found that going from the only density model to the complicated one leads to a p value of `1.6e-08`. Since this is a significant p value, we wanted to see if we can simplify this model to only density and density^2 since that is simpler. We checked the summary of the complicated model and found significant p values from the t test for density^2 (0.0357), density:Nprec (0.0071), and density:runoff (0.0148), but not for the main effects of density, NPrec, or runoff. Hence, we are wary of using this complicated model that has

significant interaction terms but insignificant main effects. We check if we can instead use the simpler model of NO3~density + density ^2 since we know density is an important term to keep, and density^2 is suggested by the regression tree. The R^2 squared for this model is high at 0.826, and the density^2 term is significant. Hence, perhaps we did not lose much predictive power by using this simple model. However, the conditions of linear regression for linearity and equal variance were not satisfied. Hence, we took the natural log of NO3. We reran our regression tree with the transformed ln.NO3 on all the variables, which again indicated a model of ln.NO3~density + density^2. However, the conditions were still not satisfied. Hence, we tried adding a density cubed term, which did satisfy the conditions. The anova test on the model of ln.NO3 vs. density, density^2, and the one with the addition of density^3 confirms this is an appropriate model to use because the p-value for this model is significant (0.0042). We plotted ln.NO3 versus density, density^2, and density^3 in Figure 2.

Table 3: *summary statistics (mean, standard deviation, median, and IQR) for NO3 , density, density^2, and density^2*

Variable Name:	**Mean**	**Standard Deviation**	**Median**	**IQR**
ln.NO3	3.3216	1.3622	3.3842	1.9709
Density	85.36	104.31	32.50	105.5
Density ^2	17908	33478	1062	13301
Density ^3	4748713	11626802	34938	1560302

Figure 1: Distribution of ln(NO3) in ln(mcmol/L)



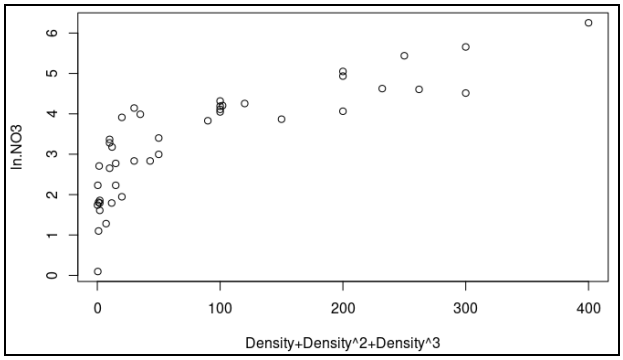


Figure 2: ln.NO3 vs Density + Density^2 + Density^3

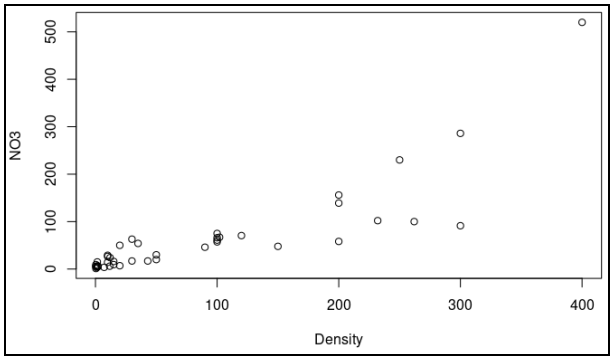


Figure 3: NO3 vs Density

Discussion:

In conclusion, we found that the best model to predict nitrate concentration in a given river is ln.NO3 ~ density + density^2 + density^3. The p-value for density, density^2 and density^3 were statistically significant from the t test. The third order polynomial model is useful. After testing assumptions of the full main effects model, the main effects suggested by the regression tree with their interaction terms, and the second and third order polynomial models, we found that the model that proved to be the most statistically significant was the third order polynomial model. We could have considered the 4th and 5th order polynomial models but chose to maintain this simpler model for interpretability and to avoid overfitting. The model ln.NO3 ~ density + density^2 + density^3 suggests a non-linear relationship between population density and nitrate concentration in rivers. The inclusion of the density^2 and density^3 terms acknowledges the possibility that as density ^2 or density^3 increases (or in general as density increases), the ln(N03) starts to stabilize and plateau. This could be due to certain environmental factors, such as the rivers can only absorb ln.NO3 up to a certain amount. Also, to obtain the predicted NO3 concentrations, the output from this model will become an exponent for base e to undo the natural log transformation on the response variable. This calculation is the predicted NO3 concentration. A limitation is that while it has a good predictability, its generalizability to other rivers may be influenced by unmeasured variables that could influence nitrate concentration, such as the number of surrounding farms. This highlights the need for caution when extrapolating to new rivers. In future

research, we may want to explore new variables to see if there are other impacts on NO3 concentration that are not accounted for in this dataset. In conclusion, the best model that can be used to predict nitrate concentration in a given river is ln.NO3 ~ density + density^2 + density^3.