**Analytical Report**

1. **Executive summary**

This report entails preparation of biostatistics data on Dolphins to fit a GLM (Generalized linear model) to the data. The objective is to create GLMs to analyze toxin concentrations in different demographic groups of dolphins. The methods applied include descriptive statistics to explore the trend of the data. There is imputation of missing data using multiple imputation method. Then a logistic regression is fitted to the complete imputed dataset.

Findings from the report:

* Findings of the summary statistics show that there were more male dolphins, more adult dolphins, more North-Carolina stranded dolphins, and most dolphins without Morbillivirus.
* Another finding was that there was similar mercury (Hg) toxin concentration between; (male and female, Florida-stranded and North Carolina-stranded dolphins) also both affected and unaffected dolphins by Morbillivirus had same mercury (Hg) toxin concentration.
* However, juvenile dolphins had different mercury concentrations from adult dolphins.
* The total number of missing values in the data was 206, missing not completely at random in the following variables; "PCB\_1268", "DEP", "NPE", "Triclosan", "BPA", "Atrazine", "Hg", "Cu", "Se", "Cd", "Tl", "Pb", Mn", "Co", "Zn", "As", "Fe".
* The missing values were imputed using the ‘mice’ package in R and a GLM model carried out on the complete datasets and the pooled coefficients were generated.
* No independent variable was statistically significant in modelling demographic groups with toxin concentration. However the coefficients show positive relationship between demographic groups and the toxin concentration, as shown by the odds ratios.
* The comparison of demographic groups based on toxin concentrations show that most adult, female dolphins from Florida contained Hg, Cu, Se and Cd toxins. Also most male, juvenile dolphins from North-Carolina had Morbillivirus.

1. **Descriptive statistics**

The *table* function was used to perform frequency analysis of the categorical variables.

Table 1

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Count** | |
| Sex | Male | 29 |
| Female | 17 |
| Age class | Adult | 31 |
| Juvenile | 16 |
| Location | NC | 30 |
| FL | 17 |
| Morbillivirus | N | 35 |
| Y | 12 |

The *summary* function was used to calculate the measures of central tendency for the toxins concentration variables.

*Man Whitney U test* was used to test the difference in toxin concentration between levels of factor variables. Mercury (Hg) variable was used as a sample for the test. P-value less than 0l05 shows that there was a difference in mercury concentration between levels of a categorical variable, otherwise the levels would be similar.

Table 3: Man Whitney test involving mercury (Hg) and 4 variables

|  |  |
| --- | --- |
| **Variable name** | **P-value** |
| Sex | 0.6954 |
| Age Class | 0.00002903 |
| Location | 0.05741 |
| Morbillivirus | 0.5725 |

1. **Imputation of missing data and Modelling: GLM model (Logistic)**

The imputation was done by the mice package in R. The “mice” stands for Multivariate Imputation by Chained Equations. We use this package in order to replace missing values with plausible values to estimate more realistic regression coefficients that are not affected by missing values. The mice package allows us to create a number of imputed datasets that replace missing values with plausible values and conduct our analysis on these separate, complete datasets in order to obtain one regression coefficient.

The first step was to check the pattern of missing values in the variables, and the output shows the proportion of missing values in each variable. The ones with more than 25% are dropped from the dataset because they will interfere with the imputation process. The variables dropped are as follows: PCB\_1268: (0.59574468), DEP: (0.59574468), NPE: (0.59574468), Triclosan: (0.59574468), BPA: (0.59574468), and Atrazine: (0.59574468). Then the mice package generates 5 imputed datasets. Then we run the logistic regression on each of the 5 datasets and pool the estimates together to get average regression coefficients and correct standard errors.

**Model Assumptions**

1. The dependent variable should be binary (all the demographic groups are binary). Yes they are binary; all are factors with 2 levels.
2. Observations to be independent of each other: This can be done by checking for multicollinearity. The Hg, Cd and Pb variables were correlated with at least one other variables hence dropping them from the model.
3. The sample size (N) should be large enough for instance the formula;

N = (10\*k/p) = (10\*10/0.26) = 385

Where k=10 independent variables, p=smallest proportion of the demographic variables: p=0.26 courtesy of the smallest proportion of (12/47 for Morbillivirus).

**Evaluation of Model fit**

The p-values of the independent variables shows their statistical significance in the model. The p-values are greater than 0.05 (a level of statistical significance) therefore the model was not a good fit. This can be attributed to the small sample size of the data because one of the assumptions of a GLM model is to have a large sample size especially when we have many independent variables.

**Interpretation of the coefficients**

In the logit model the response variable is log odds: ln(odds) = ln(p/(1-p)) = a\*x1 + b\*x2 + … + z\*xn. The odds ratios which are the exponentials of the coefficients that result from running the model. Will have a model like; odds = p/(1-p) = a\*x1 + b\*x2 + … + z\*xn. The odds ratios tables contain the odds of having one level of the response variable on relation to the other level based on the explanatory variables (toxin concentrations). For instance as the concentration of mercury (Hg) increases by one unit then a dolphin is 1.001 times more likely to be affected by the Morbillivirus. Similarly it is 9.75 times likely with Cu and 9.895 times with Se. This same interpretation can be used for the location, age class and sex demographic group variables.

**Comparison of toxin concentration based on different demographic**

In terms of location; Florida dolphins had the highest concentration of Hg, Cu, Se, Cd, Pb and As. While North-Carolina had the highest concentration of Tl, Mn, Co, and Zn.

In terms of Sex; female dolphins had the highest concentrations of Hg, Cu, Se, Cd, Tl, Co, Zn As. While had the high concentration of Pb and Mn.

In terms of Age class; adult dolphins had highest concentration of (Hg, Cu, Cd,Se, Pb, and Co) while juvenile dolphins had (Cu, Tl and Mn) in highest concentration.

In terms of Morbillivirus; dolphins with the virus had highest concentration of (Tl, Mn and Zn) while those without had highest concentration of (Hg, Cu, Se, Cd, Pb and Co).