Dirichlet Multinomial Regression –

Data:

The data used for the catch comparison analysis was compiled and obtained from an Bord Iascaigh Mhara. Importing data into R can be done in numerous ways as this programme is quite dynamic with regard to the formats which it accepts. As the dataset came in an excel format, the method used here was to uses the xlsx package which allow the loading in of .xlsx files into R as a dataframe.

In the excel file there were two sheet, one in relation to the length of nephrops in each haul given the mesh size and the side which the net was hauled and the second sheet was with regard to the weights associated with each haul. Both of these data sheets needed to be imported into excel separately, as the xlsx package doesn’t allow for more than one sheet to be loaded at a time. This is done in the function specification which has a method referred to as “sheetName” which selects the particular sheet by name and loads it into R as a dataframe.

The variable in the dataset for the nephrops length included Haul, Compartment, Gear No., Species, Carapace Length, Count, Raising Factor and Raised Counts. The dataset for nephrops weight was very similar to the nephrops length data, except that count and raised count were replaced by total wt (total weight) and sample wt (sample weight). The investigation into the variable for each data sheet was done to determine if there were any variable by which the two datasets could have been merged, given the similarities there were.

Some data cleaning was required afterward which involved the removal of some columns which were added unnecessarily with the import.

With using datasets in R it can be important to use the command str(), or structure, to determine what values are associated with each variable, e.g. factor, numeric, character, etc. Some of the values associated with variables can cause issues with some functions, for example the number of the hauls in this dataset was identified as 1 to 12. This can cause issues as Haul could be treated as a continuous numeric variable as opposed to a factor or categorical variable which is how it is meant to be used in this context. The column called Compartment, which identified the net and haul position of the net, also required that spaces between the net size and haul position (e.g. 100mm Port, changed to 100mmPort).

Some of the cleaning methods used in for this included:

1. subset() – This function allows for dataframe to broken into smaller pieces. Subsetting datasets was used to breakup dataframes or removing columns.
2. names() – This function is used is in conjunction with R syntax so that changing the names of columns. Renaming variables was necessary for avoiding errors when coding.
3. gsub() – This function allows for replacement of patterns, such as spaces, from values in columns or variables. This function was used for removing spaces from the compartment column in the nephrops length sheet of the data.
4. paste() – This functions allows for a character or characters to be pasted in front of numeric variables. This can change how R interprets a variable which is why it was applied to the Haul variable

When the data is imported into R there can be a need to rename some of the variables of importance so that mistakes can be avoided when writing out function or referencing columns in other equations. Within both datasets there is an associated raising factor with it as only some of the haul was counted or measured. A check was carried out to see if both the raising factor for length and weight are the same. From comparison between the excel dataset and the data imported into R it shows that they are in fact the same datasets.

The two cleaned sheets were then merged together, using the merge function which comes with the base R package, by the Haul and Compartment variable.

The MGLM package and Data recommendations:

The package used to run a Dirichlet Multinomial Regression was the MGLM package developed by Yiwen Zhang and Hua Zhou. The MGLM package allows for “(1) multivariate discrete distributions, (2) generating random numbers from multivariate discrete distributions and (3) running regressions and penalized regressions on the multivariate categorical response data”. There are a number of distributions which can be implemented, including the Dirichlet multinomial, generalized Dirichlet multinomial, multinomial and negative multinomial distribution. This package also comes with a number of different functions related to analysis of multivariate analysis, however not all of these applicable for this analysis.

The function “MGLMreg” requires that the dependent variable and the data associated with it is laid out in a particular fashion. To reorganise the data into a form more suitable for the “MGLMreg” function the reshape package was used. The reshape package has two functions; melt and cast, which facilitate easy data frame reorganising. The melt function takes a particular column or columns, which are identified in the melt function, and then reduces or “melts” this column into two columns, which contain the name of the variable and the value associated with it.

The cast function uses the “melted” data to reshape it into a form that is usable with the “MGLMreg” function. Cast sets the melt columns into new columns; this function uses a syntax similar to that of the formula syntax used for developing models, with the variable which remain unchanged being listed on the right hand side (RHS) separated by a tilde (~) while the columns to be reshaped are listed on the left hand side (LHS), which in this case were columns of values associated with each compartment (70mm port/starboard and 100mm port/starboard) for each variable (e.g. counts, raised counts, sample wt, total wt, raising factor etc.).

“MGLMreg” function utilises the standard R formula syntax, where the dependent variable is on the right hand side while the independent variable are on the left hand side, this function also requires that the distribution and dataset be identified as well, with the distribution being a Dirichlet and the dataset being named “AA.cast” (which is the short hand R code for the cast version of the arkh angel data).

The formula below is the regression for the simple model, using only the proportions of carapace length:

reg.simp <- MGLMreg(cbind(c100mm\_Port, c100mm\_Starboard, c70mm\_Port, c70mm\_Starboard) ~pCarapace.length, data=cast.rf, dist="DM")

The “MGLMreg” function is different to some of the other GLMM methods which are available in R. Most of the other packages that execute GLMMs can utilise some of the base functions in R to further investigate the estimate of their model. Functions such as:

* AIC() - providing the Akaike information criterion,
* vcov() - providing the variance covariance matrix of the model, cannot be used with the “MGLMreg” function.
* coeff() - providing the coefficients for each independent variable
* summary() - generally for GLMMs the summary() function gives the main estimates of the models parameters as well as the log likelihood, AIC, Bayesian information criterion (BIC) and deviance associated with the model

All of these options are available to the “MGLMreg” function. By running a summary() of the model the user is given a list of options for printing which can be selected and viewed using R syntax that is more similar to selecting columns from a dataframe, such as model$AIC or model$coefficients, etc.

Regression Models:

The model was developed with the intention to investigate the composition of the nephrops in each haul of an average fishing vessel over the course of the average day at sea. The dependent variable of the multivariate model are made up of the 2 different net sizes (70mm and 100mm mesh size) with two different haul positions (port and starboard). Three additional variable were added using data that was already available, the first of these columns was for the proportions of carapace length for each net size (“pCarapace.length”), the second was this “pCarapace.length” variable squared and finally “pCarapace.length” cubed. These columns were added to investigate the possibility that the relationship between the counts of nephrops in each net and the size of these nephrops was not a linear relationship.

As was stated earlier there are different values associated with the dependent variable, including raised count and observed counts. The first model which was used was a simple model including the observed counts as the independent variable only.

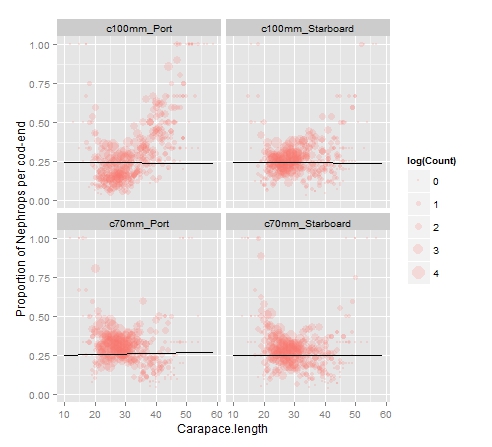


Figure - Basic Model

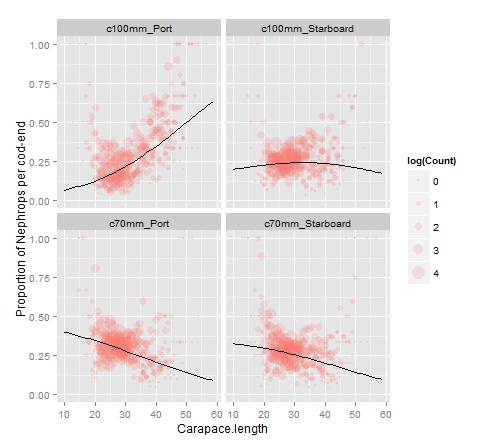


Figure - Simple Model

After plotting the fit and investigating the parameter estimates, an extension of the independent variable was made. In the second model carapace length was added as a quadratic (carapace length squared) and for the third model carapace length was added as a polynomial (including carapace length squared and carapace length cubed).

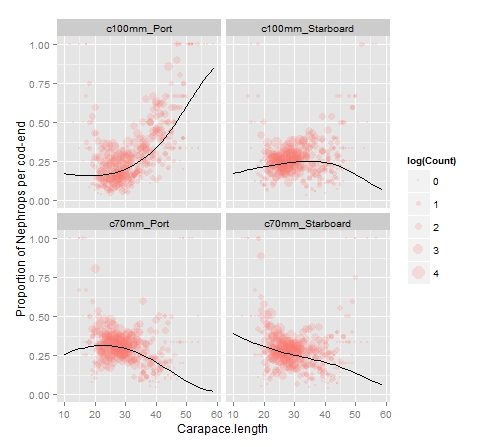


Figure - Quadratic Model

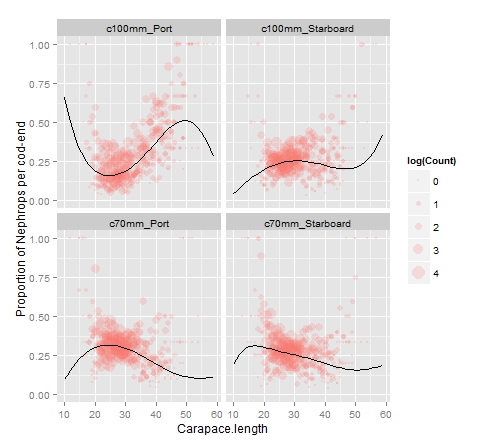


Figure - Polynomial Model

One important aspect of the observed counts which needs to be addressed is that with the observed counts a raising factor was also calculated. The raising factor that is associated with these counts can be incorporated using the offset function, which is usable in other GLMM packages; however in the MGLM package this function isn’t currently available. An alternative to this was to use the dependent variable with raised counts instead of observed counts. This can cause issues with parameter estimates as the model is using an estimation of the total number of nephrops at a particular size rather than using the observed counts and the raising factor to provide a more accurate estimation given the data.

The intention for the MGLM function will be to incorporate the offset function into the model using alternative means.