**Generalised Linear Model assignment**

Researchers are interested in which variables affect the reproductive success (number of eggs) of a fish species. They collected data in two years, 2010 and 2011, and recorded if the female who laid the eggs was dominant or not, and her body length and age. The data set is available as ‘data.fish.csv’.

Start with reading the data into R and look at the effects of age on reproductive success. Keep it simple and look at the data from 2011 only.

a. Visualise the relationship. Write down any patterns you observe which may affect your choice of test.

-There is a high concentration of zeros.

-It is bounded by zero

-no homogeneity in the variance

-we have count data (which suggest using a log-link function, )

- a non-linear relationship

b. What type of data is the response variable and what distribution does it follow?

The data is count data, and is discrete.

It is also highly right-skewed in it’s distribution.



c. One potential problem are the many zeros in the data set. What is the term for this type of data sets and how can you deal with it?

The term for this type of data set is

How to deal with it:

Exclude the zeros

Log Transform data

Apply a generalized linear regression with a Poisson Distribution.

d. Let’s take a pragmatic approach and only select the fish with eggs (number of eggs > 0).

Let’s explore two approaches to deal with our non-normal data type.

First let’s take a linear approach modelling approach.

Approach 1, LM

e. log transform the response variable and perform a linear model. Check the model fit. Do you think we can use a linear model for the transformed data set? Is age a significant predictor of reproductive success?

Model fit:

The model fits the data as the p-value for the F-test is p<0.001 indicating that the model is likely significant. Beta(age) is likely different from zero

Adjusted R^2: 45.34 % of the variance in log(egs) can be explained by age.

Age is a significant predictor of reproductive success, p-value for the t-test < 0.001 meaning that the estimator age is different from zero (when you exclude 0’s.)

Also the F-Test shows that age I significant at an alpha = 1% significance level

(special case as we have only one estimator otherwise the F-test would test for joint significance/model fit)

If we increase age by one year, holding everything constant, the number of eggs increase by 39.82%.

Log transformation is not appropriate as:

Addition of an arbitrary amount to the response because one cannot take a log of zero.

f. Make a plot of the transformed data and fit the predicted values. Take care making the graph and pay attention to detail.

The predicted values, indicated by the linear regression, suggest that reproductive age begins at approximately 6 years of age, but this is not true, given that our data > 0, only occurs at ages greater than ~12 years.

Approach 2, GLM

g. The second approach is to use a generalised linear model approach. What are the similarities and what are the differences with the earlier linear model approach?

**We will do this after we’ve read the reading.**

Why would you prefer this approach over a transformation and linear model?

**Same ^**

h. Perform a generalised linear model looking at the effects of age on the number of eggs and interpret the output. What scale are the parameters estimates in? Report the slope estimate in the correct units.

Estimates from summary(glm1): slope = 0.336106, and intercept = -1.342074. These estimates are in a log scale, given that the Poisson family applies a log link function.

Considering this, and the inverse of the log link function, our parameters can be estimated in the original scale given the following equation:

Thus slope estimate is equal to 1.3995

And intercept estimate is equal to 0.2613

**Interpretation**

i. Plot the data and the fitted curve. Try plotting the fitted model in two ways:

• use the ‘$fitted.values’ to extract them for the values in the model output\*

Line isn’t displaying on graph.

• use the predict() function where you need to input the model output, a list of values for which you want to have predicted values and set type = “response”). The latter method can give a much smoother curve if the fitted values in the response variable are very far apart.

Overdispersion occurs if the variance increases more than predicted.

j. Is overdispersion a problem in this data set? Estimate by dividing the $deviance by $df/residual.

k. Such overdispersion can be taken into account by running a so called quasipoisson link function. Re run with , family = "quasipoisson".

Compare both “poisson” and "quasipoisson" outputs and add the predicted curve to the plot.

Are they different? (HINT if you used predict() you can subtract the values from the poisson model from the quasipoisson to see how different they are).

We stick with the ‘poisson’ error link for now.

l. Let’s look at another variable and see if it can explain the relationship better. Add body.length

Run different subsets of models including both variables, either and non. Compare using anova() and set the test to Chisq (), anova(m1, m2, tets = “Chisq). Which variables should you keep and how did you determine this?

\*remember that lines()follows the order of x and y coordinates you give it.