Follow-up course regression, GLM & GAM

Highland Statistics Ltd www.highstat.comn

Exercise: Red squirrels and the offset variable in a GLM

Data description

Flaherty et al. (2012) looked at the forest stand structure on red squirrel habitat use.

Fieldwork was carried out in two areas in Scotland: Abernethy Forest and Aberfoyle. Within each sampling area, multiple 'plots' were selected. These plots are of dimension 14 m by 14 m. There are 20 plots in Aberfoyle and 32 in Abernethy.

Within each plot the following variables were measured:

- Diameter at breast height (DBH)
- Canopy closure
- Tree height

A bit more detail on this:

- …"Canopy closure measurements were made at the central point and at each of the four corners of the plot and subsequently averaged. DBH was recorded for all trees per plot. For tree height, all trees were measured. In plots where the number of trees was greater than 30, the height of the four trees with the largest DBHs was measured and recorded."…

In the spreadsheet these three variables are labelled as: DBHav, CanopyCover and T_heightav. We also have the number of trees per plot.

Squirrel presence and local abundance were assessed using cone transect lines. The response variable is the number of cones within a plot that were stripped by squirrels (feeding remains).

Flaherty et al. (2012) used a GLM and modelled the dependent variable (number of cones stripped by squirrels) as a combination of three independent variables collected in the field: canopy closure, tree density and tree height. They also included the number of trees per plots as a covariate.

The data are in the file RedSquirrel.txt and Table 1 gives the variables. See Flaherty et al. (2012) for a biological justification for using these variables. Each row in the text file represents a plot

Table 1. Variables in the file RedSquirrels.txt.

Variable	Description	Type	
Squcones	Number of cones stripped by squirrels	Count	Response variable
Ntrees	Number of trees in a plot	Continuous	Covariate
DBHav	Average DBH in a plot	Continuous	Covariate
T_heightav	Average tree height in a plot	Continuous	Covariate
CanopyCover	Canopy cover in a plot	Continuous	Covariate

References:

- FLAHERTY, PATENAUDE , CLOSE AND LURZ (2012) The impact of forest stand structure in vitro adventitious shoots formation on red squirrel habitat use. Forestry, doi: 10.1093/forestry/cps042

Underlying question and task

Just like the Flaherty et al. (2012) paper we would like to model the number of stripped cones as a function of the 4 covariates.

First follow the Flaherty et al. (2012) approach and apply a Poisson GLM using Squcones as response variable. Use all 4 covariates (provided there is no collinearity; Flaherty et al. (2012) dropped DBHav because of collinearity). Obviously you also need to do a data exploration. No interactions were used.

We would like to use these data to dive a little bit deeper in the offset. Below is some text that was copied and pasted from "A Beginner's Guide to Generalized Additive Models" by AF Zuur (publication date is the autumn of 2012). The text has not been edited on grammar and style yet.

4.4 Using the offset in a GLM or GAM

In the previous subsection we ignored the fact that the sampling effort (swept area) differs per haul, and we need to take this into account. Our failure to do so in the previous subsection makes those analyses faulty. So how do we include sampling effort into the model?

The Poisson distribution can used to describe the probability for the number of events in a time interval (t, $t + \lambda$] where t refers to time and the λ determines the size of the time interval. The density function for such a Poisson distribution is given by:

$$P(Y = y \mid \mu, \lambda) = \frac{e^{-\lambda \times \mu} \times (\lambda \times \mu)^y}{y!}$$

This density function can be used to calculate the probability that we count 0, 1, 2, or any other number of observations in the time interval (t, $t + \lambda$], assuming that we know the rate parameter μ and length of the time interval λ. The expected value of Y is $\lambda \times \mu$. If $\lambda = 1$ then μ is the expected number of events in one unit time. But if the time interval is twice as long, you expect twice as much counts, hence the $\lambda \times \mu$.

For the fishery data we do not have time units but we have sizes of sampling areas, denoted by the variable swept area. Out total abundance TA_i is essentially the number of events in a sampling area

of size SA_i (swept area). The Poisson density function can be used to quantify the probabilities of all possible values of total abundance for given swept area (λ_i) and expected numbers in a swept area of unit size (μ_i):

$$P(TA_i = y_i \mid \mu_i, \lambda_i) = \frac{e^{-\lambda_i \times \mu_i} \times (\lambda_i \times \mu_i)^{y_i}}{y_i!}$$

As before the expected total abundance is $\lambda_i \times \mu_i$. The rate parameter μ_i represents the expected total abundance for a sampling area of unit size.

If we look at the total abundance *per* swept area, that is TA_i / SA_i, we can write the expression for the mean of TA_i / SA_i as:

$$E\left(\frac{TA_i}{SA_i}\right) = \frac{\mu_i}{SA_i}$$

It is perhaps slightly confusing that we change from a product to a ratio but that is because we first looked at the number of counts in the interval (t, $t + \lambda$] and now we look at the total abundance *per unit* swept area (which is a ratio). Using the log-link function for the Poisson or negative binomial distribution we get:

$$\log\left(\frac{\mu_i}{SA_i}\right) = \alpha + \beta_1 \times Depth_i + \beta_2 \times Period_i + \beta_3 \times Depth_i \times Period_i \quad \Leftrightarrow$$
$$\log(\mu_i) = \alpha + \beta_1 \times Depth_i + \beta_2 \times Period_i + \beta_3 \times Depth_i \times Period_i + \log(SA_i)$$

Note that there is no unknown regression parameter in front of the log(SA_i) term. This term is also called the offset variable.

Summarising, to model the total abundance while taking into account the differences in sampling effort per haul we can use an offset variable. The advantages are that we obtain positive fitted values and allow for heterogeneity in total abundance.

In Section 4.4 a different data set was used, but for the squirrel data set we can argue that we have a similar issue, namely the number of trees. The more trees there are, the more cones, hence the more food. We can do three things:

1. Use the number of stripped cones divided by the number of trees as response variable. This is some sort of density. Obviously you would not use the number of trees as covariate. A possible analysis is a multiple linear regression using the remaining 3 covariates.

2. Use the number of stripped cones as response variable and use the log of the number of trees as an offset. This assumes that if you have twice the number of trees, you would have twice the number of stripped cones (and so does the density approach).

3. Follow the Flaherty et al. (2012) paper and use the number of trees as a covariate.

In this example we are interested in the second and third approaches; hence apply both. And try to understand the difference between them.

As to the actual analysis, at some point try to investigate whether the authors should have considered forest area as a categorical variable.