Data exploration, regression, GLM and GAM course

Highland Statistics Ltd   www.highstat.comn

Exercise 10: GLM applied on the Bailey data

Data description

See exercise 2. In the data exploration we decided to remove 2 sites. You need to stick to this decision.

Underlying question and task

The aim of this exercise is to model the total abundance (count) as a function of mean depth and period. The problem is that the sampling effort (SweptArea) differs per site. One option is to model the density (TotAbund / SweptArea), but in earlier analysis we have seen that this gives negative fitted values and heterogeneity. Instead we will analyse the TotAbund with a Poisson (or negative binomial) GLM using the natural log of SweptArea as an offset. This works as follows. The starting point is the following equation:

$$\frac{TotAbund_i}{SweptArea_i} = e^{\alpha + \beta_1 \times Depth_i + \beta_2 \times Period_i + \beta_3 \times Depth_i \times Period_i}$$

$$TotAbund_i = SweptArea_i \times e^{\alpha + \beta_1 \times Depth_i + \beta_2 \times Period_i + \beta_3 \times Depth_i \times Period_i}$$

$$TotAbund_i = e^{\ln(SweptArea_i)} \times e^{\alpha + \beta_1 \times Depth_i + \beta_2 \times Period_i + \beta_3 \times Depth_i \times Period_i}$$

$$TotAbund_i = e^{\alpha + \beta_1 \times Depth_i + \beta_2 \times Period_i + \beta_3 \times Depth_i \times Period_i + \ln(SweptArea_i)}$$

This is a log link function, but note that there is no parameter in front of the ln(SweptArea$_i$) term! Instead of modelling the density we will model the TotAbund (which is a count) with a Poisson (or NB) GLM and tell the glm function that there should not be a parameter in front of the ln(SweptArea) term.

The advantages are:

1. Fitted values are always positive.
2. We allow for heterogeneity

To fit this in R, use:

```
> Fish3$LSA <- log(Fish3$SweptArea)
> M1 <- glm(TotAbund ~ MeanDepth * factor(Period) + offset(LSA),
            data = Fish3, family=poisson)
```

The underlying model is:

$$TotAbund_i \sim Poisson(\mu_i)$$
$$\log(\mu_i) = \alpha + \beta_1 \times Depth_i + \beta_2 \times Period_i + \beta_3 \times Depth_i \times Period_i + \ln(SweptArea_i)$$

Now apply the usual steps:

- Is there overdispersion?
- Do you need Poisson, quasi-Poisson or NB GLM?
- Is everything significant?
- What is the optimal model?
- Apply a model validation.
- Sketch the fit of the optimal model.

Half an hour later…….here is the optimal model

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)         5.913e+00  1.365e-01   43.31   <2e-16 ***
MeanDepth          -7.129e-04  4.806e-05  -14.83   <2e-16 ***
as.factor(Period)2 -1.336e+00  1.258e-01  -10.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.9307) family taken to
be 1)
```

$$
\begin{aligned}
&Abundance_i \sim NB(\mu_i, 1.93)\\
&E(Abundance_i) = \mu_i\\
&\text{var}(Abundance_i) = \mu_i + \frac{\mu_i^2}{1.93}
\end{aligned}
$$

$$\log(\mu_i) = 5.91\text{-}0.00071 \times MeanDepth_i + \log(SweptArea_i) \qquad \text{if Period} = 1$$

$$\log(\mu_i) = 5.91\text{-}1.33\text{-}0.00071 \times MeanDepth_i + \log(SweptArea_i) \qquad \text{if Period} = 2$$

$\text{Re}\,write:$

$$\mu_i = e^{5.91\text{-}0.00071 \times MeanDepth_i + \log(SweptArea_i)} \qquad \text{if Period} = 1$$

$$\mu_i = e^{5.91\text{-}1.33\text{-}0.00071 \times MeanDepth_i + \log(SweptArea_i)} \qquad \text{if Period} = 2$$

$\text{OR Rewrit}e:$

$$\frac{\mu_i}{SweptArea_i} = e^{5.91\text{-}0.00071 \times MeanDepth_i} \qquad \text{if Period} = 1$$

$$\frac{\mu_i}{SweptArea_i} = e^{5.91\text{-}1.33\text{-}0.00071 \times MeanDepth_i} \qquad \text{if Period} = 2$$

To get predicted values for a graph, we need to choose a value for SweptArea.