Day 2 Exercises without Solutions

Generalized Additive Models with Binary Data

1) GAMs are particularly valuable with binary response variables. You can explore and demonstrate this GAM facility using the **`isolation.txt`** data set (enclosed in data folder with zipped class material for today) that we used in the class slides today to implement the binary response variable example (the Incidence Function) for GLMs. Model and estimate a GAM to further understand how the isolation of an island and its area influence the probability (i.e. the '0' or '1' incidence) that the island is occupied by the species in question.

   In the logistic regression in the class slides, **`area`** (island size) had a significant positive effect of the likelihood of occupancy, and **`isolation`** had a highly significant *negative* effect on the probability that the island will be occupied by the species. But there was no *a priori* reason to believe that the logit of the probability should be linearly related to either of the explanatory variables. Use a GAM to fit smoothed functions of both area and isolation to the incidence data. Use the same error family and link functions as with the logistic regression example using the same data.

   In the GAM results, does **`isolation`** have a significant effect on occupancy? Does **`area`** have a significant effect? Use the function **`plot.gam()`** to plot the GAM model output to look at the residuals. Adjust the plotted graphics frame with this preceding R statement to allow two plots in the same graphics frame:

   ```
   > par(mfrow=c(1,2))
   ```

   Look at the residual plots. What, if anything, do they tell you about the effect of area on incidence, particularly above a threshold of about area = 5?

   Create an **Analysis of Deviance** table and compare the model containing both of the terms **`s(area)`** and **`s(isolation)`** with a second model containing only the **`s(isolation)`** term as the explanatory variable. What do the results of the analysis of deviance indicate about the relative effect of **`area`**? How does this compare with the *p* value for **`area`** in the summary table of the first model you created and estimated in this exercise (above)?

   Fit a third model with area estimated as a parametric (linear) term and isolation as a smoothed term. What do the results of this third model indicate?

   Considering all of the results above, what do you ultimately conclude about the effect of **`area`** on **`incidence`**? About the relative value of: i) judging parameter terms in the summary() function output; ii) 'deletion analysis' using the anova table; and iii) examining residual plots?