Ecologia Numérica

Componente Teórica - Prática

Ficha de trabalho 7

Esta é uma ficha de trabalho mais longa que as anteriores, para ser realizada de forma independente (ou seja, sem apoio directo nas aulas TPs). Pode ser realizada de forma individual ou, caso assim o desejem, a pares. Pretende integrar os conhecimentos adquiridos ao longo da primeira metade da cadeira de Ecologia Numérica, com enfase nos modelos de regressão linear (*Linear Model*, LM) e modelos lineares generalizados (*Generalized Linear Models*, GLMs), e chegando a explorar um modelo aditivo generalizado (*Generalized Additive Models*, GAM).

A regressão é uma metodologia chave na análise de dados ecológicos. O objectivo é explicar uma variável dependente (que depende de, ou variável resposta) à custa de variáveis independentes (também denominadas de preditoras, explicativas, ou covariáveis). Se a resposta puder ser considerada Gaussiana temos um LM (implementado pela função `lm`), se não for Gaussiana mas a relação for linear (mesmo que apenas na escala da função de ligação) podemos usar um GLM (implementado pela função `glm`), e no caso de a resposta não ser linear podemos usar um GAM (implementado pela função `gam`, do package `mgcv`).

No caso de `glm` e `gam` temos de definir o argumento `family`, que define a família de distribuições que se pretende para a variável resposta. As famílias mais comuns, alem da Gaussiana, são a Binomial, Poisson, Binomial Negativa e Gama. `?family` dá-nos acesso à sintaxe para usar cada uma destas famílias.

A sintaxe dos modelos no R é a seguinte

- Y~X quer dizer Y em função de X
- Y~X+Z quer dizer Y em função de X e de Z
- Y~X*Z quer dizer Y em função de X e de Z, incluindo a interacção entre X e Z.
- Y~s(X) que dizer Y como uma função não linear de X (apenas válido no `gam`)

Alguns recursos uteis (alem de, claro, os slides das aulas teóricas):

- https://www.statmethods.net/advstats/glm.html
- https://www.r-bloggers.com/generalised-linear-models-in-r/
- https://www.r-bloggers.com/an-intro-to-models-and-generalized-linear-models-in-r/
- https://www.theanalysisfactor.com/r-tutorial-glm1/

Entrega: envio por e-mail de um relatório dinâmico em R Markdown, enviando o pdf e o respectivo .Rmd

Data limite para entrega: 19 de Novembro

Nome dos ficheiros: FT7A*****.Rmd ou FT7A1*****A2*****.Rmd e FT7A***** Ou FT7A1*****A2*****.Rmd, onde ***** corresponde ao(s) número(s) de aluno correspondente(s) ao(s) autor(es).

1. Efectue uma análise de regressão (Função `lm`) sobre os dados *DataTP7densidade.csv*, relativos à densidade de uma espécie em função de variáveis ambientais, considerando a densidade como a variável dependente a temperatura como independente. Faça uma representação gráfica adequada dos dados e do modelo. Explore os resultados e efectue uma avaliação do cumprimento dos pressupostos do modelo linear. Qual o valor da correlação e do coeficiente de determinação entre densidade e temperatura?

2. Com base nos dados da alínea anterior, efectue uma análise de regressão múltipla usando a função `lm`, sendo a variável dependente a densidade e as independentes todas as restantes. Comente os resultados (quais as variáveis importantes para explicar a densidade?) e verifique se os pressupostos são cumpridos. (Há funções muito práticas para extrair sub-componentes de modelos. Sobre o modelo que usar, experimente usar as funções `coef`, `fitted`, `predict.lm` e `AIC`).

3. Estimou a mortalidade de aves em alguns parques eólicos, em função de variáveis ambientais (*DataTP7parqueseolicos.csv*). Verifique se a variável "altitude" é um bom preditor da mortalidade, através de um modelo de regressão simples (Função `lm`). Caracterize a relação entre as variáveis. Elabore uma função que gere estimativas de mortalidade em função da altitude (Relembrar como criar funções, tutorial 1). Avalie o desempenho do modelo de regressão (Por exemplo, faça um `plot` do objecto de regressão criado e avalie os pressupostos do modelo de regressão).

4 Realize uma regressão múltipla aos dados *DataTP7parqueseolicos.csv* e interprete os resultados. Pretende obter um sub-modelo para utilizar noutras regiões com parques eólicos. Obtenha um sub-modelo para esse efeito. (Dica: começar por criar um modelo com todas as variáveis disponíveis e depois retirar as que não parecem importantes).

5. Aplique um GLM aos dados *DataTP7presenca*, os quais descrevem a abundância de uma espécie em função de algumas variáveis ambientais. Utilize a distribuição Gama como família de distribuição do erro (Gamma) e a função inversa como função de ligação (inverse, in a `glm` call, use argumente `family=Gamma(link="inverse")`). Compare os resultados com um GLM Poisson (in `glm` call, `family=poisson`). Compare as predições dos dois modelos num gráfico. Qual seria o valor previsto num caso em que estradas tomasse o valor 7. Qual dos dois modelos escolheria para modelar os dados (ver a função `AIC`)?

6. Aplique um GLM aos dados *DataTP7anchova.csv*, referentes à abundância de anchova (variável dependente, ANC) em função de várias variáveis ambientais (variáveis independentes). Explore os resultados.

7. Explore os dados disponíveis em https://data.mendeley.com/datasets/r3xpn3mccc/2, usados num trabalho para prever o tamanho de grupos de baleias de bico em função da sua pegada acústica. Há uma boa descrição dos dados em "data4Mendeley.pdf", mas por agora ignore as variáveis "gID", "conf" e "cs0" como variáveis explicativas (na realidade, cs0=cs-1, o que pode ser interessante porque não há grupos de tamanho 0, mas uma Poisson assume que existem 0's). Use o ficheiro "modeldata.txt" para tentar explicar o tamanho do grupo de animais ("cs") em função das outras variáveis disponíveis.

8. Usando os dados *data4GAM.txt*, ajuste um modelo linear, um GLM e um GAM e escolha o melhor modelo para representar a relação que explica "y" em função de "x".