8. Análisis de datos en ecología (IV): selección de modelos

Ecología Metodológica y Cuantitativa (5C1)
Departamento de Ecología e Hidrología

Curso 2008–09

Índice

1. Introducción

En esta sesión trataremos los aspectos relacionados con la utilización de técnicas de regresión para realizar modelos algo más complejos que los desarrollados en la práctica anterior. En este caso analizaremos respuestas ecológicas (variables dependientes) considerando más de una variable ambiental (variables independientes). Además, contemplaremos la utilización conjunta de variables ambientales cualitativas y cuantitativas en los modelos. El procedimiento general requiere seleccionar el "mejor modelo" entre los diversos modelos considerados. Aunque en los ejercicios utilizaremos sólo técnicas de regresión discreta o de Poisson, el procedimiento es igualmente aplicable con los otros tipos (lineal y logística).

2. Colinealidad

El concepto de colinealidad se refiere a la existencia de correlación entre las variables independientes, lo que puede llevar a problemas en la estimación de los parámetros de regresión. Analizaremos en primer lugar un ejemplo con datos procedentes del muestreo de aves que ya conocemos de la práctica anterior. En este caso utilizaremos otro archivo (`aves.dat`) que contiene la información correspondiente a cuatro especies y tres variables ambientales.

Ejercicios. Bloque 1:

1. Leer el archivo y examinar las características de la matriz

   ```
   read.table("http://www.um.es/docencia/emc/datos/aves.dat")->aves
   attach(aves)
   ```

2. En este primer ejercicio utilizaremos la especie *Loxia curvirostra* (Piquituerto, variable `loxcur`) para analizar su relaciones con la altitud y la temperatura. Emplearemos la función `glm()` con la opción `family=poisson`. Observa que la relación es significativa con ambas variables.

   ```
   summary(glm(loxcur~alt, family=poisson))
   summary(glm(loxcur~temp, family=poisson))
   ```

   Para visualizar gráficamente la relación podemos utilizar las funciones `plot()` y `points()`. No obstante, a diferencia de la práctica anterior, en esta utilizaremos un método más directo, con la sintaxis `glm()$fit`:

```
plot(alt, loxcur)
points(alt, glm(loxcur~alt, family=poisson)$fit, col=2)
plot(temp, loxcur)
points(temp, glm(loxcur~temp, family=poisson)$fit, col=2)
```

3. Introduce ahora en un mismo modelo las dos variables, utilizando para ello el signo + entre ambas:

   ```
   summary(glm(loxcur~alt+temp, family=poisson))
   ```

   La función proporciona ahora, en un único modelo, los coeficientes para `alt` y `temp`. ¿Cuáles son las diferencias en la interpretación ecológica de la respuesta?

   Comprueba, finalmente, que la correlación entre altitud y temperatura es muy elevada:

   ```
   cor.test(alt, temp)
   ```

---

3. Variables independientes cualitativas en los modelos de regresión

---

Ejercicios. Bloque 2:

Analizaremos en este apartado los datos del Escribano Montesino (*Emberiza cia*, variable `embcia`), incluidos también en la matriz `aves`. Vamos a considerar la relación de esta especie con la variable cualitativa "tipo de hábitat" (`hab`; valor 1 = pinar; valor 0 = matorral).

1. Comprobaremos primero que un modelo de regresión lineal con una variable independiente cualitativa es equivalente a un ANOVA:

   ```
   summary(lm(embcia~hab))
   summary(aov(embcia~hab))
   ```

   En el caso de utilizar regresiones discretas la equivalencia no se mantiene (ya que el modelo de regresión discreta asume una distribución no normal de los errores), pero la interpretación es similar: comparamos la abundancia de Escribanos entre los dos tipos de hábitat y concluimos que la abundancia es significativamente mayor en el pinar:

   ```
   summary(glm(embcia~hab, family=poisson))
   ```

2. Estimaremos ahora los coeficientes de un modelo más complejo, analizando conjuntamente la respuesta a las variables tipo de hábitat y altitud.

   ```
   summary(glm(embcia~hab+alt, family=poisson))
   plot(alt, embcia)
   points(alt, glm(embcia~hab+alt, family=poisson)$fit, col=2)
   ```

3. Estudiaremos también el significado del concepto de interacción entre variables utilizando la sintaxis: `hab*alt`.

   ```
   summary(glm(embcia~hab*alt, family=poisson))
   points(alt, glm(embcia~hab*alt, family=poisson)$fit, col=3)
   ```

   Dado que el coeficiente de la interacción no es significativo, la diferencia entre ambos modelos (con y sin interacción) es mínima.

4. Repetir el análisis con la Curruca Cabecinegra (variable `sylmel`). Interpretar aquí el efecto significativo de la interacción en el modelo.

---

4. Selección de modelos

Cuando disponemos de varias variables independientes, el número de posibles modelos de regresión se incrementa notablemente. En estos casos es necesario desarrollar un procedimiento de selección de los "mejores" modelos: aquéllos que proporcionan la mayor explicación sobre las variabilidad de las variables dependientes. Empezaremos con el caso del Mito (*Aegithalos caudatus*, variable `aegcau`, también incluida en la matriz aves). Para comparar los modelos utilizaremos el criterio de información de Akaike (AIC), seleccionando aquéllos para los que se obtengan los valores más bajos de este parámetro.

Ejercicios. Bloque 3:

1. Realizaremos análisis independientes para cada una de las tres variables ambientales (`hab`, `alt`, `temp`), utilizando regresiones discretas. Probaremos también los modelos cuadráticos en el caso de la altitud y la temperatura. Cabe recordar que en estos casos hay que crear una nueva variable (por ejemplo, `alt2<-alt*alt`, o bien utilizar la sintaxis `I(alt^2)`.
2. Compararemos los cinco modelos utilizando los valores de AIC. Nos fijaremos también en los niveles de significación de los coeficientes.
3. A continuación realizaremos nuevos modelos, dejando "fija" la variable (con término cuadrático o no), que en el paso anterior proporcionó el menor valor de AIC, y junto a ella iremos probando, una a una, con el resto de variables.
4. De nuevo compararemos los modelos mediante sus valores de AIC y seleccionaremos el de valor más bajo. Repetiremos el procedimiento hasta que no obtengamos una mejora (disminución) del valor de AIC.
5. Finalmente intentaremos perfeccionar el modelo introduciendo los términos de interacción entre las variables ambientales seleccionadas.
6. Representar gráficamente el modelo final junto a los datos originales.

---

5. Ejercicios adicionales

1. Siguiendo el procedimiento utilizado con el Mito, se propone realizar el mismo ejercicio de selección de modelos con las otras especies de aves que aparecen en el archivo `aves.dat`: el Escribano Montesino (*Emberiza cia*, variable `embcia`), el Piquituerto (*Loxia curvirostra* variable `loxcur`) y la Curruca Cabecinegra (*Sylvia melanocephala* variable `sylmel`).
2. Utilizando el archivo `territorios.dat`, analizar la influencia de la altitud y la orientación en la ocupación de territorios y la productividad de la Águilas Calzadas. Utilizar regresiones logísticas para la ocupación (unos y ceros) y regresiones discretas para la productividad (número de pollos). Seguir en cada caso el mismo procedimiento que el utilizado con el Mito.