7. Análisis de datos en ecología (III): regresión lineal, logística y discreta

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

Curso 2008–09

Índice

Antes de empezar

Iniciar R. Cargar el archivo de funciones de la asignatura, `funciones.R`:

```
source("http://www.um.es/docencia/emc/datos/funciones.R")
```

1. Introducción

En esta práctica se tratará el problema de analizar respuestas biológicas frente a variables ambientales de naturaleza cuantitativa. Los diferentes tipos de regresión considerados aquí (lineal, logística y discreta) representan métodos adecuados para el análisis de diferentes tipos de variables dependientes. Así, la regresión lineal se utiliza con variables cuantitativas continuas, la regresión logística con variables cualitativas, y la regresión discreta (o de Poisson) con variables cuantitativas discretas. Aunque generalmente las variables independientes utilizadas en las técnicas de regresión son de naturaleza cuantitativa, también pueden utilizarse variables independientes de tipo cualitativo. No obstante, dejaremos el estudio de estos casos para la siguiente práctica, y en todos los ejercicios de esta sesión consideraremos únicamente variables ambientales cuantitativas.

2. Regresión lineal

Ejercicios. Bloque 1:

1. En primer lugar analizaremos el comportamiento de los distintos estadísticos ligados a un análisis de regresión lineal. Consideremos primero un caso de variables independientes utilizando la función `reg.norm.indp()` con 15 datos:

   ```
   reg.norm.indp(15)
   ```

   ¿Se obtiene una regresión significativa? ¿Cual es el estadístico determinante?

2. En el caso de que dos variables presenten relación lineal pueden darse distintos casos. Analizaremos el efecto de la "proximidad" de los datos al modelo teórico. Para ello utilizaremos la función `reg.norm.dp()`,

que toma el número de datos a simular y un valor de variabilidad entre los valores esperados y los observados (el valor cero corresponde a una correlación perfecta).

```
reg.norm.dp(15,0.1)
```

Utilizar la función para 0.5 y para 1.

¿Qué efecto tiene el alejamiento de los datos del modelo teórico? ¿qué estadístico muestra este efecto de forma más acusada?

Ejercicios. Bloque 2:

Se sospecha que la abundancia de *Lavandula*, en los arenales del Parque Nacional de Doñana, está condicionada por la disponibilidad de agua. Se han obtenido datos de cobertura de esta especie en lugares con diferente profundidad de la capa freática. Para estudiar las relaciones entre la cobertura y la profundidad utilizaremos un modelo lineal.

1. En primer lugar leeremos el archivo de datos:

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

2. Para realizar la regresión utilizaremos los datos transformados logarítmicamente:

   ```
   lncob<- log(cob)
   lav.reg<- lm(lncob~prof)
   summary(lav.reg)
   ```

3. Determinar la significación estadística de los coeficientes y el valor de R^2 ajustada.

4. Escribir la ecuación de la recta de regresión obtenida y crear una nueva variable con los valores ajustados ($y = b_0 + b_1\ x$):

   `b1 <-` ... valor del coeficiente de la variable procedente de la tabla de coeficientes

   `b0 <-` ... valor del coeficiente independiente procedente de la tabla de coeficientes

   ```
   lncob.esp <- b0 + b1 * prof
   ```

5. Representar en una gráfica los valores ajustados (la recta de regresión) y los datos transformados frente a la profundidad.

   ```
   plot(prof, lncob)
   points(prof, lncob.esp, col=2)
   abline(b0, b1, col=3)
   ```

6. Probar, ahora, con un modelo cuadrático:

   ```
   prof2 <- prof*prof
   lav.reg2 <- lm(lncob~prof+prof2)
   summary(lav.reg2)
   ```
   [1]

7. Determinar la significación estadística de los coeficientes y el valor de R^2 ajustada, comparándolos con los anteriores.

8. Escribir la nueva ecuación del modelo y crear otra variable con los valores ajustados. Representa en una gráfica los valores ajustados y los datos originales frente a la profundidad ($y = b_0 + b_1\ x + b_2\ x^2$):

   `b2 <-` ... valor del coeficiente de la variable cuadrática

   `b1 <-` ... valor del coeficiente de la variable

   `b0 <-` ... valor del coeficiente independiente

   ```
   lncob.esp <- b0 + prof * b1 + prof2 * b2
   plot(prof, lncob)
   points(prof, lncob.esp, col=2)
   lines(prof, lncob.esp, col=3)
   ```

[1] También puede utlizarse la expresión `lm(lncob~prof+I(prof^2))`

9. Representar los valores y el modelo con la escala original ($y' = e^{b_0+b_1x+b_2x^2}$):

```
cob.esp<-exp(lncob.esp)
plot(prof, cob)
points(prof, cob.esp, col=2)
lines(prof,cob.esp, col=3)
```

---

3. Regresión logística

Cuando tenemos que utilizar modelos de regresión con variables dependientes de tipo cualitativo emplearemos técnicas de regresión logística. En R, la función para aplicar regresiones logísticas es `glm()` (de *Generalized Linear Models*), añadiendo al final la opción `family=binomial`. Los datos del ejemplo siguiente forman parte de un experimento en el que se estudió la respuesta germinativa de una rara especie halófila, el *Halocnemum strobilaceum*, en función de la presión osmótica.

Ejercicios. Bloque 3:

1. Leer el archivo `halocnemum.dat`

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

   Examinar las características de la matriz: en la primera columna figura el número de semillas que germinan en cada unidad de muestreo, en la segunda el número de semillas que no germinan (en este experimento el total de semillas es siempre 25), y en la tercera los valores de presión osmótica.

2. Aplicaremos la función para realizar la regresión logística. Necesitaremos una nueva matriz, que actuará como «variable» dependiente, en la que las columnas se correspondan con los éxitos y los fracasos (la especificaremos con `cbind(germ, nogerm)`), y el vector con los valores de la variable independiente `pres`.

   ```
   glm(cbind(germ, nogerm)~pres, family=binomial) ->halocnemum.glm
   summary(halocnemum.glm)
   ```

3. Para analizar los resultados escribiremos la ecuación del modelo obtenido y crearemos una nueva variable con los valores ajustados mediante la siguiente transformación:

   $$y = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$$

   Para simplificar las expresiones utilizaremos una variable intermedia y_1 siendo: $y_1 = b_0 + b_1x$, y por lo tanto

   $$y = \frac{e^{y_1}}{1+e^{y_1}}$$

   `b1 <-` valor del coeficiente de la variable

   `b0 <-` valor de la constante

   ```
   y1 <- b0+b1*pres
   ```

   Así los valores esperados según el modelo son:

   ```
   y <- exp(y1)/(1+exp(y1))
   ```

4. Para interpretar adecuadamente los resultados representaremos gráficamente el modelo.

   ```
   plot(pres, germ/25)
   lines(pres, y, col=2)
   ```

---

4. Regresión discreta o de Poisson

Analizaremos en este apartado los datos de un muestreo de una especie de ave (la Collalba Negra, *Oenanthe leucura*), en el que se anotó el número de individuos encontrados en 180 taxiados realizados en diversas sierras de la Región de Murcia. En este ejercicio consideraremos como variable independiente la altitud sobre el nivel del mar (variable `alt`). Dado que la variable dependiente es de naturaleza discreta (número de individuos), emplearemos un modelo de regresión discreta con la función `glm()`, pero eligiendo en este caso la opción `family=poisson`.

Ejercicios. Bloque 4:

1. Leer el archivo *pajaros.dat*.

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

2. Aplicaremos la función para realizar la regresión discreta:

   ```
   glm(oenleu~alt, family=poisson)->oenleu.glm
   summary(oenleu.glm)
   ```

3. Para interpretar adecuadamente los resultados representaremos gráficamente el modelo. En el caso de las regresiones discretas, el cálculo de los valores ajustados se realiza mediante:

   `b1` <- valor del coeficiente de la variable
   `b0` <- valor de la constante
   `y <- exp(b0+b1*alt)`

4. Con las siguientes funciones representaremos la gráfica. Como los datos de altitud no están ordenados, no es conveniente aplicar la función `lines()`; en su lugar utilizaremos `points()`.

   ```
   plot(alt, oenleu)
   points(alt, y, col=2)
   ```

---

5. Ejercicios adicionales

1. En el archivo `marmenor.dat` se encuentran los censos (1983-2002) de tres especies de aves acuáticas invernantes en el Mar Menor (`Pcri`: *Podiceps cristatus* - Somormujo Lavanco; `Pnig`: *Podiceps nigricollis* - Zampullín Cuellinegro; `Phca`: *Phalacrocorax carbo* - Cormorán Grande), así como los aportes estimados de nitrógenoa la laguna en esos años (TmN).Analiza para cada especie las relaciones entre el número de individuos y los aportes de nitrógeno.

2. En el archivo `cistus.dat` se presentan datos correspondientes a un estudio sobre las relaciones entre la vegetación y la profundidad de la capa freática. En este caso los valores corresponden a medidas de cobertura de la especie *Cistus libanotis*. ¿Qué modelo de regresión proporciona una mejor explicación de la respuesta de la especie?

3. En el archivo `pinos2.dat` se presentan los datos de dbh de pinos analizados en la práctica anterior. Ahora se proporcionan también los valores de altitud (sobre el nivel del mar) y pendiente de cada parcela. ¿Existe relación entre el `dbh` y las variables ambientales?

4. El Águila Calzada (*Hieraætus pennatus*) es una rapaz migradora que retorna a la Península Ibérica durante el mes de marzo, ocupando territorios en áreas forestales donde dispone de nidos para reproducirse. Queremos analizar el efecto de las perturbaciones humanas en la ocupación de territorios por esta especie, y para ello se dispone de información sobre un total de 61 territorios: ocupación (variable `ocup`); distancia a la pista forestal más próxima (`dpf`); distancia a cultivos (`dcu`); distancia al borde del bosque (`dbb`); distancia al camino o senda más cercana (`dcs`); distancia a la cantera más próxima (`dca`). Los datos se proporcionan en el archivo `aguilas.dat`.

5. En el archivo `reproduccion.dat` se presenta información relativa al éxito reproductor de una población de Águila Calzada en una zona forestal del interior de la Región de Murcia. La matriz contiene el

número de parejas que tuvieron éxito y fracaso en la reproducción entre los años 1998 y 2004, además de la fecha media de puesta (día del mes de marzo) de la población en cada año. En este ejercicio intentaremos explicar la relación éxitos/fracasos (utilizaremos la expresión `cbind(éxito, fracaso)`) en función de la fecha media de puesta (variable `fecha`).

6. En este ejercicio se propone analizar la respuesta de las otras dos especies de aves que aparecen en el archivo`pajaros.dat`: el Escribano Montesino (*Emberiza cia*, variable `embcia`) y el Pinzón Vulgar (*Fringilla coelebs* variable `fricoe`). Seguir el mismo procedimiento que el utilizado con la Collalba Negra, empleando regresiones discretas y evaluando también los modelos gaussianos (`alt + alt2`).

7. 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).