Sección de Matemáticas
Universidad de La Laguna

Paula de Quintana Gómez

Modelo lineal generalizado. Aplicación al conteo del número de ectoparásitos.

Generalized linear models. Application to counting the number of ectoparasites.

Trabajo Fin de Grado
Grado en Matemáticas
La Laguna, Junio de 2020

Dirigido por
María Mercedes Suárez Rancel

María Mercedes Suárez Rancel
Departamento de Matemáticas, Estadística e Investigación Operativa
Universidad de La Laguna
38200 La Laguna, Tenerife

Agradecimientos

Gracias a mi tutora María Mercedes Suárez Rancel por su tiempo y dedicación en la creación de esta memoria.

Gracias a María Mercedes Suárez Rancel, Miguel Molina Borja y María de Fuentes Fernández por dejarme formar parte de su grupo de investigación.

Gracias a mi padre, mi madre y mi hermano, por su apoyo incondicional y creer en mí.

Gracias a María Jesús, por sus descansos eternos en la cafetería.

Gracias a todas las personas que han formado parte de este camino.

Paula de Quintana Gómez
La Laguna, 10 de junio de 2020

Resumen · Abstract

Resumen

En este Trabajo de Fin de Grado se estudian los Modelos Lineales Generalizados. En primer lugar veremos los componentes, la estimación de parámetros y como medir la bondad de ajuste. Además se verá cómo seleccionar el mejor modelo de un conjunto de modelos mediante criterios dedicados a su comparación. Se estudian los criterios de información de Akaike (AIC), el criterio de información Bayesiano (BIC) y la Validación cruzada. Por último se realiza un estudio sobre la carga de ectoparásitos del ácaro Geckobia en dos poblaciones ecológicamente opuestas (norte y y sur de Tenerife).

Palabras clave: *Modelo Lineal Generalizado – AIC – BIC – Sobredispersión – Validación Cruzada*

Abstract

In this end-of-grade project, Generalized Linear Models are studied. First, we will see the components, the parameter estimation and how to measure the goodness of fit. In addition, it will be seen how to select the best model from a set of models using critea dedicated to its comparison. The Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the cross Validation are studied. Finally, a study is carret out on the load of ectoparasites of the Geckobia mite in two ecologically oposed populations (north and south of Tenerife).

Keywords: *Generalized Linear Models – AIC – BIC– Overdispersion – Cross Validation*

Contenido

Introducción

El objetivo de los modelos de regresión lineal es predecir el comportamiento de una variable dependiente en función de variables independientes. Para ello deben cumplirse las siguientes hipótesis: normalidad de la variable dependiente, homocedasticidad, no existencia de autocorrelación y multicolinealidad.

Debido a la alta restricción que tienen estos modelos surgen los Modelos Lineales Generalizados. Son una generalización de los modelos lineales clásicos en los que se permite que la variable de respuesta tenga un modelo de distribución que no sea la normal.

La variable de respuesta debe pertecer a la familia exponencial, por lo que están incluidas distribuciones como la exponencial, Poisson o gamma entre otras.

En el siglo XVIII surgen métodos para tratar los datos en forma de recuentos de eventos motivados por el interés en enumerar las probabilidades de configuración de cartas y dados. En este caso, la distribución básica es la Poisson.

Un ejemplo clásico es el publicado por L. Bortkiewicz en 1989. En él muestra la causa de las muertes producidas en un regimiento de caballería de las guerras prusianas. Se centra en las muertes producidas por coces de caballo. Contó cuantas muertes se produjeron durante 20 años y analizó con qué probabilidad se producían las muertes por año. Pudo observar que los datos se ajustaban a una distribución de Poisson con 0.61 muertes al año.

Con los modelos lineales generalizados se pueden desarrollar modelos para el análisis de recuentos similares a los modelos lineales clásicos para cantidades continuas.

El primer capítulo se centra en el estudio de los modelos lineales generalizados. Se definen sus tres componentes: componente aleatoria, componente sistemática y función de enlace. Se desarrollará la estimación de los parámetros y cómo evaluar los modelos mediante medidas de bondad de ajuste y el estudio de los residuos.

En el segundo capítulo se ve cómo formular el conjunto de modelos candidatos y una vez elegido el conjunto, cómo seleccionar el mejor modelo mediante criterios de información (AIC, BIC). También se estudia la validación cruzada y se comparan los diferentes criterios.

En el tercer capítulo se exponen los resultados del análisis de datos reales, al cuál he sido invitada y he participado junto a mi tutora María Mercedes Suárez Rancel, Miguel Molina Borja y María de Fuentes Fernández. El objetivo es cuantificar el número de parásitos externos de la Tarentola delalandii y analizar su variación entre estaciones y sexos. He participado en la realización del análisis de los datos, para el cuál se han usado los modelos lineales generalizados, en primer lugar con distribución de Poisson, y tras presentarse sobredispersión se ha acudido a la distribución binomial negativa.

1

Modelo lineal generalizado

Este capítulo se centra en el estudio de los GLM. Estos modelos fueron formulados por John Nelder y Robert Wedderburn.

Son una extensión de los modelos lineales clásicos en los que se permiten variables de respuesta cuyos modelos de distribución no son la normal. La variable de respuesta debe seguir cualquier distribución que sea de la familia exponencial, entre las que se encuentran la distribución normal, Poisson, binomial o exponencial entre otras.

Para la realización de este capítulo nos hemos apoyado en [5] y [8].

1.1. Modelos lineales clásicos

Los datos estarán ordenados de la siguiente manera:

- Vector de columnas de las observaciones, $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$.
- Una matriz $\boldsymbol{X}$, $n \times p$ con los valores de p covariables $\boldsymbol{x_1, x_2, ..., x_p}$ como las columnas.
- Parámetros desconocidos $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ asociados a las covariables y con valores arbitrarios. Podemos definir un vector de residuos:

$$\boldsymbol{\epsilon = y - X\beta} \tag{1.1}$$

El modelo quedaría de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon_i, i = 1, ..., n \tag{1.2}$$

O lo que es lo mismo:

$$\boldsymbol{y = X\beta + \epsilon} \tag{1.3}$$

De forma matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Hipótesis

Deben verificarse las siguientes hipótesis:

- $E[\epsilon] = 0$
- La varianza del error debe ser constante.
- Los errores deben ser independientes entre sí.
- $\epsilon \sim N(0, \sigma)$
- Todas las observaciones poseen igual importancia en la estimación de los resultados mínimos.
- Todas las variables deben ser cuantitativas continuas.

En ocasiones los datos no cumplen todas las restricciones que tiene el modelo lineal clásico, en ese caso, se debe aplicar el Modelo Lineal Generalizado.

1.2. Modelos lineales generalizados

1.2.1. Tipos de variables

Se consideran variables dependientes, cuyos valores están afectados por otros factores y covariables. Las variables dependientes pueden ser continuas o discretas, o tomar forma de factores. Los factores son variables cualitativas y las covariables son cuantitativas.

1.2.2. Componentes

El modelo lineal generalizado tiene 3 componentes:

- **Componente aleatoria:** es la variable de respuesta ($\boldsymbol{Y}$) y su distribución de probabilidad. Por ejemplo, en la regresión lineal sería una distribución normal o en la regresión logística binaria una distribución binomial.
- **Componente sistemática:** son las variables explicativas $\boldsymbol{x_1, x_2, ..., x_p}$ que se utilizan en el predictor lineal $\boldsymbol{\eta}$.
- **Función enlace:** especifica el enlace entre componentes aleatorios y sistemáticos. Relaciona el valor esperado de la respuesta con el predictor lineal.

Componente aleatoria

Se trata de la variable de respuesta $\boldsymbol{Y}$ con observaciones independientes $(\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_N})$ con medias $\boldsymbol{\mu}$.

La variable de respuesta $\boldsymbol{Y}$ debe tener una distribución perteneciente a la familia exponencial. La función de densidad debe poder escribirse como:

$$f_Y(y; \boldsymbol{\theta}, \phi) = \exp[\{y\boldsymbol{\theta} - b(\boldsymbol{\theta})\}/a(\phi) + c(y, \phi)] \tag{1.4}$$

donde a, b y c son funciones específicas, ϕ un parámetro de dispersión y $\boldsymbol{\theta}$ un parámetro canónico de la distribución.

En la familia exponencial se encuentran la distribución normal, exponencial, gamma, beta o Poisson entre otras.

Para la distribución normal:

$$f_Y(y; \theta, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[\frac{-(y-\mu)^2}{2\sigma^2}] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[\frac{-(y^2+\mu^2-2y\mu)}{2\sigma^2}] =$$

$$= \exp[\frac{-(y^2+\mu^2-2y\mu)}{2\sigma^2} + \frac{ln(2\pi\sigma^2)}{-2}] = \exp[\frac{-y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{\mu y}{\sigma^2} - \frac{1}{2}(ln(2\pi\sigma^2))] =$$

$$= \exp[(y\mu - \frac{\mu^2}{2})/\sigma^2 - \frac{1}{2}[\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2)]]$$

donde $\theta = \mu$, $b(\theta) = \mu^2/2, \phi = \sigma^2, a(\phi) = \phi$ y $c(y, \phi) = \frac{1}{2}[y^2/\sigma^2 + ln(2\pi\sigma^2)]$.

Se escribe $l(\theta, \phi; y) = lnf_Y(y; \theta, \phi)$ para el logaritmo de verosimilitud considerada como una función de θ y ϕ, entonces la media y la varianza de Y se derivan fácilmente de las relaciones conocidas.

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0$$

y

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\frac{\partial^2 l}{\partial \theta}\right) = 0$$

de la función de densidad (1.4) se tiene

$$l = [y\theta - b(\theta)]/a(\phi) + c(y, \phi)$$

de donde

$$\left(\frac{\partial}{\partial \theta}\right) = [y - b'(\theta)]/a(\phi)$$

y

$$\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -b''(\theta)/a(\phi)$$

luego se tiene

$$0 = E\left(\frac{\partial l}{\partial \theta}\right) = [\mu - b'(\theta)]/a(\phi)$$

entonces

$$E(Y) = \mu = b'(\theta)$$

por lo tanto

$$0 = -\frac{b''(\theta)}{a(\phi)} + \frac{\mathrm{Var}(Y)}{a^2(\phi)}$$

y entonces

$$\mathrm{Var}(Y) = b''(\theta)a(\phi)$$

donde la prima denota diferenciación respecto de θ. Luego la varianza de Y es el producto de una función que depende únicamente del parámetro canónico (por lo tanto de la media) y de otra función independiente de θ y depende solo de ϕ.

A continuación se muestran las características de algunas distribuciones de la familia exponencial:

Tabla 1.1. Características de las distribuciones de la familia exponencial.

Distribución	Rango	a()	b( )	c()	$\mu = E(Y)$	θ	Varianza
Normal	$(-\infty, \infty)$	$\phi = \sigma^2$	$\frac{1}{2}\theta^2$	$-\frac{1}{2}(\frac{y^2}{\phi} + ln(2\pi\phi))$	θ	μ	1
Poisson	Ent$[0, \infty]$	1	e^θ	$-lny!$	e^θ	$ln(\mu)$	μ
Binomial	$[0, n]$	$\frac{1}{n}$	$ln(1 + e^\theta)$	$ln\left[\binom{n}{ny}\right]$	$\frac{e^\theta}{1+e^\theta}$	$ln(\frac{\mu}{1-\mu})$	$\mu(1-\mu)$
Gamma	$(0, \infty)$	$\phi = \upsilon^{-1}$	$-ln(-\theta))$	$(\upsilon - 1)ln(y\upsilon) + +ln\upsilon - ln\tau(\upsilon)$	$-1/\theta$	$-1/\mu$	μ^2
Inversa Gausiana	$(0, \infty)$	$\phi = \sigma^2$	$-(-2\theta)^{1/2}$	$-\frac{1}{2}(ln(2\pi\phi y^3) + \frac{1}{\phi y})$	$(-2\theta)^{1/2}$	$-\frac{1}{2}\mu^2$	μ^3

Componente sistemática

La combinación lineal de las variables explicativas $\boldsymbol{x_1, x_2, ..., x_p}$ cuyos valores se conocen y $\boldsymbol{\beta s}$ cuyos valores se desconocen y son estimados a partir de los datos producen el predictor lineal:

$$\boldsymbol{\eta} = \sum_1^p \beta_j \boldsymbol{x_j}.$$

Si se indexan las observaciones la parte sistemática se escribe del siguiente modo:

$$\eta_i = \sum_{1}^{p} \beta_j x_{ij}; i = 1, ..., N,$$

donde x_{ij} es el valor de la j-ésima covariable para la observación i.

En forma matricial sería de la siguiente manera:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

Hay que tener en cuenta que $\boldsymbol{\eta}$ es $n \times 1$, $\boldsymbol{X}$ es $n \times p$ y $\boldsymbol{\beta}$ es $p \times 1$.

$\boldsymbol{X}$ es la matriz del modelo y $\boldsymbol{\beta}$ es el vector de parámetros.

Función enlace

Se llama $\boldsymbol{\mu}$=$\mathbf{E(Y)}$ al valor esperado de $\mathbf{Y}$. La función de enlace relaciona $\boldsymbol{\mu}$ con el predictor lineal como

$$g(\mu_i) = \eta_i$$

y $g(\ )$ se llama la función de enlace.

Si $g(\mu) = \mu$ esto da lugar al modelo de regresión lineal clásico:

$$\mu = E(Y) = \beta_1 x_1 + ... + \beta_k x_k.$$

Elegir la función de enlace en ocasiones puede ser complicado ya que pueden existir varias funciones de enlaces aplicables en una distribución. Las distribuciones de la familia exponencial tienen una función de enlace especial llamado enlace canónico, que es el que se aplica por defecto a cada distribución. Para elegir la función de enlace que se va a utilizar se compara entre varias funciones de vínculo para el mismo modelo y se utiliza la que produzca el mejor ajuste del modelo a los datos.

El enlace canónico de las distribuciones de la Tabla 1.1 es:

Tabla 1.2. Función de enlace de las distribuciones de la familia exponencial.

Distribución	Función de enlace
Normal	$\boldsymbol{\eta = \mu}$
Poisson	$\boldsymbol{\eta} = ln(\boldsymbol{\mu})$
Binomial	$\boldsymbol{\eta} = ln[\boldsymbol{\mu}/(1-\boldsymbol{\mu})]$
Gamma	$\boldsymbol{\eta = \mu^{-1}}$
Inversa Gausiana	$\boldsymbol{\eta = \mu^{-2}}$

1.2.3. Estimación de parámetros

Esta sección se ha basado en [5, Sección 3.1.].

Una vez elegido el modelo se estiman los parámetros de interés βs.

El logaritmo de verosimilitud de un vector $\boldsymbol{y} = (y_1, ..., y_n)$ escrito en forma canónica es

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi).$$

Por lo tanto

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (\frac{(y_i - b'(\theta_i)}{a(\phi)})\frac{\partial \theta_i}{\partial \boldsymbol{\beta}}$$

pero $\eta_i = g(\mu_i) = \boldsymbol{\beta'} \boldsymbol{x_i}$ y debido a que

$$E(y) = b'(\boldsymbol{\theta}), Var(y) = b''(\boldsymbol{\theta})a(\phi)$$

se tiene que

$$g(b'(\theta_i)) = \boldsymbol{\beta'} \boldsymbol{x}_i$$

$$g(b'(\theta_i))b''(\theta_i)\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_i$$

$$g'(\mu_i)b''(\theta_i)\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_i$$

luego

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{a(\phi)g'(\mu_i)b''(\theta_i)}\boldsymbol{x}_i$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{g'(\mu_i)V_i}\boldsymbol{x}_i$$

donde $V_i = Var(y_i) = a(\phi)b''(\theta_i)$. $\hat{\boldsymbol{\beta}}$ es la solución de $\frac{\partial l}{\partial \boldsymbol{\beta}} = 0$.

Este sistema de ecuaciones en general se resuelve de forma iterativa usando el algortimo de Newon-Raphson:

$$\left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta_1}} \approx \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta_0}} + \left.\frac{\partial^2 l}{\partial \boldsymbol{\beta}\boldsymbol{\beta'}}\right|_{\boldsymbol{\beta_0}} (\boldsymbol{\beta_1} - \boldsymbol{\beta_0})$$

$$0 = \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta_0}} + \left.\frac{\partial^2 l}{\partial \boldsymbol{\beta}\boldsymbol{\beta'}}\right|_{\boldsymbol{\beta_0}} (\boldsymbol{\beta_1} - \boldsymbol{\beta_0})$$

se calcula $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ a partir de $\boldsymbol{\beta_0}$ y así sucesivamente. Este proceso da lugar a $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \to \hat{\boldsymbol{\beta}}$.

El valor de la segunda derivada se sutituye por su valor esperado, a esto se le llama Mínimos cuadrados iterativamente reponderados. Usando

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta'}}\right) = -E\left(\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^2.$$

Se tiene que

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right) = -E\left(\sum_{i^1}^{n} \frac{(y_i-\mu_i)^2}{(g'(\mu_i)V_i)^2}\boldsymbol{x_i}\boldsymbol{x_i'}\right) = -\sum_{i=1}^{n} \frac{V_i}{(g'(\mu_i))^2 V_i^2}\boldsymbol{x_i}\boldsymbol{x_i'} = -\sum_{i=1}^{n} w_i \boldsymbol{x_i}\boldsymbol{x_i'}$$

donde $w_i = 1/V_i(g'(\mu_i))^2$. La expresión anterior en forma matricial puede escribirse como:

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right) = -\boldsymbol{X'WX}$$

donde $\boldsymbol{W}$ es una matriz diagonal con elementos w_i. Por lo tanto, si $\hat{\boldsymbol{\beta}}$ es solución de $\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, entonces $\hat{\boldsymbol{\beta}}$ es asintóticamente normal con media $\boldsymbol{\beta}$ y matriz de covarianzas $(\boldsymbol{X'WX})^{-1}$.

La ecuación ahora es,

$$\boldsymbol{\beta_{new}} = \boldsymbol{\beta_{old}} - \left(E\left(\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right)\right)^{-1}_{\boldsymbol{\beta_{old}}} \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta_{old}}} =$$

$$\boldsymbol{\beta_{old}} + (\boldsymbol{X'WX})^{-1}\boldsymbol{X'W}(\boldsymbol{y}-\boldsymbol{\mu_{old}})\boldsymbol{g'}(\boldsymbol{\mu}_{old}) = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'Wz}$$

donde $\boldsymbol{z} = \boldsymbol{X\beta_{old}} + (\boldsymbol{y}-\boldsymbol{\mu}_{old})\boldsymbol{g'}(\boldsymbol{\mu}_{old}) = \boldsymbol{\eta}_{old} + (\boldsymbol{y}-\boldsymbol{\mu}_{old})\boldsymbol{g'}(\boldsymbol{\mu}_{old})$ y $\boldsymbol{z}$ es el working vector.

1.2.4. Bondad de ajuste

El ajuste de un modelo a los datos se evalua mediante diferentes medidas de bondad de ajuste.

Devianza

La discrepancia se mide entre los valores ajustados y en estudio. La función de verosimilitud del modelo de interés es $l(\hat{\mu}, \phi; y)$. La máxima verosimilitud alcanzada en un modelo completo es $l(y, \phi; y)$. La discrepancia del ajuste será proporcional al doble de la diferencia entre la función de verosimilitud máxima alcanzada y la función de verosimilitud del modelo en estudio.

Denotando $\hat{\theta} = \theta(\hat{\mu})$ y $\tilde{\theta} = \theta(y)$ la estimación de los parámetros canónicos bajo los dos modelos, la discrepancia, asumiendo $a_i(\phi) = \phi/w_i$ sería:

$$\sum 2w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi = D(y;\hat{\mu})/\phi$$

donde $D(y;\hat{\mu})$ es la devianza para el modelo actual.

La devianza escalada se define como:

$$D^*(y;\hat{\mu}) = \frac{D(y;\hat{\mu})}{\phi}.$$

El modelo estará bien ajustado si

$$D^*(y;\hat{\mu}) \sim \chi^2_{n-p}.$$

A continuación se muestran las devianzas para las distribuciones de la Tabla 1.1, los sumatorios de $i = 1, ..., N$

Tabla 1.3. Davianzas de las distribuciones de la familia exponencial.

Distribución	Desviación
Normal	$\sum(y-\hat{\mu})^2$
Poisson	$2\{\sum[yln(y/\hat{\mu}) - (y-\hat{\mu})]\}$
binomial	$2\sum\{[yln(y/\hat{\mu})] + (n-y)ln[(n-y)/(n-\hat{\mu})]\}$
gamma	$2\sum[-ln(y/\hat{\mu}) + (y-\hat{\mu})/\hat{\mu}]$
inversa gausiana	$\sum(y-\hat{\mu})^2/(\hat{\mu}^2 y)$

Estadístico de Pearson

El estadístico de Pearson X^2 es otra medida importante de discrepancia, se define

$$X^2 = \sum(y-\hat{\mu})^2/V(\hat{\mu}),$$

donde $V(\hat{\mu})$ es la estimación de la varianza en la distribución supuesta.

En la distribución normal, X^2 es la suma residual de los cuadrados. En Poisson o binomial es la estadística original de Pearson X^2. En modelos anidados es preferible usar la desviación. En otras circunstancias podría usarse X^2 porque tiene una interpretación más directa.

1.2.5. Residuos

En aquellos modelos cuya variable dependiente sigue una distribución normal los residuos se expresan como $y - \hat{\mu}$. Los residuos se utilizan para comprobar la presencia de valores anómalos. En los modelos lineales generalizados se necesitan residuos generalizados que se puedan aplicar a todas las distribuciones que puedan reemplazar a la normal.

Residuo de Pearson

El residuo de Pearson se define como,

$$r_p = \frac{y-\hat{\mu}}{\sqrt{V(\hat{\mu})}}.$$

El nombre se debe a que para la distribución de Poisson el residuo de Pearson es la raíz cuadrada del estadístico de bondad de ajuste de Pearson X^2, de modo que,

$$\sum r_p^2 = X^2.$$

Residuo de Anscombe

Una desventaja del residuo de Pearson es que la distribución de r_P para las distribuciones no normales a menudo están sesgadas, por lo que pueden no tener propiedades parecidas a la del residuo normal.

Anscombe propuso definir un residual usando una función $A(y)$ en lugar de y, donde $A()$ es elegido para hacer a la función $A(y)$ lo más normal posible.

La función $A()$ viene dada por

$$A() = \int \frac{d\mu}{V^{1/3}(\mu)}.$$

Para la distribución de Poisson se tiene

$$\int \frac{d\mu}{\mu^{1/3}} = 3/2\mu^{2/3},$$

el residual se basa en $y^{2/3} - \mu^{2/3}$. La transformación que 'Normaliza' no estabiliza la varianza, de modo que se debe dividir por la raíz cuadrada de la varianza de $A(Y)$, que es, en primer orden $A'(\mu)\sqrt{V(\mu)}$. Entonces, para la distribución de Poisson, el residuo de Anscombe es dado por

$$r_A = \frac{3/2(y^{2/3} - \mu^{2/3})}{\mu^{1/6}}.$$

Residuo deviance

Si se utiliza la deviance como medida de discrepancia, entonces cada unidad aporta una cantidad d de medida, de modo que

$$\sum d_i = D.$$

Por lo tanto, se define

$$r_D = sgn(y - \hat{\mu})\sqrt{d_i},$$

se tiene una cantidad que aumenta o disminuye con $y - \hat{\mu}$ y para la cual $\sum r_D^2 = D$.

Para la distribución de Poisson se tiene

$$r_D = sgn(y - \hat{\mu})[2(yln(y/\hat{\mu}) - y + \hat{\mu}]^{1/2}.$$

1.3. Ventajas de los GLM sobre la regresión tradicional

- La variable dependiente Y no necesariamente sigue una distribución normal.
- Las herramientas de inferencia y de verificación de modelos que se utilizan en otros modelos como los modelos log-lineal o logística también se pueden usar aquí. Por ejemplo: pruebas de razón de Wald, etc.
- Hay softwares donde se puede utilizar el GLM variando las tres componentes como glm() en R o SPSS.

1.4. Ejemplo

Este ejemplo se encuentra en [9, Ejemplo 15.2]. Una compañía de productos de consumo está estudiando los factores que afectan a la posibilidad de que un cliente canjee un cupón por uno de sus productos de cuidado personal. Se realizó un experimento factorial 2^3 para investigar las siguientes variables:

A - Valor del cupón (Alto, bajo).

B - Periodo de tiempo para la cual el cupón es válido.

C - Facilidad de uso (Fácil, difícil).

Se seleccionó al azar a un total de 1000 clientes para cada una de las 8 celdas del diseño 2^3, y la respuesta es la cantidad de cupones canjeados. Resultados en la Tabla 1.4.

Tabla 1.4. Cantidad de cupones canjeados.

A	B	C	Cantidad de cupones canjeados
-	-	-	200
+	-	-	250
-	+	-	265
+	+	-	347
-	-	+	210
+	-	+	286
-	+	+	271
+	+	+	326

La variable dependiente Y es cantidad de cupones canjeados.

Y los factores son x_1 el valor del cupón, x_2 la longitud de tiempo para el cual el cupón es válido y x_3 la facilidad de uso.

El modelo es el siguiente

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

Se puede pensar en la respuesta como el número de éxitos de 1000 ensayos de Bernouilli en cada celda de diseño, por lo que un modelo razonable para la respuesta es un modelo lineal generalizado con una distribución de respuesta binomial y un enlace logit. Esta forma particular de GLM generalmente se llama regresión logística.

En un GLM se tiene

$$f_\theta(Y) = exp[\{y\theta - b(\theta)\}/a(\phi) + c(y,\theta)],$$

$$\boldsymbol{\eta} = \sum_{1}^{p} \beta_j \boldsymbol{x_j}.$$

Para la distribución binomial con enlace logit:

$$\begin{aligned}
&a() = \frac{1}{n} \\
&b() = ln(1 + e^\theta) \\
&c() = ln[\binom{n}{ny}] \\
&\mu = E(Y) = \frac{e^\theta}{1 + e^\theta} = \frac{1}{1 + exp(-X\beta)} \\
&\eta = log[\frac{\mu}{1-\mu}]
\end{aligned}$$

Los experimentadores decidieron ajustar un modelo que involucra solo los efectos principales y las interacciones de dos factores. Por lo tanto, el modelo para la respuesta esperada es

$$E(Y) = \frac{1}{1 + exp[-(-\beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i<j}^{2} \sum_{j=2}^{3} \beta_{ij} x_i x_j)]}. \tag{1.5}$$

En la Tabla 1.5 se tiene una salida de los datos. La tabla se ajusta al modelo completo que involucra los tres efectos principales y las tres interacciones de dos factores.

Los p-valores en la tabla indican que la intercepción, los efectos principales de A y B, y la interacción BC son significativos, pues los p-valores son menores que 0.05.

En la Tabla 1.6 se muestra la sección de bondad de ajuste, que presenta dos estadísticas de prueba diferentes (Pearson y Deviance) que miden la adecuación general del modelo. Todos los p-valores para estas estadísticas de bondad de ajuste son grandes, lo que implica que el modelo es satisfactorio.

Tabla 1.5. Resultados del análisis.

Predictor	Coef.	SE Coef	p-valor
Constant	-1.01154	0.0255150	0.000
A	0.169208	0.0255092	0.000
B	0.169622	0.0255150	0.000
C	0.0233173	0.0255099	0.361
A*B	-0.0062654	0.0255122	-0.25
A*C	-0.0027726	0.0254324	0.913
B*C	-0.0410198	0.0254339	0.107

Tabla 1.6. Bondad de ajuste.

Método	Chi-cuadrado	DF	p-valor
Pearson	1.46458	1	0.226
Devianza	1.46451	1	0.226

En la Tabla 1.7 se presenta el análisis de un modelo reducido que contiene los tres efectos principales y la interacción BC (se incluyó el factor C para mantener la jerarquía del modelo).

Tabla 1.7. Resultados del análisis.

Predictor	Coef.	SE Coef	p-valor
Constant	-1.01142	0.255076	0.000
A	0.168675	0.0254235	0.000
B	0.169116	0.0254321	0.000
C	0.0230825	0.0254306	0.364
B*C	-0.04109711	0.0254307	0.107

Tabla 1.8. Bondad de ajuste.

Método	Chi-cuadrado	DF	p-valor
Pearson	1.53593	3	0.674
Devianza	1.53593	3	0.674

El modelo ajustado es

$$\hat{y} = \frac{1}{1 + exp[-(-1.01 + 0.169x_1 + 0.169x_2 + 0.023x_3 - 0.041x_2x_3)]} = \frac{1}{1 + exp(+1.01 - 0.169x_1 - 0.169x_2 - 0.023x_3 + 0.041x_2x_3)}$$

Dado que los efectos de C y la interacción BC son muy pequeños, estos términos probablemente podrían eliminarse del modelo sin mayores consecuencias.

2

Selección de modelos

En el capítulo anterior se ha visto cómo estimar los parámetros y estudiar su presición. En este capítulo se determina qué conjunto de modelos candidatos se deben considerar y cómo seleccionar el mejor modelo entre todos los candidatos.

Hasta principio de los años setenta no se consideraba especificar el modelo como un trabajo de los estadísticos. Encontrar un modelo adecuado es muy importante ya que si el modelo es inadecuado la inferencia será deficiente. A continuación se muestran métodos de selección de modelos con gran apoyo teórico y fáciles de usar.

Hemos empleado como bibliografía de apoyo [1], [2], [4] y [7].

2.1. Formulación de modelos candidatos

Formular el conjunto de modelos candidatos es una de las tareas más importantes y difíciles. Para ello, los científicos deben estar capacitados y tener experiencia ya que se trata de una tarea subjetiva. Debe coordinarse la información del científico y la información del estadístico.

El conjunto de modelos candidatos debe incluir un modelo global. Se trata de un modelo con muchos parámetros que incluye todos los efectos relevantes. Debemos comprobar si el modelo se ajusta a los datos y realizar el análisis si el ajuste es aceptable. Del modelo global derivaremos modelos con menos parámetros. Se trata de crear alternativas basadas en lo que se conoce. Estos modelos tendrán diferente número de parámetros.

Chatfield considera importante que para elegir el conjunto de modelos candidatos debe influir la información previa, el conocimiento de los expertos, la limitación en los parámetros del modelo, las variables, etc. Si un modelo no responde a la realidad no debemos incluirlo.

El conjunto de modelos candidatos debe ser pequeño y con hipótesis plausibles, pero lo suficientemente grande como para no omitir un modelo muy bueno.

Una vez elegido el conjunto de modelos candidatos, debemos escoger el mejor, para ello hacemos uso de diferentes criterios.

2.2. Criterio de información de Akaike, AIC

En primer luegar se expone el criterio de información de Akaike. Tanto este criterio como el criterio de información Bayesiano miden la capacidad explicativa del modelo y a su vez penalizan su complejidad. Se utiliza cuando se tiene que elegir un modelo entre una gran cantidad de modelos.

Mide la bondad de ajuste a partir de la máxima verosimilitud. La complejidad la mide a partir del número de parámetros. No es muy restrictivo en cuanto a complejidad, pues si el modelo tiene muchos datos no nos penalizará. La fórmula es la siguiente:

$$AIC = 2K - 2\ln(\mathrm{L}),$$

donde K es el número de parámetros independientes que hay en el modelo estadístico y L es la función de máxima verosimilitud.

El término $-2\ln(\mathrm{L})$ es conocido como la desviación. Akaike pensó en AIC como desviación más 2K.

Se debe calcular AIC para cada modelo y seleccionar el modelo con el menor valor de AIC como el mejor.

Con el primer término se penaliza el número de parámetros con el Principio de Parsimonia, cuanto más aumenta el número de parámetros mayor será el valor de AIC. Con el segundo término se mide la bondad de ajuste.

Penalizar 2K es similar a hacer validación cruzada dejando un dato fuera.

AIC es un criterio bastante fácil a la hora de implementarse.

La diferencia entre los valores de AIC es significativa. Se define Δ_i como

$$\Delta_\mathrm{i} = \mathrm{AIC_i} - \mathrm{AIC_{min}}, \mathrm{i} = 1...\mathrm{r}$$

donde AIC_{min} es el menor valor de AIC para los modelos comparados. Ese modelo será el mejor del conjunto de modelos.

Si $\Delta_i \leq 2$ tiene soporte empírico sustancial. Si Δ_i está entre 4-7 tienen soporte empírico considerablemente menor. Si $\Delta_i \geq 10$ el modelo no compite por ser el mejor.

2.3. AICc

AICc es una variante de AIC que se utiliza cuando los tamaños de muestra son pequeños, pues AIC deja de ser fiable en estos casos.

La fórmula es la siguiente:

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2\mathrm{K} \cdot (\mathrm{K}+1)}{\mathrm{n}-\mathrm{K}-1}$$

donde K es el número de parámetros independientes que hay en el modelo estadístico y n es el tamaño de la muestra.

Burham y Anderson recomiendan su uso cuando $\frac{n}{K} < 40$. Para tomar esta decisión se debe elegir el K de las dimensiones más altas, es decir, del modelo global del conjunto de candidatos.

A medida que aumenta el tamaño de la muestra AICc converge a AIC, por lo que si la muestra es muy grande se comportan de forma parecida.

Para clasificar los modelos según la perdida de información debemos calcular las diferencias AICc, que son los valores Δ.

$$\Delta_\mathrm{i} = \mathrm{AICc_i} - \mathrm{AICc_{min}}, \mathrm{i} = 1...\mathrm{r}$$

$\mathrm{AICc_{min}}$ es el mínimo valor de AICc para los modelos comparados.

Los modelos cuyo Δ es superior a 9-11 tienen poco soporte, es decir, pierden demasiada información. Los que tengan Δ mayor que 20 no tienen soporte empírico.

Se debe calcular todos los valores de AICc y Δ y seleccionar el que tenga la menor pérdida de información.

2.4. Criterio de información Bayeasiano, BIC

El criterio de información bayesiano se trata de un criterio parecido a AIC propuesto por Schwarz. Se trata de una modificación de AIC pero más restrictivo. Se utiliza para seleccionar un modelo entre una cantidad finita.

Tiene la siguiente fórmula:

$$BIC = -2 \cdot \ln(\mathrm{L}) + \mathrm{K}\ln(\mathrm{n})$$

donde,

- n es el número de datos o tamaño de la muestra,
- K es el número de parámetros,
- L es la función de máxima verosimilitud del modelo.

Al igual que AIC, cuanto menor sea su valor mejor será el modelo. En este caso, el número de parámetros se penaliza con $K\ln(\mathrm{n})$, por lo que es mayor a la penalización que hace AIC.

2.5. Validación cruzada

La validación cruzada se utiliza para la comparación de modelos y por tanto para la selección del modelo.

Se trata de un método estadístico que evalúa y compara algoritmos diviendo datos en dos segmentos. Uno de los segmentos se utiliza como datos de entrenamiento y otro como datos para validar el modelo. La validación cruzada es práctica si tenemos menos de 20 modelos ya que computacionalmente es muy dificil.

A continuación se muestran los tres tipos de validación cruzada.

2.5.1. Validación cruzada de k iteraciones

En la validación cruzada de k iteraciones los datos se dividen en k segmentos de igual tamaño. Posteriormente se realizan k iteraciones de entrenamiento y validación. En cada iteración se utiliza un pliegue diferente de datos. Lo más común es utilizar k=10.

En cada iteración se usan k-1 pliegues de datos para aprender y posteriormente utilizamos los aprendidos para hacer predicciones sobre los datos.

2.5.2. Validación cruzada aleatoria

En la validación cruzada aleatoria los datos se reorganizan en cada ronda por lo que aumenta el número de estimaciones.

Así se obtiene una estimación y comparación confiables pues se hace un gran número de estimaciones.

2.5.3. Validación cruzada dejando un dato fuera

En la validación cruzada dejando un dato fuera en cada iteración todos los datos, excepto una observación, se utilizan en el entrenamiento y el modelo se prueba solo en una observación.

2.6. Diferencia entre criterios

En esta sección se presentan diferencias entre los criterios que se acaban de presentar. En primer lugar entre los criterios AIC y BIC. Posteriormente se incluye también validación cruzada.

2.6.1. Diferencias entre AIC y BIC

Son los criterios más utilizados en cuanto a la selección del modelo, pero, debemos saber cuál utilizar en cada caso.

- Una diferencia destacable es el término de penalización. La penalización que hace BIC es mucho mayor que la que hace AIC, ya que en el primer caso penaliza con ln y en el segundo con 2 por el número de observaciones. Por ello BIC se debe utilizar en modelos con menos datos que AIC.
- Otra diferencia es el objetivo de los criterios. El objetivo de BIC es acercarse al modelo real mientras AIC selecciona el mejor modelo entre los modelos dados. Por lo tanto BIC supone que el modelo real está entre los candidatos, mientras que en AIC esta afirmación no es cierta.

Se concluye que BIC penaliza los modelos con muchos parámetros por lo que al utilizarlo se obtienen modelos con menos parámetros que si se utiliza AIC.

2.6.2. Diferencias entre AIC, BIC y Validación Cruzada

- Una de las cosas que se debe tener en cuenta es que tanto AIC como BIC son mucho más fáciles de usar que Validación Cruzada, que computacionalmente lleva más tiempo.
- AIC y BIC sólo se pueden utilizar en modelos que se estiman con la función de máxima verosimilitud. Si este no es el caso entonces se debe utilizar Validación Cruzada.
- Anteriormente se ha visto que AIC es asintóticamente equivalente a Validación Cruzada si penalizamos $2k$. BIC también puede ser equivalente a Validación Cruzada k-veces bajo la aplicación de algunos supuestos.

3

Aplicación del modelo lineal generalizado al conteo del número de ectoparásitos

La carga individual del parásito depende de varios factores como el sexo, el tamaño corporal o las condiciones climáticas. A su vez, los parásitos pueden presentar varias patologías a corto y largo plazo. En este trabajo, se analiza la carga de ectoparásitos del ácaro Geckobia en dos poblaciones ecológicamente opuestas (norte y sur de Tenerife) del geco (Tarentola delalandii). Para este propósito, se realizan transsectos aleatorios para capturar geckos debajo de las piedras en cada población y se cuentan todos los ácaros encontrados en cualquier parte del cuerpo de cada gecko. Los resultados de la aplicación de modelos lineales generalizados mostraron que no hubo efectos significativos en el número de ectoparásitos de: población, estación, sexo (dentro de la población), índice de condición o temperatura de los refugios donde se encontraron geckos. Sin embargo, hubo efectos significativos de las interacciones de estación por índice de condición y de estación por sexo (anidado dentro de la población). El parasitismo fue mayor en individuos con valores de índice de cuerpo inferior en otoño-invierno que en aquellos con mayor condición corporal en primavera-verano. Además, las hembras de la población del norte estaban más parasitadas que los machos y los juveniles en los meses más fríos del año.

3.1. Introducción

La parasitosis en los reptiles es una de las principales causas de mortalidad entre estos animales cuando se mantienen en recintos; sin embargo, en el ambiente natural, la relación parásito-huésped tiende a estar en equilibrio.

Los principales factores que se han descrito como influyentes en el grado de parasitismo de una muestra son: su sexo, estado hormonal, condición reproductiva y comportamiento. Además, el número de ectoparásitos también puede verse influido por la interacción física entre los individuos, como a través del contacto sexual, la lucha, la anidación comunitaria o los lugares de retiro.

En el caso de los geckos, se han reportado varios parásitos para diferentes géneros.

La región geográfica y la estación o el año también pueden afectar las cargas parasitarias.

Tarentola delalandii es un pequeño gecko, predominantemente nocturno, que se extiende por toda Tenerife y La Palma (Islas Canarias). Este gecko es actualmente muy común en la zona costera, aunque rara vez se ve por encima de los 1500 m de altitud. Existen pocos estudios sobre este gecko, pero sus principales depredadores son: búhos, cernícalos, ratas y erizos durante el anochecer y la noche, cernícalos y alcaudones durante el día. Y gatos en cualquier momento.

El objetivo para este estudio fue cuantificar el número de parásitos externos y analizar su variación entre estaciones y sexos en dos poblaciones seleccionadas (norte y sur de la isla) que tienen características ecológicas contrastadas.

Se presupone que el número de parásitos externos debería variar a lo largo de los meses de estudio, con valores más altos en primavera-verano. Además, los machos estarían más parasitados que las hembras y los juveniles. Específicamente, en reptiles, la intensidad y el dominio de las infestaciones de ectoparásitos tienden a ser menores en las hembras. Se ha informado que la testosterona masculina puede tener un efecto inmunosupresor y los machos que tienen niveles altos de testosterona tienden a tener una mayor movilidad, lo que implicaría una mayor probabilidad de estar expuesto a parásitos de otros congéneres durante encuentros con hembras o durante peleas con otros individuos. Por otro lado, una mayor cobertura de plantas se ha correlacionado con una mayor probabilidad de infestación por ácaros, probablemente porque la vegetación anual aumenta la exposición pero varía según la temporada.

También se plantea la hipótesis de que los individuos de la población del sur (temperaturas ambientales más altas) podrían estar más parasitados que los de la población del norte. En general, se acepta que la temperatura es una de las variables ambientales más importantes que afectan la distribución del parásito, ya que el aumento de las temperaturas afecta positivamente el desarrollo, la reproducción y la tasa de transmisión del parásito. Por lo tanto, las condiciones climáticas de la zona sur podrían favorecer a los ácaros Geckobia más que los de la zona norte.

Se plantea la hipótesis de que los individuos con una condición corporal inferior estarían más infestadas que aquellas con una condición corporal superior, especialmente en la temporada otoño-invierno después de que los geckos hayan alcanzado un alto nivel esfuerzo reproductivo durante el período anterior primavera-verano.

3.2. Datos y variables

Antes de ver en detalle las variables que se utilizan en el análisis se describe como fueron recogidos los datos.

3.3. Áreas de estudio

El trabajo se centra en dos ecosistemas contrastantes en la isla, uno en latitud media en el norte y otro en una latitud baja en el sur. Se procede a recopilar datos en dos poblaciones: 1) Geneto (San Cristóbal de La Laguna, norte de la isla) y 2) El Médano (Granadilla de Abona, en el sur). El muestreo de campo se realizó durante los meses de abril a julio de 2013 y 2015 y de octubre de 2014 a enero de 2015. Duró 30 horas, aproximadamente, por mes en cada área. El trabajo de campo se realizó entre las 10:00 h y las 17:00 h, cuando los geckos podían estar inactivos debajo de las piedras.

Cada localidad tiene un clima, geología y vegetación diferentes, representantes de las partes norte y sur de la isla.

3.4. Variables

La base de datos está formada por 538 individuos. Se utiliza el número de parásitos de cada individuo como variable dependiente en función de las variables independientes que se dividen en factores y covariables. En los factores se encuentran la población, el sexo (anidado en población) y el período del año. En las covariables el índice de condición y la temperatura del refugio.

3.5. Análisis de datos

Como había una gran proporción de geckos que no presentaban ningún ectoparásito, estos casos no se consideraron para los siguientes análisis.

Una vez incorporados los archivos a la computadora, los datos se analizaron utilizando Modelos Lineales Generalizados con distribución Poisson y enlace logaritmo, tomando el número de ácaros como variable dependiente y población, sexo (anidado dentro de la población) y período del año (estación) como factores e índice de condición (IC) y temperatura del refugio como covariables.

Con este análisis se descubre que existe sobredispersión.

Una forma de solucionar la sobredispersión es recurrir al modelo binomial negativo. En este caso, se analizan los datos utilizando Modelos Lineales Generalizados con distribución binomial negativa y enlace de logaritmo.

En los análisis que se realizan se consideran los siguientes modelos: 1) incluyendo todos los factores, covariables, sexo anidado dentro de la población e interacciones de orden 2; 2) incluyendo solo interacciones significativas que aparecen en el modelo 1; 3) lo mismo que en el modelo 1 pero esta vez, sin tener en cuenta algunos especímenes que tenían un gran número de ectoparásitos (más de 100); 4) incluyendo solo interacciones significativas que aparecen en el modelo 3.

Para la selección del modelo, se utiliza el criterio de información de Akaike (AIC) que permitió seleccionar qué modelo respaldaba sustancialmente los datos. Se utiliza una derivada de segundo orden (AICc) que contiene un término de corrección de sesgo para tamaños de muestra pequeños que debe usarse cuando el número de parámetros libres, K, excede n / 40 (donde n es el tamaño de la muestra). Se presentan las diferencias de AICc (ΔAICc = AICc - AICmin) para comparar los resultados de múltiples modelos y se utiliza un límite de $\Delta \leq 2$ para incluir solo aquellos modelos con un apoyo sustancial de los datos.

Para detectar si los gecos diferían en la condición corporal entre las poblaciones, las estaciones y el sexo, y como los datos no cumplían con los requisitos paramétricos, se aplica nuevamente un GLM con distribución normal y enlace identidad usando IC como la variable dependiente y la población, la estación y el sexo como factores. Se realizan análisis considerando los siguientes modelos: 1) incluyendo los tres factores e interacciones de orden 2; 2) incluyendo solo los tres factores; 3) lo mismo que en el modelo 1 pero considerando el sexo anidado dentro de la población; 4) incluyendo solo la estación, población y sexo anidados dentro de la población.

3.6. Resultados

En la Tabla 3.1 se presenta la media, el error estándar, el mínimo, máximo y el tamaño de la muestra de la variable dependiente número de parásitos.

Se observa que la media es 10,09 y la varianza 350,749, por lo que se empieza a cuestionar el problema de sobredispersión.

Tabla 3.1. Media, error estándar (S.E.), valores mínimos y máximos y tamaño de muestra (N) de la variable dependiente Número de parásitos.

Media	S.E.	Min-Max	N	Varianza
10,09	1,100	1-194	290	350,749

En la Tabla 3.2 se presenta la media, el error estándar, el mínimo, máximo y el tamaño de la muestra de geckos muestreados en cada sitio y estación. De un total de 538 individuos muestreados, 239 (44.4 %) no tenían ningún ácaro.

En la Tabla 3.3 se presenta la proporción de individuos infestados de cada población y período de muestreo.

Tabla 3.2. Media, error estándar (S.E.), valores mínimos y máximos y tamaño de muestra (N) de SVL y BM en machos, hembras y juveniles de las dos poblaciones muestreadas de T. delalandii.

Población	Sexo		SVL(mm)	BM(g)
Norte	Machos	Media	60.67	8.2
		S.E.	0.71	0.29
		Min-Max	50-74	4.3-15.6
		N	54	54
	Hembras	Media	53.86	5.7
		S.E.	0.63	0.21
		Min-Max	46-62	3.2-9.5
		N	50	50
	Juveniles	Media	39.75	2.72
		S.E.	1.69	0.32
		Min-Max	27-48	0.9-4.6
		N	16	16
Sur	Machos	Media	49.95	4.41
		S.E.	0.54	0.15
		Min-Max	39-60	2-7.5
		N	59	59
	Hembras	Media	46.88	3.8
		S.E.	0.37	0.13
		Min-Max	39-53	1.6-10
		N	72	72
	Juveniles	Media	35.34	2.09
		S.E.	0.76	0.17
		Min-Max	25-45	0.7-4.6
		N	44	44

Tabla 3.3. Número total de individuos muestreados en cada período de estudio, por categorías de individuos y población; entre paréntesis el porcentaje de los infestados por ectoparásitos externos.

		Machos	Hembras	Juveniles	Total
Otoño-invierno	Norte	38	24	22	84
		(57.9)	(70.8)	(31.8)	(54.8)
	Sur	26	32	35	93
		(76.9)	(65.6)	(40)	(59.1)
Primavera-verano	Norte	59	66	36	161
		(55.9)	(51.5)	(25)	(47.2)
	Sur	61	82	57	200
		(63.9)	(64.6)	(52.6)	(61)

En la Tabla 3.4 se presenta el número de geckos que se esperaba que estuvieran parasitados y en número de geckos realmente parasitados en otoño-invierno y primavera-verano de la población del norte. En otoño-invierno se esperaban parási-

tados 41.8 y estaban parasitados 46 por lo que hubo hubo más geckos parasitados de lo esperado. En primavera-verano hubo 76 parasitados y se esperaba que hubiera 80.2 por lo que en este caso hay menos geckos no parasitados de lo esperado.

En la Tabla 3.5 se presentan los mismos datos que en la Tabla 3.4 pero con la problación del sur. En este caso se obtienen los resultados al contrario.

Tabla 3.4. Número de gecos parasitados o no, recuentos y porcentajes esperados, en cada estación muestreada de la población del norte.

	Machos	Parasitos	No parasitos	Total
Estación Otoño-invierno	Recuento	46	38	84
	Recuento esperado	41.8	42.2	84
	% dentro de la estación	54.8 %	45.2 %	100 %
Primavera-verano	Recuento	76	85	161
	Recuento esperado	80.2	80.8	161
	% dentro de la estación	47.2 %	52.8 %	100 %

Tabla 3.5. Número de gecos parasitados o no, recuentos y porcentajes esperados, en cada estación muestreada de la población del sur.

	Machos	Parasitos	No parasitos	Total
Estación Otoño-invierno	Recuento	55	38	93
	Recuento esperado	56.2	36.8	93
	% dentro de la estación	59.1 %	40.9 %	100 %
Primavera-verano	Recuento	122	78	200
	Recuento esperado	120.8	79.2	200
	% dentro de la estación	61.0 %	39.0 %	100 %

Ahora se aplica GLM con distribución normal y enlace logarítmico sobre el índice de condición de geckos. También se aplica distribución normal y enlace identidad, que se observa que produce un mejor ajuste del modelo a los datos, por lo que se utiliza esta función de enlace. Los resultados mostraron que el modelo 2 (Tabla 3.6) era el más apropiado (AICc más bajo). Con base en este modelo, el IC difirió principalmente entre estaciones y entre poblaciones, sin embargo, no difirió entre sexos. (Tabla 3.7).

Tabla 3.6. Valores de AICc para los diferentes modelos analizados con índice de condición de gecko como variable dependiente.

Modelo	AICc	ΔAICc
1	548,65	9,97
2	538,68	0
3	552,65	13,97
4	542,67	3,99

Tabla 3.7. Resultados del análisis del modelo lineal generalizado con distribución normal y enlace identidad aplicado al índice de condición usando el modelo con los tres factores. En negrita efectos significativos de factores.

	Chi-cuadrado de Wald	df	p-valor
Intersección	0,042	1	0,838
Población	7,865	1	0.005
Estación	17,420	1	0.000
Sexo	1,873	2	0,392

A partir de ahora se considera el número de ectoparásitos como variable dependiente. A continuación se analizan a los datos con distribución de Poisson y enlace logaritmo.

En la Tabla 3.8 se presenta el estadístico de desviación, el estadístico chi-cuadrado de Pearson, el criterio de información de Akaike (AIC), entre otros, para el modelo con distribución de poisson y enlace logaritmo.

Los estimadores de dispersión más utilizados son la relación entre el estadístico de Pearson χ^2 a sus correpondientes grados de libertad y la relación de la función de desviación D y sus grados de libertad (gl):

$$\frac{\chi^2}{gl}$$

ó

$$\frac{D}{gl}.$$

La relación será 1 si se cumple la propiedad de equidispersión, es decir, la $Var(Y) = E(Y)$. Si es mayor que 1 indica sobredispersión y si es menor que 1 indica infradispersión.

El estadístico de desviación tiene un valor de $3206,421$ que se evalua en la siguiente relación

$$\frac{D}{gl} = \frac{3206,421}{263} = 12,192 > 1$$

El modelo presenta sobredispersión.

Tabla 3.8. Resultados de la bondad de ajuste utilizando la distribución de Poisson con enlace logaritmo.

Bondad de ajuste			
	Valor	gl	Valor/gl
Desvianza	3206,421	263	12,192
Desvianza escalada	207,328	263	
Chi-cuadrado de Pearson	4067,421	263	15,465
Chi-cuadrado de Pearson escalado	263,000	263	
Logaritmo de verosimilitud	-2097,846		
Criterio de información de Akaike (AIC)	4249,692		
AIC corregido para muestras finitas (AICC)	4348,779		
Criterio de información bayeasiana (BIC)	4348,779		

Para solucionar la presencia de sobredispersión se acude a la distribución binomial negativa con enlace logaritmo.

Se presentan a continuación los resultados del primer modelo.

En la Tabla 3.9 se muestran las pruebas de efectos del modelo que incluye los factores, covariables, sexo anidado en población e interacciones de orden 2.

El análisis mostró que el número de ectoparásitos en T. delalandii se ve significativamente afectado por la interacción de la estación por IC.

En la Tabla 3.10 se presentan las estadísticas de bondad de ajuste del modelo anterior, el estadístico de desviación, el estadístico chi-cuadrado de Pearson y el criterio de información de Akaike (AIC), entre otros.

Se observa que el valor de AICc que se utilizará para la comparación es 1899,096.

Tabla 3.9. Resultados del análisis del modelo lineal generalizado con distribución binomial negativa y enlace logaritmo aplicado a los números de ectoparásitos de T. delalandii usando el modelo con factores, covariables, interacciones de orden 2 y sexo anidado en población. En negrita efectos significativos de factores e interacciones.

	Chi-cuadrado de Wald	df	p-valor
Intersección			
Población	0.122	1	0.727
Estación	0.077	1	0.781
Temperatura	0.259	1	0.611
IC	0.037	1	0.848
Población*Estación	1.011	1	0.315
Población*Temperatura	0.024	1	0.876
Población*IC	3.045	1	0.081
Estación*Temperatura	0.093	1	0.761
Estación*IC	7.832	1	0.005
Temperatura*IC	0.000	1	0.993
sexo(población)	3.389	4	0.495
estación*sexo(población)	7.863	4	0.097
sexo(población)*Temperatura	5.618	4	0.230
Sexo(Población)*IC	3.067	4	0.547

Tabla 3.10. Resultados de la bondad de ajuste utilizando la distribución de Binomial Negativa con enlace logaritmo usando el modelo con factores, covariables, interacciones de orden 2 y sexo anidado en población.

Bondad de ajuste			
	Valor	gl	Valor/gl
Desvianza	289,783	263	1,102
Desvianza escalada	223,767	263	
Chi-cuadrado de Pearson	340,591	263	1,295
Chi-cuadrado de Pearson escalado	263,000		
Logaritmo de verosimilitud	-919,663		
Criterio de información de Akaike (AIC)	1893,325		
AIC corregido para muestras finitas (AICC)	1899,096		
Criterio de información bayeasiana (BIC)	1992,412		

Ahora se analiza un nuevo modelo que incluye solo los efectos significativos del modelo anterior, es decir, estación por IC.

En la Tabla 3.11 se presentan las pruebas de efectos del modelo que incluye los efectos significativos del modelo anterior. En la 3.12 se muestran sus pruebas de bondad de ajuste.

Tabla 3.11. Resultados del análisis del modelo lineal generalizado con distribución binomial negativa y enlace logaritmo aplicado a los números de ectoparásitos de T. delalandii usando el modelo con factores, covariables, interacciones de orden 2 y sexo anidado en población significativas. En negrita efectos significativos de factores e interacciones.

	Chi-cuadrado de Wald	df	p-valor
Intersección	562,784	1	0,000
Estación*IC	7,899	2	0.019

Tabla 3.12. Resultados de la bondad de ajuste utilizando la distribución de Binomial Negativa con enlace logaritmo usando el modelo con factores, covariables, interacciones de orden 2 y sexo anidado en población significativas.

Bondad de ajuste			
	Valor	gl	Valor/gl
Desvianza	379,818	287	1,323
Desvianza escalada	158,105	287	
Chi-cuadrado de Pearson	689,463	287	2,402
Chi-cuadrado de Pearson escalado	287,0000	287	
Logaritmo de verosimilitud	-964,680		
Criterio de información de Akaike (AIC)	1935,360		
AIC corregido para muestras finitas (AICC)	1935,444		
Criterio de información bayeasiana (BIC)	1946,370		

Ahora se repiten los modelos pero no tenemos en cuenta algunos especímenes que tenían un gran número de ectoparásitos (más de 100).

En la Tabla 3.13 se presentan la media, el error estándar, el mínimo, máximo y el tamaño de la muestra de la variable dependiente número de parásitos con los nuevos datos.

La media es 8.69 y la varianza 149,538 por lo que se empieza a sospechar que vuelve a haber sobredispersión.

Tabla 3.13. Media, error estándar (S.E.), valores mínimos y máximos y tamaño de muestra (N) de la variable dependiente Número de parásitos en la muestra sin los especímenes con mas de 100 ectoparásitos.

Media	S.E.	Min-Max	N	Varianza
8.69	0.722	1-84	287	149.538

En la Tabla 3.14 se presentan el estadístico de desviación, el estadístico chi-cuadrado de Pearson y el criterio de información de Akaike (AIC), entre otros, para el modelo con distribución de poisson y enlace logaritmo con los nuevos datos.

Se comprueba así si los datos presentan sobredispersión.

El estadístico de desviación tiene un valor de $2544,597$ que se evalua en la siguiente relación,

$$\frac{D}{gl} = \frac{2544,597}{260} = 9,787 > 1.$$

El modelo presenta sobredispersión.

Tabla 3.14. Resultados de la bondad de ajuste utilizando la distribución de Poisson con enlace logaritmo en la muestra sin los especímenes con mas de 100 ectoparásito.

Bondad de ajuste

	Valor	gl	Valor/gl
Desvianza	2544,597	260	9,787
Desvianza escalada	211,168	260	
Chi-cuadrado de Pearson	260,000	260	12,050
Chi-cuadrado de Pearson escalado	-1756,754		
Logaritmo de verosimilitud	-145,788		
Criterio de información de Akaike (AIC)	3567,509		
AIC corregido para muestras finitas (AICC)	3573,347		
Criterio de información bayeasiana (BIC)	3666,315		

Para solucionarla se acude nuevamente a la distribución binomial negativa con enlace logaritmo.

En la Tabla 3.15 se presentan las pruebas de efectos del modelo que incluye los factores, covariables, sexo anidado en población e interacciones de orden 2 para la muestra sin los especímenes con mas de 100 ectoparásitos.

El análisis mostró que el número de ectoparásitos en T. delalandii se ve significativamente afectado por la interacción de la estación por sexo anidado en la población. También hubo un efecto significativo de la interacción de la estación por IC.

En la Tabla 3.16 se presentan las estadísticas de bondad de ajuste del modelo anterior, el estadístico de desviación, el estadístico chi-cuadrado de Pearson y el criterio de información de Akaike (AIC), entre otros.

El valor de AICc es 1840,390.

Tabla 3.15. Resultados del análisis del modelo lineal generalizado con distribución binomial negativa y enlace logaritmo aplicado a los números de ectoparásitos de T. delalandii usando el modelo con factores, covariables, interacciones de orden 2 y sexo anidado en población para la muestra sin los especímenes con mas de 100 ectoparásitos. En negrita efectos significativos de factores e interacciones.

	Chi-cuadrado de Wald	df	p-valor
Intersección			
Población	0.103	1	0.748
Estación	1.800	1	0.180
Temperatura	0.018	1	0.892
IC	0.333	1	0.564
Población*Estación	1.715	1	0.190
Población*Temperatura	0.060	1	0.806
Población*IC	0.583	1	0.445
Estación*Temperatura	0.516	1	0.472
Estación*IC	4.259	1	0.039
Temperatura*IC	0.060	1	0.806
sexo(población)	3.357	4	0.500
estación*sexo(población)	9.661	4	0.047
sexo(población)*Temperatura	4.221	4	0.377
Sexo(Población)*IC	4.999	4	0.287

Tabla 3.16. Resultados de la bondad de ajuste utilizando la distribución de Binomial Negativa con enlace logaritmo usando el modelo con factores, covariables, interacciones de orden 2 y sexo anidado en población para la muestra sin los especímenes con mas de 100 ectoparásitos.

Bondad de ajuste			
	Valor	gl	Valor/gl
Desvianza	266,714	260	1,026
Desvianza escalada	227,465	260	
Chi-cuadrado de Pearson	304,863	260	1,173
Chi-cuadrado de Pearson escalado	260,000		
Logaritmo de verosimilitud	-890,276		
Criterio de información de Akaike (AIC)	1834,552		
AIC corregido para muestras finitas (AICC)	1840,390		
Criterio de información bayeasiana (BIC)	1933,358		

En la Tabla 3.17 se presentan los resultados del análisis para el modelo que muestra las interacciones de orden dos que afectan significativamente los números de ectoparásitos en geckos para la muestra sin los especímenes con mas de 100 ectoparásito.

En la Tabla 3.18 se presentan las estadísticas de bondad de ajuste del modelo anterior.

Tabla 3.17. Resultados del análisis del modelo lineal generalizado con distribución binomial negativa y enlace logaritmo aplicado a los números de ectoparásitos de T. delalandii usando el modelo con factores, covariables, interacciones de orden 2 y sexo anidado en población significativas para la muestra sin los especímenes con mas de 100 ectoparásitos. En negrita efectos significativos de factores e interacciones.

	Chi-cuadrado de Wald	df	p-valor
Intersección	627,519	1	0,000
Estación*IC	7,442	2	0.024
Estación*sexo(población)	34,874	11	0.000

Tabla 3.18. Resultados de la bondad de ajuste utilizando la distribución de Binomial Negativa con enlace logaritmo usando el modelo con factores, covariables, interacciones de orden 2 y sexo anidado en población significativas para la muestra sin los especímenes con mas de 100 ectoparásitos.

Bondad de ajuste			
	Valor	gl	Valor/gl
Desvianza	281,576	273	1,031
Desvianza escalada	232,299	273	
Chi-cuadrado de Pearson	330,911	273	1,212
Chi-cuadrado de Pearson escalado	273,000		
Logaritmo de verosimilitud	-897,707		
Criterio de información de Akaike (AIC)	1823,414		
AIC corregido para muestras finitas (AICC)	1824,958		
Criterio de información bayeasiana (BIC)	1874,646		

Los valores de AICc calculados para diferentes modelos de GLM (Tabla 3.19) informan que el valor más bajo fue para el modelo que incorpora solo los efectos significativos de las interacciones estación por IC y estación por sexo (anidado dentro de la población), para la muestra sin los especímenes con mas de 100 ectoparásito, por lo tanto, se elige el modelo 4 de esa tabla como el mejor modelo. Antes de obtener las conclusiones se debe comprobar que este modelo corrige la sobredispersión y se ajusta mejor que el modelo de Poisson sobredispersado.

Tabla 3.19. Valores de AIC para los diferentes modelos analizados con números de parásitos gecko como variables dependientes, covariables (SVL y temperatura de refugio), factores sexo (anidado dentro de la población), población y la estación. El modelo 4 fue seleccionado para interpretación.

Modelo	AICc	ΔAICc
1	1899,096	74,138
2	1935,444	110,486
3	1840,390	15,432
4	1824,958	0

A continuación se comprueba que el modelo 4 corrige la sobredispersión. Para ello se acude a los resultados de la Tabla 3.18.

El estadístico de desviación tiene un valor de $281,576$. Se evalua en la siguiente relación:

$$\frac{D}{gl} = \frac{281,576}{273} = 1,031,$$

por lo que el modelo no presenta sobredispersión.

Se comprueba ahora si se ajusta mejor que el modelo de Poisson sobredispersado. Para ello se utiliza una prueba de Razón de Verosimillitud. Se relaciona el mejor modelo con el modelo de Poisson con las mismas variables seleccionadas. La Var(Y) $= \mu$ para distribuciones de Poisson y Var(Y) $= \mu(1 + \alpha\mu)$ para modelo binomial negativo.

Se realiza el contraste de hipótesis:

$$\begin{cases} H_0 : \alpha = 0 \\ H_1 : \alpha > 0 \end{cases}$$

Así se comprueba la importancia de α. Si $\alpha = 0$ la distribución binomial negativa se convierte en la de Poisson.

El estadístico es el siguiente:

$$RV = -2(l(\hat{\mu}) - l(\hat{\mu}, \hat{\alpha})) \sim \chi^2_1$$

donde $l(\hat{\mu})$ y $l(\hat{\mu}, \hat{\alpha})$ son respectivamente, los logaritmos de verosimilitud bajo los modelos de regresión de Poisson y de binomial negativo y χ^2_1 es una distribución Chi-cuadrado con 1 grado de libertad.

Se rechaza H_0 si el estadístico es mayor que $\chi^2_{1,1-\alpha}$, con α nivel de significación.

Se acude a la Tabla 3.18 para observar el logaritmo de verosimilitud bajo el modelo binomial negativo cuyo valor es -897,707. En la Tabla 3.20 se observa el logaritmo de verosimilitud bajo el modelo de Poisson, cuyo valor es -1869,087.

Luego, $RV = -2(-1869,087 + 897,707) = 1942,76$.

Por lo que el modelo Binomial Negativo se ajusta mejor que el modelo de Poisson.

El modelo 4 es el mejor modelo, en él el número de ectoparásitos en T.delalandii se ve significativamente afectado por la interacción de la estación por sexo anidado en la población y la interacción de la estación por IC.

Tabla 3.20. Resultados de la bondad de ajuste utilizando la distribución de Poisson con enlace logaritmo usando las mismas interacciones que el modelo de la Tabla 3.18.

Bondad de ajuste			
	Valor	gl	Valor/gl
Desvianza	2769,262	273	10,144
Desvianza escalada	208,477	273	
Chi-cuadrado de Pearson	3626,344	273	13,283
Chi-cuadrado de Pearson escalado	273,000	273	
Logaritmo de verosimilitud	-1869,087		
Criterio de información de Akaike (AIC)	3766,174		
AIC corregido para muestras finitas (AICC)	3767,718		
Criterio de información bayeasiana (BIC)	3817,407		

3.7. Conclusiones del análisis de conteo del número de ectoparásitos

El estudio demuestra que el número de ectoparásitos Geckobia en T. delalandii no difirió significativamente entre las dos poblaciones analizadas.

El efecto significativo de estación por sexo (dentro de la población) refleja el hecho de que los ácaros fueron más frecuentes en otoño-invierno que en primavera-verano y que existen diferencias entre los tipos de individuos.

El efecto significativo en el número de ectoparásitos de la interacción de la estación y el índice de condición del gecko refleja que los ácaros fueron más frecuentes en individuos con un índice corporal más bajo y en otoño-invierno que en geckos con una condición corporal más alta en primavera-verano.

Conclusiones

En este trabajo se ha realizado un estudio teórico de los Modelos Lineales Generalizados y se han aplicado a un conjunto de datos reales. En el análisis de los datos en primer lugar se seleccionó el conjunto de modelos candidatos, posteriormente se analizaron haciendo uso de Modelos Lineales Generalizados con distribución de Poisson. Debido a los incovenientes, como la presencia de sobredispersión en los datos, se ha procedido a realizar el análisis haciendo uso de Modelos Lineales Generalizados con distribución binomial negativa. Por último se eligió el mejor modelo y se obtuvieron conclusiones.

En este trabajo de fin de grado se han conseguido los objetivos propuestos, estudiar los Modelos Lineales Generalizados y poder aplicarlos sobre un conjunto de datos reales. Se proponen, como línea de continuación de este trabajo las siguientes tareas, estudiar los Modelos Lineales Generalizados Mixtos, así como en mayor profundidad los Modelos Lineales Generalizados para datos de conteo. También se propone seguir estudiando acerca de la validación cruzada y comparar los resultados obtenidos con los criterios de AIC y BIC.

Bibliografía

[1] ANDERSON, D.R; BURNHAM, K.P. *Model selection and multimodel inference. A Practical Information-Theoretic Approach*, Springer-Verlag, Nueva York, 2002.

[2] ANDERSON, D.R; BURNHAM, K.P; HUYVAERT, K.P. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav Ecol Sociobiol*, 2011, vol. 65, pp. 23–35.

[3] COLIN CAMERON, A; TRIVEDI, P.K. Econometric Society Monographs: Regresion Analysis of Count Data, Cambridge University Press, Nueva York, 1998.

[4] DUNKLER, D; HEINZE, G; WALLISCH, C. Variable selection-A review and recommendations for the practicing statiscian. *Biometrical Journal*, 2018, vol. 65, pp. 431–449.

[5] DURBÁN, M. "Modelos Lineales Generalizados", Universidad Carlos III, Madrid, 2014. Disponible en: http://halweb.uc3m.es/esp/Personal/personas/durban/esp/web/GLM/curso_GLM.pdf

[6] HARDIN, J.W; HARRIS, T; YANG, Z. Modeling underdispersed count data with generalized Poisson regression. *The Stata Journal*, 2012, vol. 12, Nº 4, pp. 736–747.

[7] HUAN, L; REFAEILZADEH, P; TANG, L."k-fold Cross-Validation" Arizona State University, 2008.

[8] MC CULLAGH, P; NELDER, J.A. *Generalized Linear Models*, Chapman and Hall, Londres, 1989.

[9] MONTGOMERY, D.C. *Design and Analysis of Experiments*, John Wiley & Sons, Inc., 2013.

[10] MOLLA, D.T; MUNISWAMY, B. Power of Test Overdispersion Parameter in Negative Binomial Regresion Model. *IOSR Journal of Mathematics*, 2012, vol. 1, Nº 4, pp. 29–36.

[11] DE FUENTES FERNÁNDEZ, M; SUÁREZ RANCEL, M.M; DE QUINTANA GÓMEZ, P; MOLINA BORJA, M. *Season, body condi-*

tion and sex affect variation in ectoparasite number of Tarentola delalandii from two ecologically contrasting populations of Tenerife. 2020. Artículo en proceso de publicación.

Generalized linear models. Application to counting the number of ectoparasites.

Sección de Matemáticas
Universidad de La Laguna

Paula de Quintana Gómez
Facultad de Ciencias · Sección de Matemáticas
Universidad de La Laguna
alu01009903480@ull.edu.es

Abstract

In this end-of-grade project, Generalized Linear Models are studied. First, we will see the components, the parameter estimation and how to measure the goodness of fit. In addition, it will be seen how to select the best model from a set of models using critea dedicated to its comparison. The Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the cross Validation are studied. Finally, a study is carret out on the load of ectoparasites of the Geckobia mite in two ecologically oposed populations (north and south of Tenerife).

1. Generalized Linear Models

The objetive of linear regression models is to predict the behavior of a dependent variable based on independent variables. For this certain hypotheses must be fulfilled, for example, normalcy of the dependent variable, continuos variables, etc... The Generalized Linear Models are an extension of the classical linear models in which the response variable can follow any distribution of the exponencial family.
GLM's have three components:

- Random component: it is the response variable Y and its probability distribution.
- Systemic component: these are the explanatory variables $x_1, x_2, ..., x_p$ that are used in the linear predictor η,

$$\eta = \sum_{1}^{p} \beta_j x_j.$$

- Link function: it specifies the link between random and system components.

To estimate the parameters we will use the likelihood logarithm:

$$\boldsymbol{\beta_{new}} = \boldsymbol{\beta_{old}} - \left(E\left(\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \boldsymbol{\beta'}} \right) \right)^{-1}_{\boldsymbol{\beta_{old}}} \left. \frac{\partial l}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta_{old}}} =$$

$\boldsymbol{\beta_{old}} + (\boldsymbol{X'WX})^{-1}\boldsymbol{X'W}(\boldsymbol{y} - \boldsymbol{\mu_{old}})\boldsymbol{g'}(\boldsymbol{\mu}_{old}) = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'Wz}$
where $\boldsymbol{z} = \boldsymbol{X\beta_{old}} + (\boldsymbol{y} - \boldsymbol{\mu}_{old})\boldsymbol{g'}(\boldsymbol{\mu}_{old}) = \boldsymbol{\eta}_{old} + (\boldsymbol{y} - \boldsymbol{\mu}_{old})\boldsymbol{g'}(\boldsymbol{\mu}_{old})$
and $\boldsymbol{z}$ is the working vector.
To find out if the model is well adjusted, the Deviance statistic or the Pearson statistic are used:

$$\sum 2w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi = D(y;\hat{\mu})/\phi$$

$$X^2 = \sum (y - \hat{\mu})^2 / V(\hat{\mu}),$$

2. Model selection

To choose the set of candidate models, previous information, expert knowledge, limitation of model parameters and variables, etc, must influence.
Once the set of candidate models has been chosen, we must select the best one, for this, AIC, BIC and cross validation are used.

- Akaike information criterion (AIC):

$$AIC = 2K - 2\log(\mathrm{L})$$

- AICc

$$\mathrm{AICc} = \mathrm{AICc} + \frac{2\mathrm{K} \cdot (\mathrm{K}+1)}{\mathrm{n} - \mathrm{K} - 1}$$

- Bayesian information criterion (BIC):

$$BIC = -2 \cdot \ln(\mathrm{L}) + \mathrm{K}\ln(\mathrm{n}).$$

- Cross validation. It is a statistical method that evaluates and compares algorithms dividing data into two segments. One of the segments is used as trainning data and the other as data to validate the model. There are three types:
  Cross vaildation of k iterations..
  Aleatory cross validation.
  Cross validation leaving a data out.

3. Application of the generalized linear model to the count of the number of ectoparasites

Our objetive for this study is to qualitify the number of external parasites and analyze their variation between seasons and sexes in two isolated populations (north and south of the island) that have opposite ecological characteristics.
From different models were made, which after analysis were compared using the second order derivative (AICc).
The data was analyzed using generalized linear models with Poisson distribution and logarithmic link.
However, overdispersion was observed in the sample:

$$\frac{D}{gl} = \frac{3206,421}{263} = 12,192 > 1$$

To solve the presence of overdispersion, we resorted to the negative binomial distribution with a logarithmic link.
Parasitism was higher in individuals with lower body index values in autumn-winter than in those with greater body condition in spring-summer. Furthermore, the females of the northern population were more parasitized than males and juveniles in the colder month of the year.

References

[1] MC CULLAGH, P; NELDER, J.A. *Generalized Linear Models*, Chapman and Hall, Londres, 1989.

[2] ANDERSON, D.R; BURNHAM, K.P. *Model selection and multimodel inference. A Practical Information-Theoretic Approach*, Springer-Verlag, Nueva York, 2002.

[3] DURBÁN, M. "Modelos Lineales Generalizados", Universidad Carlos III, Madrid, 2014.