Sección de Matemáticas
Universidad de La Laguna

Ignacio Domínguez Espinosa

Modelos Lineales Generalizados: análisis, medidas de comparación y aplicación al conflicto entre reptiles

Generalized Linear Models: analysis, comparison metrics, and application to reptile conflicts

Trabajo Fin de Máster
Máster en Modelización
e Investigación Matemática
Estadística y Computación
La Laguna, Mayo de 2024

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

Quiero expresar mi agradecimiento a mi tutora, la profesora María Mercedes Suárez Rancel por su profundo interés y preocupación en mi aprendizaje y en la redacción de esta memoria. También agradezco a Miguel Molina Borja por sus consejos en la partes biológicas y prácticas de este documento.

Quiero agradecer sobretodo a mi familia y amigos por su apoyo durante todo mi periodo de posgrado, y por los buenos momentos vividos en el transcurso de esta etapa.

Ignacio Domínguez Espinosa
La Laguna, 18 de mayo de 2024

Resumen · Abstract

Resumen

En esta memoria se realiza una revisión de los Modelos Lineales Generalizados, tratando la teoría encontrada en la literatura. Estos engloban los modelos lineales clásicos, mediante la definición de tres componentes: una aleatoria, otra sistemática y una función de enlace para conectar las dos. Esto permite la modelización de una gran variedad de problemas, en los cuales los datos pueden seguir diversas distribuciones. Además, se muestra un algoritmo para la estimación de los parámetros del modelo.

Luego, se presentan diferentes herramientas para la selección del mejor conjunto de variables, como la Devianza o el AIC. También es común realizar infrerencia con diferentes tests de hipótesis e intervalos de confianza, y verificar las hipótesis del modelo mediante el cálculo de residuos o representaciones gráficas.

Por último, dada la versatilidad de aplicación de este tipo de modelos, y habiendo tenido la posibilidad de realizar una investigación matemática con el grupo de ecología de la Universidad de La Laguna, se aborda una de sus publicaciones como aplicación de los Modelos Lineales Generalizados. Esta consiste en el análisis de conflictos entre reptiles en función de su morfología, su comportamiento y la reflectancia de su piel. Para ello, se realizan dos experimentos: uno considerando todos estos aspectos y otro donde se aplica crema solar a los lagartos, para reducir la reflectancia. Los resultados revelan que, en el primer escenario, el resultado del conflicto depende de factores como la masa corporal y la frecuencia de ciertos comportamientos específicos, mientras que en el segundo caso, la aplicación de crema solar influye en el resultado de las confrontaciones entre los reptiles.

Palabras clave: *Modelos Lineales Generalizados – AIC – Devianza – Análisis de residuos – Conflictos entre reptiles*

Abstract

This dissertation conducts a review of Generalized Linear Models, addressing the theory found in the literature. These encompass classic linear models, by defining three components: a random one, a systematic one, and a link function to connect the two. This allows modeling a wide variety of problems, in which the data can follow various distributions. Additionally, an algorithm for estimating the model parameters is presented.

Subsequently, different tools are introduced for selecting the best set of variables, such as Deviance or AIC. It is also common to perform inference with different hypothesis tests and confidence intervals, and to verify the model hypotheses by calculating residuals or graphical representations.

Finally, given the versatility of application of this type of models, and having had the opportunity to conduct mathematical research with the ecology group of the University of La Laguna, one of their publications is addressed as an application of Generalized Linear Models. This consists of analyzing conflicts among reptiles based on their morphology, behavior, and skin reflectance. For this purpose, two experiments are conducted: one considering all these aspects and another where sunscreen is applied to the lizards to reduce reflectance. The results reveal that, in the first scenario, the outcome of the conflict depends on factors such as body mass and the frequency of certain specific behaviors, while in the second case, the application of sunscreen influences the outcome of confrontations among reptiles.

Keywords: *Generalized Linear Models – AIC – Deviance – Residual analysis – Lizard contests*

Contenido

Introducción

La modelización estadística siempre ha sido una herramienta básica y esencial para el análisis de datos. Desde modelos de predicción en el campo de la economía, hasta en el estudio del comportamiento animal. En el ámbito científico, los modelos lineales se han utilizado ampliamente, ya que emplean una combinación lineal para relacionar las diferentes variables, y así predecir el resultado de una variable respuesta. Esta última suele ser una variable aleatoria con una cierta distribución, como la distribución Normal.

Muchas técnicas fueron desarrolladas para el tratamiento de modelos lineales, comenzando con los trabajos de Gauss y Legendre [26] en estimaciones de mínimos cuadrados, y aplicándolas a modelos con variables predictoras normales y no normales (como la regresión logística con variables binarias o la regresión de Poisson con distribución de Poisson).

El concepto de Modelos Lineales Generalizados ("Generalized Linear Models" o GLM) fue introducido por John Nelder y Robert Wedderburn en 1972 [19], con un artículo que tenía como objetivo unificar los modelos lineales, aprovechando las similaridades de estructura y de estimación de parámetros por métodos de log-verosimilitud. Estas metodologías han sido recogidas en muchos libros, como [16] en 1983, permitiendo una visión más clara y extensa de los GLM.

En la actualidad, estos modelos siguen en constante desarrollo. Se han creado nuevas extensiones para poder modelizar una mayor cantidad de situaciones, como la predicción de datos de supervivencia [14] o la incorporación de inferencia Bayesiana [28]. También se han diseñado aplicaciones que permiten utilizar los modelos, como SPSS de IBM [13], y se han programado librerías en lenguajes de programación como Python [21] o R [8], dando mayor accesibilidad a los Modelos Lineales Generalizados.

El objetivo de este trabajo de fin de máster consiste en realizar una revisión de los Modelos Lineales Generalizados, tratando su estructura tripartita

clásica, formas de estimación, métricas de comparación de modelos, inferencia y métodos de diagnóstico de las hipótesis. Por último, se realiza un análisis similar al del artículo [3], donde se estudian de los conflictos que tienen los machos de dos especies de lagartos en la isla de Tenerife, en función de sus cualidades morfológicas, de comportamiento y la reflectancia de las manchas de la piel.

En el documento, se pueden ver aplicadas competencias del máster, como las adquiridas en la asignatura de Modelización Estadística. Entre las actividades llevadas a cabo en la asignatura, destaca una aproximación a los GLM en el caso logístico, que fue útil para comprender algunas de las ideas principales de los modelos. Otra competencia a destacar es "Saber realizar un análisis de regresión lineal (estimación, inferencia y diagnóstico)", en la que se inspira la estructura de este trabajo.

Esta memoria está organizada en los siguientes capítulos:

- **Capítulo 1.** Modelos Lineales Generalizados 1. Se presentan los conceptos básicos en este tipo de modelos, varios resultados teóricos y un algoritmo para estimar los parámetros.

- **Capítulo 2.** Métricas de comparación de Modelos e Inferencia 2. Se desarrollan diversas medidas que permiten encontrar las variables explicativas que mejor describen la variable dependiente. Se enuncian algunos resultados inferenciales para determinar la significación de las estimaciones.

- **Capítulo 3.** Diagnóstico de los Modelos Lineales Generalizados 3. Se definen diferentes métodos para verificar las hipótesis del modelo y para la detección de valores anómalos.

- **Capítulo 4.** Aplicación de Modelos Lineales Generalizados en el análisis de comportamiento de reptiles 4. Se adaptan los aspectos teóricos del modelo para realizar un estudio con datos reales. Estos provienen del artículo [3], y relacionados con los conflictos entre lagartos.

1

Modelos Lineales Generalizados

En este primer capítulo, se abordan los principios de los Modelos Lineales Generalizados. Se exploran resultados clásicos de los GLM, así como su estructura, y algunas propiedades. Por último, se describe un algoritmo para la estimación de parámetros del modelo.

1.1. Fundamentos y estructura de los GLM

Los Modelos Lineales Generalizados son una agrupación de múltiples modelos lineales clásicos en una única clase de modelos. Incluyen casos como: la regresión lineal, el análisis de la varianza, los modelos logit y probit, los modelos de respuesta multinomial, entre otros. Se agrupan debido a la cantidad de propiedades que tienen en común, como la linealidad o los métodos de estimación de parámetros. Por tanto, tiene sentido examinar esta diversidad de modelos dentro de una categoría única, la cual denominamos Modelos Lineales Generalizados.

Dado que es una extensión, algunas hipótesis de los modelos originales se conservan. Una de ellas es la necesidad de observaciones independientes o sin correlación, es decir, que el valor de cada uno de los datos no está influenciado de ninguna manera con el resto de los datos de la muestra, evitándose así situaciones donde la muestra presente alta autocorrelación o multicolinealidad. En la práctica, estas restricciones suelen ser bastante permisivas.

No obstante, en los Modelos Lineales Generalizados, no es necesario imponer otras condiciones más restrictivas como la normalidad. Este hecho amplía considerablemente la aplicabilidad de estos modelos en el ámbito científico, pudiendo estudiar conjuntos de datos con distribuciones no gaussianas.

Para desarrollar estos modelos, se parte del siguiente conjunto de datos: $(\mathbf{y}; \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$. El vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ es un conjunto de n observaciones independientes que provienen de un vector de variables aleatorias $\mathbf{Y}$ independientes y con medias $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$. Los vectores $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, donde el vector columna $\mathbf{x}_j' = (x_{1j}, x_{2j}, \ldots, x_{nj})$, contienen n mediciones de una

variable X_j, para cada observación y_i. Se puede reescribir este conjunto de vectores de forma matricial mediante la matriz $X \in M_{n\times p}(\mathbb{R})$, que se denomina *matriz modelo*.

A continuación, se definen las tres componentes del modelo, dando lugar al siguiente enfoque tripartito:

- *Componente aleatoria*: Representada por las variables $\mathbf{Y}$, independientes y con la misma distribución de la familia exponencial, con medias $E[\mathbf{Y}] = \boldsymbol{\mu}$ y varianza constante.
- *Componente sistemática*: Consiste en una combinación lineal de un vector de parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ y una matriz modelo X, que contiene las p variables con las que se quiere explicar el comportamiento de las esperanzas de $\mathbf{Y}$. Para ello, se crea un predictor lineal $\boldsymbol{\eta}$, definido de la siguiente manera:

$$\boldsymbol{\eta} = \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j$$

- *Función de enlace*: Es una función $g(\cdot)$ que relaciona las dos componentes mediante una igualdad. Tiene la siguiente expresión:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu})$$

En las siguientes subsecciones se trata en mayor profundidad cada una de las componentes, explicando sus características y mostrando algunos ejemplos.

1.1.1. La Componente aleatoria y la Familia Exponencial de distribuciones

Uno de los aspectos fundamentales de los Modelos Lineales Generalizados es la capacidad de manejar diferentes distribuciones de probabilidad de la variable dependiente. En los modelos clásicos, como la regresión lineal, surgía la limitación de exigir a la variable respuesta un único tipo de distribución. Si esta no verificaba dicho comportamiento, los científicos se veían obligados a recurrir a modelos más débiles o a probar con transformaciones de la variable, que pueden afectar a otras condiciones o hipótesis del modelo.

Con los Modelos Lineales Generalizados, se amplía el número de distribuciones de $\mathbf{Y}$ a aquellas provenientes de la familia de distribuciones exponencial.

Definición 1.1. *Sea Y una variable aleatoria continua o discreta con función de distribución F_Y, sea y una observación de la variable. Decimos que la distribución de Y pertenece a la familia exponencial de distribuciones de parámetro θ*

si la función de densidad $f_Y(y)$ (o distribución de probabilidad $p_Y(y)$), se puede expresar de la siguiente forma:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{1.1}$$

ó

$$p_Y(Y = y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{1.2}$$

Donde las funciones $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$ son conocidas, con $a(\cdot) \neq 0$

Comúnmente, a θ se le identifica como el *parámetro canónico* (o "canonical parameter"), mientras que a ϕ se le conoce como el *parámetro de dispersión* (o "dispersion parameter") en la literatura. En este trabajo, se asume que se conoce el valor del parámetro ϕ, pues si se desconoce, se aplican métodos para estimarlo (véase la sección 6.8. de [8]).

A continuación, se muestran algunos ejemplos de distribuciones conocidas que se pueden expresar de la anterior forma.

Ejemplo 1.2. Para la distribución Normal, si Y sigue una distribución $N(\mu,\sigma)$, se pueden tomar los siguientes parámetros y funciones para llegar a la función de densidad de la normal.

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

Para ello, sustituyendo en la función (1.1) se obtiene:

$$\begin{aligned} f_Y(y;\mu,\sigma) &= \exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]\right\} \\ &= \exp\left\{\frac{y\mu - \frac{\mu^2}{2} - \frac{y^2}{2}}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} \\ &= \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

Llegando a la expresión de la distribución normal.

Ejemplo 1.3. Para una distribución discreta como la Poisson, si Y sigue una distribución $Poiss(\lambda)$, los parámetros que hay que elegir son los siguientes:

- $\theta = \ln(\lambda)$,
- $a(\phi) = 1$
- $b(\theta) = e^{\theta}$
- $\phi = 1$
- $c(y, \phi) = -\ln(y!)$

Sustituyendo en la función (1.2), también se consigue llegar a la distribución de probabilidad de la variable.

Ejemplo 1.4. Otras distribuciones como la Bernoulli también pertenecen a la familia exponencial. Si Y sigue una distribución $Be(p)$, los valores a elegir son:

- $\theta = \ln[\frac{p}{1-p}]$,
- $a(\phi) = 1$
- $b(\theta) = \ln[1 + \exp\{\theta\}]$
- $\phi = 1$
- $c(y, \phi) = 0$

1.1.2. Construcción de la Componente sistemática mediante predictores lineales

Para conseguir que las variables X_j intervengan en el estudio del valor esperado de $\mathbf{Y}$, al igual que en los modelos lineales, se utiliza un predictor lineal de las variables. Este lo podemos expresar de la siguiente forma:

$$\boldsymbol{\eta} = \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j$$

Donde, se buscan estimar los parámetros β_i del modelo para garantizar una buena aproximación.

También hay que destacar que las variables X_j pueden ser cuantitativas o cualitativas. En el primer caso, se denominan covariables y en el segundo caso factores. Esta diferenciación es necesaria pues, al igual que en otros modelos lineales, hay que realizar cambios en las variables cualitativas para que los algoritmos de resolución funcionen correctamente.

Una variable cualitativa X_j que tome valores $\{v_1.v_2, ..., v_l\}$ se puede sustituir por $l - 1$ variables indicadoras B_{v_k}:

$$B_{v_k} = \begin{cases} 1, & \text{si } X_j = v_k \\ 0, & \text{si } X_j \neq v_k \end{cases}$$

donde $k \in \{1, 2, ..., l-1\}$. Nótese que son $l-1$ variables, pues si se incluyen todas las variables indicadoras para cada nivel, se genera un problema de multicolinealidad.

1.1.3. Función de enlace

La función de enlace (o "link function") de un GLM es una función $g(\cdot)$ que conecta la componente aleatoria con el predictor lineal. Para ello, se relacionan los valores esperados de las variables, denotado como $E[\mathbf{Y}] = \boldsymbol{\mu}$, con el predictor lineal definido en 1.1.2, de la siguiente forma:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu})$$

En general, la función de enlace $g(\cdot)$ debe ser monótona y diferenciable para permitir la estimación de los parámetros del modelo.

Exiten diversas funciones de enlace que pueden aplicarse en estos modelos. Cuando se elige la función que verifica que $\boldsymbol{\eta} = \boldsymbol{\theta}$, donde $\boldsymbol{\theta}$ representa los parámetros canónicos, llamamos a esa función la *función de enlace canónica*. En la siguiente tabla se muestran diferentes funciones de enlace canónicas para algunas distribuciones de la familia exponencial.

Distribución	Enlace Canónico
Normal	$g(\mu) = \mu$
Poisson	$g(\mu) = \ln(\mu)$
Binomial	$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$
Gamma	$g(\mu) = \mu^{-1}$
Gaussiana inversa	$g(\mu) = \mu^{-2}$

Tabla 1.1. Enlaces canónicos para algunas distribuciones

Se pueden utilizar otras funciones distintas de las canónicas, como:

1. El enlace probit:
$$\boldsymbol{\eta} = \Phi^{-1}(\boldsymbol{\mu})$$
Donde Φ representa la función de distribución acumulada de la normal.
2. El enlace complementario log-log:
$$\boldsymbol{\eta} = ln[ln(1 - \boldsymbol{\mu})]$$

3. La familia de enlaces exponencial:

$$\boldsymbol{\eta} = \begin{cases} \boldsymbol{\mu}^{\lambda}, & \text{si } \lambda \neq 0 \\ \ln(\boldsymbol{\mu}), & \text{si } \lambda = 0 \end{cases}$$

La selección de una función de enlace puede compararse a la elección de una transformación en la variable respuesta. Sin embargo, es importante destacar que la función de enlace afecta exclusivamente a la media de la variable respuesta, a diferencia de las transformaciones usuales que influyen directamente en la función de distribución. Al igual que con las transformaciones, optar por una función de enlace inadecuada puede generar problemas en la estimación de los parámetros del modelo.

1.2. Análisis y resultados teóricos

Esta sección tiene como objetivo mostrar varios resultados teóricos y definiciones relacionadas con los Modelos Lineales Generalizados. Se estudia la log-verosimilitud de los modelos, así como las funciones generadoras, y como calcular la media y la varianza de cada Y.

1.2.1. Log-Verosimilitud

La función de log-verosimilitud desempeña un papel fundamental en la teoría de los Modelos Lineales Generalizados, no solo como medida de concordancia entre los datos observados y las predicciones, sino también como herramienta para poder calcular estimadores de máxima verosimilitud.

Al suponer que la distribución de la variable respuesta pertenece a la familia exponencial, la expresión de esta función se simplifica drásticamente, facilitando mucho el cálculo y la interpretación de la expresión.

A continuación, se introduce una definición de la función de verosimilitud y de log-verosimilitud, que se utiliza en diferentes secciones del trabajo.

Definición 1.5. *Sea $y = \{y_1, y_2, ..., y_n\}$ una muestra de una población con densidad $f_Y(y;\theta)$, siendo θ los parámetros. La función de verosimilitud de la muestra asociada a θ es:*

$$L_\theta(y) = \prod_{i=1}^{n} f_Y(y_i;\theta)$$

Y la función de log-verosimilitud de la muestra asociada a θ es:

$$l_\theta(y) = ln(L_\theta(y))$$

Con esta definición, se puede determinar la log-verosimilitud de los Modelos Lineales Generalizados. Para ello, se sustituyen las expresiones (1.1) y (1.2) en la fórmula anterior.

- Para el caso de una variable continua, utilizamos la expresión (1.1):

$$\begin{aligned} L_{\theta,\phi}(y) &= \prod_{i=1}^{n} f_Y(y_i;\theta,\phi) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right\} \\ &= \exp\left\{\sum_{i=1}^{n}\left[\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right]\right\} \\ &= \exp\left\{\frac{\theta\sum_{i=1}^{n} y_i - nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\phi)\right\} \end{aligned}$$

  Por lo tanto,

$$L_{\theta,\phi}(y) = \exp\left\{\frac{\theta\sum_{i=1}^{n} y_i - nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\phi)\right\} \tag{1.3}$$

$$l_{\theta,\phi}(y) = \ln(L_{\theta,\phi}(y)) = \frac{\theta\sum_{i=1}^{n} y_i - nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\phi) \tag{1.4}$$

- Para el caso de variables discretas, seguimos un desarrollo similar, utilizando la función de masa de probabilidad (1.2), concluyendo con las mismas expresiones:

$$L_{\theta,\phi}(y) = \prod_{i=1}^{n} p_Y(Y = y_i;\theta,\phi) = \exp\left\{\frac{\theta\sum_{i=1}^{n} y_i - nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\phi)\right\} \tag{1.5}$$

$$l_{\theta,\phi}(y) = \ln(L_\theta(y)) = \frac{\theta\sum_{i=1}^{n} y_i - nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\phi) \tag{1.6}$$

En la práctica, tal y como indica el libro de P. McCullagh y J. A. Nelder, F.R.S. [16], estas expresiones se simplifican, pues la función $a(\phi)$ suele ser de la siguiente forma:

$$a(\phi) = \frac{\phi}{w}$$

donde w es un peso conocido a priori y ϕ es el parámetro de dispersión.

Ejemplo 1.6. En la distribución normal, $a(\phi) = \frac{\sigma^2}{n}$, donde $w = n$

1.2.2. Funciones generadoras

La familia de distribuciones exponencial tiene bastantes particularidades. Una de ellas es que la función generadora de momentos tiene una expresión bastante sencilla.

En este apartado se van a desarrollar las diferentes funciones generadoras aplicadas a los Modelos Lineales Generalizados.

Definición 1.7. *La función generadora de momentos de una variable Y , denotada por $M_Y(t)$, con función de densidad $f_Y(y)$ (o distribución de probabilidad $p_Y(y)$), se expresa de la siguiente forma:*

$$M(t) = E[e^{tY}] = \begin{cases} \int_S f_Y(y)e^{ty}dy, & si\ Y\ es\ continua \\ \sum_{y\in S} p_Y(y)e^{ty}, & si\ Y\ es\ discreta \end{cases}$$

para todos los valores de t donde la esperanza exista y siendo S el conjunto de todos los valores que puede tomar Y.

Definición 1.8. *La función generadora acumulada tiene la expresión siguiente:*

$$K(t) = \log(M(t)) = \log(E[e^{tY}])$$

Definición 1.9. *El momento r-ésimo se calcula de la siguiente forma:*

$$\kappa_r = \left.\frac{d^r K(t)}{dt^r}\right|_{t=0}$$

En la fórmula de los momentos, la esperanza y la varianza de Y son los momentos κ_1 y κ_2 respectivamente. En la sección 1.2.3 se calculan.

A continuación, se va a desarrollar la fórmula de la función generadora de momentos para el caso donde la variable Y es continua. Se puede realizar el mismo procedimiento para cuando la variable es discreta y se llega a la misma expresión.

Sabiendo que Y es continua y tiene una distribución de la familia exponencial, se pueden calcular las funciones generadoras de la siguiente forma:

$$\begin{aligned} M(t) = E[\exp(tY)] &= \int_S \exp(ty)\exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}dy \\ &= \exp\left\{\frac{-b(\theta)}{a(\phi)}\right\}\int_S \exp\left\{ty + c(y,\phi) + \frac{y\theta}{a(\phi)}\right\}dy \end{aligned}$$

Multiplicando y dividiendo por $\exp\left\{\frac{b(\theta+ta\phi)}{a(\phi)}\right\}$ se obtiene:

$$
\begin{aligned}
M(t) &= \exp\left\{\frac{b(\theta+ta\phi)-b(\theta)}{a(\phi)}\right\}\int_S \exp\left\{ty+c(y,\phi)+\frac{y\theta}{a(\phi)}-\frac{b(\theta+ta\phi)}{a(\phi)}\right\}dy \\
&= \exp\left\{\frac{b(\theta+ta\phi)-b(\theta)}{a(\phi)}\right\}\int_S \exp\left\{\frac{y(\theta+ta\phi)-b(\theta+ta\phi)}{a(\phi)}+c(y,\phi)\right\}dy \\
&= \exp\left\{\frac{b(\theta')-b(\theta)}{a(\phi)}\right\}\int_S \exp\left\{\frac{y(\theta')-b(\theta')}{a(\phi)}+c(y,\phi)\right\}
\end{aligned}
$$

Donde $\theta' = \theta + t\phi$. La integral de la derecha de la expresión es igual a 1 pues el integrando es la función de densidad con θ'. Por tanto, la fórmula final de la función generadora de momentos es:

$$M(t) = \exp\left\{\frac{b(\theta+ta(\phi))-b(\theta)}{a(\phi)}\right\} \tag{1.7}$$

La expresión de la función generadora acumulada es:

$$K(t) = \frac{b(\theta+ta(\phi))-b(\theta)}{a(\phi)} \tag{1.8}$$

Ejemplo 1.10. Cuando Y sigue una distribución normal $N(\mu,\sigma)$, obtenemos la siguiente función generadora de momentos:

$$M(t) = \exp\left\{\frac{(\mu+t\sigma^2)^2-\mu^2}{2\sigma^2}\right\} = \mu t + \frac{(\sigma t)^2}{2}$$

Ejemplo 1.11. Para la Poisson, si Y sigue una distribución $Poiss(\lambda)$, la función queda:

$$M(t) = \exp\left\{e^{ln(\lambda)+t} - e^{ln(\lambda)}\right\} = e^{\lambda(e^t-1)}$$

Ejemplo 1.12. Si es una Bernoulli, con Y siguiendo una distribución $Be(p)$, la función es:

$$M(t) = 1 - p + pe^t$$

Por último, se describe una propiedad única de las variables con distribución de la familia exponencial, que sirve para el cálculo de momentos.

Proposición 1.13. *Si Y proviene de una distribución con familia exponencial, el momento r-ésimo se puede expresar de la siguiente forma:*

$$\kappa_r = a(\phi)^{r-1}\frac{d^r b(\theta)}{d\theta^r}$$

Demostración. Sabemos por la definición 1.9 que la fórmula del momento r-ésimo viene determinada por:

$$\kappa_r = \left.\frac{d^r K(t)}{dt^r}\right|_{t=0}$$

Primero, se demuestra por inducción que:

$$\frac{d^r K(t)}{dt^r} = a(\phi)^{r-1}\frac{d^r b(\theta + ta(\phi))}{d\theta'^r}$$

Siendo $\theta' = \theta + ta(\phi)$.

Para $r = 1$, utilizando la expresión (1.8) y aplicando la regla de la cadena se obtiene:

$$\frac{dK(t)}{dt} = \frac{1}{a(\phi)}b'(\theta + ta(\phi))a(\phi) = b'(\theta + ta(\phi)) = a(\phi)^0\frac{db(\theta + ta(\phi))}{d\theta'}$$

Tomamos como cierta la hipótesis de inducción $\frac{d^r K(t)}{dt^r} = a(\phi)^{r-1}\frac{d^r b(\theta+ta(\phi))}{d\theta'^r}$, y se demuestra para $r + 1$.

$$\frac{d^{r+1}K(t)}{dt^{r+1}} = \frac{d}{dt}\left(\frac{d^r K(t)}{dt^r}\right)$$

Por hipótesis inductiva:

$$\frac{d^{r+1}K(t)}{dt^{r+1}} = \frac{d}{dt}\left(a(\phi)^{r-1}\frac{d^r b(\theta + ta(\phi))}{d\theta'^r}\right) = a(\phi)^r\frac{d^{r+1}b(\theta + ta(\phi))}{d\theta'^{r+1}}$$

Luego se demuestra que:

$$\frac{d^r K(t)}{dt^r} = a(\phi)^{r-1}\frac{d^r b(\theta + ta(\phi))}{d\theta'^r}$$

Siendo $\theta' = \theta + ta(\phi)$.

Tomando $t = 0$, $\theta' = \theta$ y:

$$\kappa_r = \left.\frac{d^r K(t)}{dt^r}\right|_{t=0} = a(\phi)^{r-1}\frac{d^r b(\theta)}{d\theta^r}$$

□

1.2.3. Media y varianza de variables de la familia exponencial

Dado que Y proviene de una distribución con familia exponencial, y sabiendo que la media y la varianza son el primer y el segundo momento, se pueden calcular con la proposición anterior, llegando a lo siguiente:

$$\kappa_1 = b'(\theta)$$

$$\kappa_2 = b''(\theta)a(\phi)$$

Por tanto, $E[Y] = \mu = b'(\theta)$ y $var[Y] = b''(\theta)a(\phi)$.

Nótese que en el cálculo de la varianza, sabiendo que la media se puede expresar como $\mu = \frac{db(\theta)}{d\theta}$, se tiene que:

$$\frac{d^2b(\theta)}{d\theta^2} = \frac{d}{d\theta}\left(\frac{db(\theta)}{d\theta}\right) = \frac{d\mu}{d\theta}$$

Puesto que la varianza es positiva, es decir, $\frac{d^2b(\theta)}{d\theta^2} > 0$, entonces $\frac{d\mu}{d\theta} > 0$. Se concluye que μ debe ser una función monótona creciente de θ. Este resultado es bastante importante, pues la relación entre las variables permite calcular los estimadores de máxima verosimilitud mediante el uso de derivadas parciales (ver la sección 1.3.1). Además, se puede definir una función denominada función de varianza mediante la siguiente expresión:

$$V(\mu) = \frac{d\mu}{d\theta}$$

Luego, la varianza de Y se puede escribir como $var[Y] = a(\phi)V(\mu)$.

Ejemplo 1.14. Para cuando Y sigue una distribución normal $N(\mu, \sigma)$, sabemos que $b(\theta) = \frac{\mu^2}{2}$, con $\theta = \mu$. La esperanza de Y sería $E[Y] = \frac{d(\frac{\mu^2}{2})}{d\mu} = \mu$, la función de varianza $V(\mu) = \frac{d^2b(\theta)}{d\theta^2} = 1$ y la varianza de Y es $var[Y] = a(\phi)V(\mu) = \sigma^2$, que son los valores usuales de la esperanza y la varianza de la distribución gaussiana.

1.3. Estimación de parámetros del modelo

En este apartado se intentan estimar los parámetros de los Modelos Lineales Generalizados mediante los estimadores clásicos de máxima verosimilitud. Para ello, primero se introduce una sección donde se indican algunas fórmulas que influyen en el desarrollo del algoritmo y luego se muestra el procedimiento iterativo para alcanzar las estimaciones.

1.3.1. Funciones de puntuación e información de los parámetros

El algoritmo de estimación de β_j se basa en la definición de dos funciones: una función de puntuación (o "scoring"), que es la derivada de la log-verosimilitud respecto de β_j, y una función de información de Fisher, basada en el valor esperado de la matriz hessiana de la log-verosimilitud respecto de β_j y β_k.

Para llegar a sus expresiones, se parte de Y, cuya distribución pertenece a la familia exponencial, y sea y una observación de Y. La función de log-verosimilitud, tomando $n = 1$ en (1.4), es la siguiente:

$$ln(L_{\theta,\phi}(y)) = \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)$$

Sustituyendo, como en el final de la sección 1.2.1, $a(\phi) = \frac{\phi}{w}$, siendo w el peso de la observación y. Queda la siguiente función:

$$ln(L_{\theta,\phi}(y)) = w\frac{\theta y - b(\theta)}{\phi} + c(y, \phi)$$

La derivada de la log-verosimilitud en función de θ es la siguiente:

$$\frac{\partial ln(L_{\theta,\phi}(y))}{\partial \theta} = w\frac{y - b'(\theta)}{\phi} = w\frac{y - \mu}{\phi}$$

Sustituyendo $\mu = db(\theta)/d\theta$ y $V(\mu) = \frac{d\mu}{d\theta}$ se tiene que:

$$\frac{\partial ln(L_{\theta,\phi}(y))}{\partial \mu} = \frac{\partial ln(L_{\theta,\phi}(y))}{\partial \theta}\frac{\partial \theta}{\partial \mu} = w\frac{y - \mu}{\phi V(\mu)}$$

Luego, si tenemos el predictor lineal

$$g(\mu) = \eta = \sum_{j=0}^{p} \beta_j x_j$$

donde $x_0 = 1$, la derivada de este con respecto de β_j es:

$$\frac{\partial \eta}{\partial \beta_j} = x_i$$

$$\frac{\partial \eta}{\partial \mu}\frac{\partial \mu}{\partial \beta_j} = x_i$$

Por tanto, la derivada de $lnL_{\theta,\phi}(y))$ respecto de β_j es:

$$\frac{\partial ln(L_{\theta,\phi}(y))}{\partial \beta_j} = \frac{\partial ln(L_{\theta,\phi}(y))}{\partial \mu}\frac{\partial \mu}{\partial \beta_j} = w\frac{(y - \mu)x_j}{\phi V(\mu)\frac{\partial \eta}{\partial \mu}}$$

Para calcular la información de Fisher de β_j, se necesitan las esperanzas de las segundas derivadas:

$$\frac{\partial^2 ln(L_{\theta,\phi}(y))}{\partial\beta_k\partial\beta_j} = \frac{\partial}{\partial\beta_k} w\frac{(y-\mu)x_j}{\phi V(\mu)\frac{\partial\eta}{\partial\mu}} + (y-\mu)\frac{\partial}{\partial\beta_k}\left(\frac{wx_j}{\phi V(\mu)\frac{\partial\eta}{\partial\mu}}\right)$$

El término de la derecha se hace 0 al hacer la esperanza de $(y-\mu)$. La expresión que queda es la siguiente:

$$E\left[\frac{\partial^2 ln(L_{\theta,\phi}(y))}{\partial\beta_k\partial\beta_j}\right] = \frac{-wx_jx_k}{\phi V(\mu)(\frac{\partial\eta}{\partial\mu})^2}$$

Con estas identidades ya se puede tratar el caso con $\mathbf{y} = (y_1, y_2, ..., y_n)$ observaciones. Nótese que se busca predecir $p+1$ parámetros utilizando los estimadores de máxima verosimilitud. Para ello, se deriva la función de log-verosimilitud, respecto de los parámetros $\beta_j, \forall j \in \{0, 1, ..., p\}$ y se iguala a 0. Con ello se define la función de puntuación para β_j:

$$U(\beta_j) = \frac{\partial l(\beta_0, \beta_1, ..., \beta_p, \phi; \mathbf{y})}{\partial\beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} W_i\frac{\partial\eta_i}{\partial\mu_i}(y_i-\mu_i)x_{ij}, \forall j \in \{0, 1, ..., p\}$$

donde

$$W_i = \frac{w_i}{V(\mu_i)(\frac{\partial\eta_i}{\partial\mu_i})^2}$$

y se busca β_j tal que $U(\beta_j) = 0$, $\forall j \in \{0, 1, ..., p\}$

El criterio de información de Fisher se expresa mediante la suma de las esperanzas calculadas previamente. Esta da la siguiente fórmula:

$$I_{jk}(\beta_i) = -\frac{1}{\phi}\sum_{i=1}^{n} W_i x_{ij}x_{ik}, \forall i, j, k \in \{0, 1, ..., p\}$$

1.3.2. Algoritmo de estimación de los parámetros

Uno de los métodos para la estimación de los parámetros β_j es el algoritmo de puntuaciones de Fisher ("Fisher scoring algoritm"). Este proporciona un método de cálculo de los estimadores de máxima verosimilitud de los parámetros β_j, que denotamos por $\hat{\beta}_j$.

Cada iteración del algoritmo debe encontrar soluciones para las ecuaciones $U(\beta_j) = 0, \forall j \in \{0, 1, ..., p\}$, utilizando la información de Fisher $I_{jk}(\beta)$ para el la mejora de soluciones. Una alternativa que ofrece el libro [16] es utilizar un

proceso equivalente a resolver las ecuaciones. Para ello, se construye una variable z, llamada *variable dependiente ajustada*, con la cual se resuelve un problema de regresión lineal formada por z como variable dependiente y con covariables x_j y pesos W. En general, la expresión de z es de la siguiente forma:

$$z = \eta + \left(\frac{\partial \eta}{\partial \mu}\right)(y - \mu) \tag{1.9}$$

Luego, el procedimiento para la estimación de los coeficientes β_j es el siguiente:

1. Definimos $\hat{\eta_{(0)}}$ como la estimación actual del predictor tomando valores aleatorios de $\hat{\beta}_{(0)}$, y $\hat{\mu_{(0)}}$ como el valor obtenido de la función de enlace $\eta = g(\mu)$.
2. Se calcula z_0:
$$z_0 = \hat{\eta_{(0)}} + \left(\frac{\partial \eta}{\partial \mu}\right)_{(0)} (y - \hat{\mu_{(0)}})$$
donde la derivada $(\frac{\partial \eta}{\partial \mu})$ está evaluada en $\hat{\mu_{(0)}}$.
3. Se calculan los pesos $W_{(0)}^{-1}$:
$$W_{(0)}^{-1} = \frac{V(\hat{\mu_{(0)}})}{w} \left(\frac{\eta_{(0)}}{\mu_{(0)}}\right)^2_{(0)}$$
4. Se hace regresión en z_0, con las covariables x_j y con los pesos $W_{(0)}$, resultando en la estimación $\hat{\beta}_{(1)}$.
5. Con $\hat{\beta}_{(1)}$ se itera desde el principio del procedimiento hasta convergencia o hasta un número determinado de iteraciones.

Ejemplo 1.15. Este ejemplo proviene del libro [8], donde trata con un conjunto de datos llamado "nminer". Se utiliza un GLM con distribución de Poisson y con componente sistemática $log(\mu) = \beta_0 + \beta_1 x$. Utilizando la fórmula (1.9), tenemos que el regresor z es:

$$z = log(\hat{\mu}) + \frac{y - \hat{\mu}}{\hat{\mu}}$$

Se toma como $\hat{\mu_0} = y + 0.1$ para evitar los ceros de los datos en el modelo logarítmico, y como pesos se utiliza $W = \mu$.

En la tabla 1.2. se puede ver la evolución de los parámetros en cada iteración, donde se observa que el algoritmo converge rápidamente, dando como modelo final:

$$\log \hat{\mu} = -0.8762 + 0.1140x$$

Iteración r	$\hat{\beta}_0^{(r)}$	$\hat{\beta}_1^{(r)}$
1	0.122336	0.081071
2	−0.589798	0.103745
3	−0.851982	0.113123
4	−0.876031	0.113975
5	−0.876211	0.113981
6	−0.876211	0.113981

Tabla 1.2. Tabla de los coeficientes obtenidos en cada iteración del algoritmo

2

Métricas de comparación de modelos e inferencia

En este capítulo se aborda la tarea de encontrar el mejor subconjunto de variables del GLM, mediante la creación de métricas. Además, se define inferencia en los modelos para estudiar la significación de las estimaciones.

2.1. Selección del mejor modelo

Tras la definición de los Modelos Lineales Generalizados, queda determinar qué combinación de variables son las que mejor explican los datos. Para ello se busca crear un conjunto de modelos, donde, para cada modelo, se toma un subconjunto diferente de variables explicativas.

Para desarrollar esta colección de modelos, se inicia tomando uno con todas las variables ("*full model*" o modelo completo) y otro sin ninguna variable, solo con el término β_0 ("*null model*" o modelo nulo), para tenerlos como referencia. A partir de estos, se toman distintos conjuntos de variables como posibles alternativas que puedan predecir mejor los datos.

En general, se busca una colección pequeña de variables que permitan que el modelo tenga buena capacidad de predicción. Esta idea se denomina el *principio de parsimonia*, es decir, el mejor modelo tendrá el menor número posible de variables, que garantice una representación adecuada de los datos.

Esto permite evitar los efectos y consecuencias de dos sucesos muy comunes en la construcción de modelos: *el sobreajuste* (o *"overfitting"*) y el *el subajuste* (o *"underfitting"*).

- El sobreajuste surge cuando se incluyen variables no explicativas dentro del modelo de predicción. Esto, en modelos como la regresión lineal, da lugar a estimadores insesgados, pero con alta varianza. . Para resolver el problema, se recomienda aplicar técnicas de reducción de dimensionalidad u otras medidas, como las del artículo [5].

- El subajuste aparece cuando se eliminan variables significativas dentro del modelo. En la regresión lineal, esto provoca que los estimadores tengan poca varianza, pero presenten sesgo, dando lugar a errores en los coeficientes estimados. Una solución sencilla puede ser detectar y añadir las variables que hayan sido omitidas en el estudio, y que puedan tener importancia en el modelo.

2.2. Métricas utilizadas en los Modelos Lineales Generalizados

Una forma sencilla de evaluar las capacidades del modelo es mediante el diseño de métricas o medidas con las que poder comparar los modelos y así poder determinar el modelo más adecuado a las necesidades del estudio.

Existen algunas bastante utilizadas en los GLM, que se centran en el estudio de los errores de predicción, de los residuos o de la verosimilitud, entre otros. Por otra parte, hay otras medidas inspiradas en la teoría de la información, en distancias u otras métricas, que pueden ser útiles en la comparación de modelos.

En las siguientes subsecciones se van a tratar estos tipos de medidas para los Modelos Lineales Generalizados, su origen y su interpretación.

2.2.1. Métricas basadas en el error y la verosimilitud

Una medida comúnmente utilizada en los modelos que utilizan los estimadores de máxima log-verosimilitud es el propio valor máximo de la log-verosimilitud, es decir, el valor de la función evaluada en los estimadores calculados con el método de 1.3.2.

Utilizado la fórmula de la definición 1.5, se puede expresar la medida de máxima log-verosimilitud como:

$$\text{Máx. Log-Verosimilitud} = \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \mathbf{y}) = \ell(\hat{\boldsymbol{\beta}}; \mathbf{y}) \tag{2.1}$$

siendo $\hat{\boldsymbol{\beta}}$ los estimadores calculados con 1.3.2.

Otra de las métricas más utilizadas en el contexto del los GLM para cuantificar la discrepancia del modelo es la Devianza (o "Deviance"). Esta se basa en la comparación del modelo actual con un modelo denominado *modelo saturado*. Este último, tiene tantos parámetros como observaciones tiene la muestra, donde cada parámetro corresponde a un dato y las estimaciones $\hat{\boldsymbol{\mu}}$ obtenidas equivalen exactamente con las correspondientes μ de cada dato. Por ejemplo, en el caso de regresión lineal, calcular el modelo saturado es equivalente a hacer interpolación de los datos y obtener como modelo el polinomio interpolador.

Para comparar ambos modelos, se utilizan las medidas de máxima log-verosimilitud de cada modelo, descritas en la expresión (2.1), denotando la verosimilitud del modelo saturado como $\ell_S(\mathbf{y})$. La fórmula de la Devianza es la siguiente:

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2[\ell_S(\mathbf{y}) - \ell(\hat{\boldsymbol{\beta}};\mathbf{y})]\phi \tag{2.2}$$

Donde ϕ es el parámetro de dispersión del modelo.

De esta medida, se suele emplear sobretodo su versión escalada. Esta se calcula dividiendo la Devianza entre el parámetro ϕ, obteniendo una nueva métrica denominada Devianza escalada, que se denota como $D^*(\mathbf{y};\hat{\boldsymbol{\mu}})$. Esta tiene mayor interés en el estudio comparativo de modelos pues se puede utilizar en inferencia, al distribuirse como una χ^2 de p grados de libertad, siendo p el número de parámetros del modelo.

Luego, se define la expresión de la Devianza escalada de la forma siguiente:

$$D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{\phi} = 2[\ell_S(\mathbf{y}) - \ell(\hat{\boldsymbol{\beta}};\mathbf{y})] \tag{2.3}$$

Otra medida menos utilizada es el estadístico de bondad de ajuste χ^2 de Pearson. Se suele calcular cuando se requiere de una medida con una interpretación más directa. Esta medida de bondad de ajuste se puede definir de la siguiente forma:

$$\chi^2 = \sum \frac{(\mathbf{y} - \hat{\mu})^2}{var(\hat{\mu})} \tag{2.4}$$

Con $var(\hat{\mu})$ la varianza estimada de la función de distribución.

2.2.2. Criterio de Información de Akaike y otras métricas basadas en la teoría de la información

Una alternativa para seleccionar modelos adecuados es a través de fórmulas inspiradas en algunas propiedades de la teoría de la información. Dentro de este grupo, destaca el Criterio de Información de Akaike (AIC), que es una de las medidas más reconocidas en los Modelos Lineales Generalizados.

Este criterio se basa en la suposición de que existe una distribución f que describe la realidad, es decir, la combinación todos los procesos naturales, biológicos o físicos que ocurrieron para originar los datos, junto con el muestreo, la toma y medición de datos, y el almacenamiento para el posterior análisis y tratamiento. También se considera una función g que representa el modelo aproximado que se quiere comparar con f. Este se modeliza y se construye en

función del conjunto de datos de estudio x y los parámetros del modelo g, que se denotan como θ.

El Criterio de Información de Akaike emplea una distancia llamada *distancia de Kullback-Liebler* para representar la pérdida de información cuando g es usada para aproximar f o la distancia que hay entre la realidad f con la aproximación g.

Definición 2.1. *Se define la distancia Kullback-Liebler entre f y g como:*

- *En el caso continuo:*

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x|\theta)}\right) dx \tag{2.5}$$

  siendo $f(x)$ la distribución real y $g(x|\theta)$ el modelo aproximado por g para los datos x y dados los parámetros θ.
- *En el caso discreto:*

$$I(f,g) = \sum_{i=1}^{K} p_i \log\left(\frac{p_i}{\pi_i}\right) \tag{2.6}$$

  donde K es el número de casos posibles, p_i es la probabilidad real (representa f) y π_i es la aproximada (representa g), y además se cumple que:

$$0 < p_i < 1, 0 < \pi_i < 1, \sum_{i=1}^{K} p_i = \sum_{i=1}^{K} \pi_i = 1$$

Ejemplo 2.2. Supongamos que la función f sigue una distribución gamma de parámetros $\Gamma(\alpha = 4, \beta = 4$. Se busca ver cual de las distribuciones de la tabla 2.1 es la que menos información pierde. Para ello, se calculan las distancias $I(f,g)$ según la fórmula (2.5) o (2.6), y se elige aquella con menor pérdida de información o aquella más cercana a la realidad. En esta ocasión se prefiere g_1, que es la distribución de Weibull, con $\alpha = 2$ y $\beta = 20$).

Funciones	Modelos aproximados	$I(f,g_i)$	Clasificación
g_1	Weibull ($\alpha = 2$, $\beta = 20$)	0.04620	1
g_2	Lognormal ($\theta = 2$, $\sigma^2 = 2$)	0.67235	3
g_3	Gaussiana inversa ($\alpha = 16$, $\beta = 64$)	0.06008	2
g_4	Fisher-Snedecor ($\alpha = 4$, $\beta = 10$)	5.74555	4

Tabla 2.1. Caption

Cabe destacar que, en el ejemplo anterior, para poder definir correctamente la distancia Kullback-Liebler, se necesita saber la distribución real de f, así como conocer el valor exacto de los parámetros de los modelos g_i que se quieren

comparar. Esto es una fuerte desventaja, pues en general, se desconoce como se distribuye f y se utilizan estimaciones para los parámetros de las g_i.

Para resolver el primer problema, se puede escribir la fórmula de (2.5) de la siguiente forma:

$$\begin{aligned} I(f,g) &= \int f(x)\log(f(x))dx - \int f(x)\log(g(x|\theta))dx \\ &= E_f[\log(f(x))] - E_f[\log(g(x|\theta))] \end{aligned} \tag{2.7}$$

Se puede observar que la esperanza de la izquierda es constante para cualquier función g que se quiera comparar. Luego, se puede definir una distancia relativa a la constante, tomando únicamente la segunda esperanza.

$$I(f,g) - E_f[\log(f(x))] = -E_f[\log(g(x|\theta))]$$

Esta puede ser una buena medida para cuantificar la distancia relativa entre los modelos, siempre que se pueda estimar el valor de $E_f[\log(g(x|\theta))]$.

Cabe destacar que, tanto el desarrollo anterior, como los posteriores, se pueden llevar a cabo de forma similar con las fórmulas para el caso discreto (2.6).

Por otra parte, hay que tener en cuenta que los parámetros θ de la función tomada g se suelen estimar en base a la muestra tomada. Esto afecta considerablemente al valor de la distancia pues se tendría la siguiente fórmula:

$$\hat{I}(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x|\hat{\theta(y)})}\right)dx$$

donde x es la variable de integración e y representa los datos, haciendo que $\hat{\theta}$ dependa de la muestra tomada.

Para que la información no dependa tanto de la muestra, es más conveniente tomar el valor esperado de la distancia estimada, es decir:

$$E_{\hat{\theta}}[\hat{I}(f,g)] = \int f(y)\left[\int f(x)\log\left(\frac{f(x)}{g(x|\hat{\theta}(y))}\right)dx\right]dy \tag{2.8}$$

Haciendo un desarrollo similar al utilizado en (2.7):

$$
\begin{aligned}
E_{\hat{\theta}}[\hat{I}(f,g)] &= \int f(y) \left[\int f(x) \log\left(f(x)\right) dx \right] dy \\
&\quad - \int f(y) \left[\int f(x) \log\left(g(x|\hat{\theta}(y)) \right) dx \right] dy \\
&= \left[\int f(y) dy \right] \left[\int f(x) \log(f(x)) dx \right] \\
&\quad - \int f(y) \left[\int f(x) \log(g(x|\hat{\theta}(y))) dx \right] dy \\
&= \text{Constante} - E_y E_x \left[\log(g(x|\hat{\theta}(y))) \right] \qquad (2.9)
\end{aligned}
$$

Por tanto, se puede calcular la distancia relativa esperada de Kullback-Liebler estimando:

$$T = E_y E_x \left[\log(g(x|\hat{\theta}(y))) \right] \qquad (2.10)$$

El valor de esta esperanza dependen exclusivamente de g y θ estimado por y y g (normalmente con estimadores de máxima verosimilitud). Si se tiene un conjunto de modelos $\{g_i\}$ se consideran mejores aquellos modelos con mayor valor en la esperanza.

Akaike en 1973 demostró que un estimador sesgado de T es la máxima log-verosimilitud de la fórmula (2.1) y que el sesgo, bajo ciertas condiciones, es aproximadamente el número de parámetros p del modelo.

$$\hat{T} = \ell(\hat{\theta}; y) - p$$

Basándose en los criterios clásicos de información, se multiplica por -2 la estimación de T, se obtiene el Criterio de Información de Akaike:

$$AIC = -2\ell(\hat{\theta}; y) + 2p$$

Donde se busca que el valor de AIC sea lo más pequeño posible.

En el contexto de los Modelos Lineales Gerenalizados el parámetro θ descrito son los β_j y ϕ. La fórmula final queda de la siguiente forma:

$$AIC = -2\ell(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y}) + 2p \qquad (2.11)$$

Observando la fórmula, se puede comprobar que cumple el principio de parsimonia, es decir, intenta buscar un modelo que estime correctamente con $\ell(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y})$, mientras que restringe el número de parámetros del modelo con p, haciendo que sea lo más simple que se pueda sin perder demasiada precisión.

Como AIC es una escala relativa, se suelen computar las diferencias Δ_i de AIC, para todos los modelos $\{g_i\}$ que se estén comparando.

$$\Delta_i = AIC_i - \min_i AIC_i \tag{2.12}$$

Según la literatura relacionada con este tipo de métricas, los modelos con $\Delta_i < 2$ tienen bastante probabilidad de ser correctos y deben de ser considerados en la elección del mejor modelo. Otros con $4 < \Delta_i < 7$ se pueden considerar, pero son menos plausibles. Aquellos cuyas diferencias sean mayores de 10 no se suelen considerar, salvo que provengan de opinión de expertos en la materia del estudio u otros motivos ajenos al modelo matemático.

Una dificultad que tiene el AIC en la comparación de modelos, es que se presente la situación en la que hayan muchos parámetros y poco tamaño muestral. En este caso, se utiliza una medida alternativa llamada *Criterio de información de Akaike de segundo orden*, que se denota por AIC_c y se define como:

$$AIC_c = -2\ell(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y}) + 2p\left(\frac{n}{n-p-1}\right) \tag{2.13}$$

donde n es el tamaño muestral.

Esta medida añade una penalización al número de variables, mediante el término corrector $\frac{n}{n-p-1}$. Si n es suficientemente grande y p no, $\left(\frac{n}{n-p-1}\right) \longrightarrow 1$. Se recomienda utilizar AIC_c cuando $\frac{n}{p} < 40$, y en otro caso utilizar AIC.

Existen otras medidas derivadas de AIC o basadas en otros criterios. Algunas de estas son:

- *Criterio de Información Bayesiano* (o "Bayesian Information Criteria", BIC) se basan en medidas de dimensión consistente. Su expresión es la siguiente:

$$BIC = 2\ell(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y}) + p \cdot log(n) \tag{2.14}$$

- *Quasi-Criterio de Información de Akaike*(o "Quasi-Akaike Information Criterion", QAIC) es una modificación del criterio de AIC que se centra en los casos en los que hay sobredispersión de los datos, es decir, la varianza muestral es mayor que la varianza teórica. La fórmula es la siguiente:

$$QAIC = -\left[\frac{2\ell(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y})}{\hat{c}}\right] + 2p \tag{2.15}$$

donde $\hat{c}$ es el cociente del estadístico χ^2 (2.4) de bondad de ajuste del modelo y los grados de libertad.

Ejemplo 2.3. Este ejemplo pretende aplicar las diferentes medidas vistas en el capítulo de manera más práctica, mediante la aplicación de Modelos Lineales Generalizados a un conjunto de datos sencillo. Estos datos provienen de [3] y tratan sobre la evolución del calor durante el endurecimiento del cemento en Portland. Estos datos tienen 4 variables regresoras x_i que definen la proporción en porcentaje de ciertos químicos dentro del cemento y la variable de estudio y trata de las calorías emitidas durante el endurecimiento del cemento por gramo después de 180 días.

El conjunto de datos se puede ver en la Tabla 2.2

x_1	x_2	x_3	x_4	y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

Tabla 2.2. Datos del cemento de Portland

El objetivo del estudio es aplicar un Modelo Lineal Generalizado para predecir el comportamiento esperado de la variable y.

Para ello, se utilizó un modelo con distribución normal y con función de enlace la identidad. Se utiliza SPSS para realizar las distintas estimaciones de los parámetros y para el cálculo de algunas de las medidas

La Tabla 2.3 muestra algunas de las métricas calculadas para el modelo donde se consideran todas las variables x_i.

Métrica	Valor
Devianza	47.864
Devianza escalada	13
Chi-cuadrado de Pearson	47.864
Max Log-Verosimilitud	-26.918
AIC	65.837
AIC_C	79.837
BIC	69.226

Tabla 2.3. Alguna de las métricas

La Tabla 2.4 muestra algunas de las métricas calculadas para el modelo con solo las variables x_1 y x_2.

Métrica	Valor
Devianza	57.904
Devianza escalada	13
Chi-cuadrado de Pearson	57.904
Max Log-Verosimilitud	-28.156
AIC	64.312
AIC_C	69.312
BIC	66.572

Tabla 2.4. Caption

Observando las primeras cuatro métricas de las tablas, se puede observar que toman valores similares en la máxima log-verosimilitud y en la devianza escalada. Sin embargo, en la Devianza y la Chi-cuadrado si presentan diferencias los modelos de al menos 10 unidades, luego se puede considerar inicialmente al modelo con menor valor de estas, es decir, al modelo con todas las x_i como mejor modelo.

Por otra parte, el tamaño de la muestra $n = 13$ y el número de variables es $p = 4$, entonces $\frac{n}{p} = \frac{13}{4} < 40$, luego hay que elegir AIC_C en lugar de AIC en la comparación de las medidas. Se puede observar que en esta ocasión también hay una diferencia de 10 unidades entre las medidas, eligiendo como mejor modelo aquella con menor valor del AIC_C, es decir, el segundo modelo.

A simple vista, se puede observar que los resultados de las medidas con contradictorios. No obstante, sabemos que las métricas AIC y AIC_C cumplen con el principio de parsimonia, es decir, buscan un modelo que explique bien los datos sin ser muy complicado, mientras que las medidas basadas en la verosimilitud solo se centran en mejorar la precisión, permitiendo en muchas ocasiones el sobreajuste como ocurre en este caso. Al tomar todas las variables x_i, el modelo que surge es un modelo complejo para el tamaño muestral que se tiene.

Como conclusiones del estudio, se prefiere el modelo con solo las variables x_1 y x_2, pues es un modelo suficientemente preciso y sencillo para la muestra tomada.

2.3. Inferencia y Tests de Hipótesis para GLM

La inferencia en los Modelos Lineales Generalizados constituye un aspecto fundamental en el análisis de estos modelos, brindando herramientas para la comprensión y la toma de decisiones basadas en los datos. La importancia de

la inferencia en GLM radica en la necesidad de verificar si las estimaciones realizadas tienen significación bajo un cierto nivel de confianza suficientemente fiable. De igual manera, se pueden utilizar otros tests como el de comparación de modelos anidados para contrastar las diferencias.

2.3.1. Tests de comparación de modelos anidados

Con el objetivo de cumplir con el principio de parsimonia, uno de los objetivos de los test de hipótesis es verificar si al eliminar o añadir variables al modelo afecta significativamente a la precisión del modelo. El test de comparación de modelos anidados permite verificar si la inclusión de parámetros adicionales en el modelo más complejo mejora significativamente la calidad de ajuste, o si el modelo más simple sigue siendo preferible.

Sean dos Modelos Lineales Generalizados, basados en la misma distribución pero con distinto predictor lineal:
Modelo A: $g(\hat{\mu_A}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta_{p_A}} x_{p_A}$
Modelo B: $g(\hat{\mu_B}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta_{p_A}} x_{p_A} + ... + \hat{\beta_{p_B}} x_{p_B}$
con $p_A < p_B$.

Se puede observar que el Modelo A es un caso particular del B, donde los parámetros $\hat{\beta}$, a partir de p_A, son igual a 0 en el modelo A. Cuando se da esta situación, se dice que el modelo A está anidado en el modelo B.

Para determinar si el Modelo A es tan bueno como el Modelo B, se puede realizar el siguiente test de hipótesis:

$$H_0 : \hat{\beta_{p_A+1}} = \hat{\beta_{p_A+2}} = ... = \hat{\beta_{p_B}} = 0$$

$$H_1 : \exists j \in \{p_A + 1, p_A + 2, ..., p_B\} : \hat{\beta}_j \neq 0$$

En la sección de métricas 2.2.1 anterior, se observó que la devianza escalada (2.3) se distribuía como una χ^2 de p grados de libertad, siendo p el número de parámetros de estudio. Teniendo en cuenta esta afirmación, se puede definir el estimador similar para el test:

$$L = D^*(y, \hat{\mu_A}) - D^*(y, \hat{\mu_B})$$

Este estimador se distribuye como una chi cuadrado de $p_A - p_B$ grados de libertad, es decir, $L \sim \chi^2_{p_A - p_B}$, lo que permite la definición del test y la aplicación de niveles de confianza para determinar la fiabilidad de los coeficientes estimados de los modelos anidados.

2.3.2. Contraste Omnibus

La prueba o test Omnibus es un test de hipótesis de los Modelos Lineales Generalizados en el que se compara se manera significativa si los coeficientes del modelo actual son todos 0 o existe alguno distinto de cero. Para ello, el test Omnibus se basa en el test de hipótesis anterior, donde el Modelo A es el modelo nulo (el modelo que solo tiene el coeficiente constante β_0), mientras que el modelo B es el modelo que se quiere estudiar con p parámetros. El test de hipótesis se puede redactar de la siguiente forma:

$$H_0 : \beta_1 = ... = \beta_p = 0$$

$$H_1 : \exists j \in \{1, ..., n\} : \beta_j \neq 0$$

Con este test se puede verificar si las variables elegidas X_i tienen importancia en la predicción del valor esperado de la variable de estudio, o por lo contrario, estas aportan tanto como el término constante.

Cabe destacar que si este test da como correcta la hipótesis alternativa, se desconocen qué variables son las que aportan o no en la predicción del modelo.

2.3.3. Tests de Wald

Con el objetivo de determinar qué coeficientes estimados $\hat{\beta}_j$ del modelo son verdaderamente significativos, se lleva a cabo una prueba de Wald a cada coeficiente. Esta prueba consiste en realizar el siguiente test de hipótesis a cada uno de los coeficientes del modelo:

$$H_0 : \hat{\beta}_j = \beta_j^0$$

$$H_1 : \hat{\beta}_j \neq \beta_j^0$$

El estadístico de comparación de este test es el siguiente:

$$Z = \frac{(\hat{\beta}_j - \beta_j^0)^2}{var(\hat{\beta}_j)}$$

Donde $var(\hat{\beta}_j)$ es la varianza estimada de $\hat{\beta}_j$ y este estimador se compara con una normal de media cero y varianza 1.

Para poder verificar si los coeficientes son no nulos, simplemente hay que realizar el test con $\beta_j^0 = 0$

2.3.4. Intervalos de confianza de Wald

También se pueden definir los intervalos de confianza de Wald. Para ello, se define el intervalo de confianza de Wald para

$$\hat{\beta}_j \pm z_{\alpha/2}\sqrt{var(\hat{\beta}_j)}$$

donde $z_{\alpha/2}$ proviene de la normal con media cero y varianza 1 y con α siendo el nivel de confianza del intervalo.

Ejemplo 2.4. Se utilizan los mismos datos del ejemplo 2.3 de la sección de métricas. Se tienen 4 variables regresoras x_i y una variable de estudio y que se busca estudiar su valor esperado utilizando un Modelo Lineal Generalizado.

En este caso se utiliza el modelo considerando las cuatro variables, con distribución normal y con función de enlace la identidad.

Se considera un primer modelo tomando todas las variables x_i, donde los resultados de los tests se pueden ver en la Tabla 2.5 y 2.6.

Test	Significación
Omnibus	< 0.001

Tabla 2.5. Prueba de Omnibus

Variable	Valor estimado	Significación del test de Wald	Intervalo de Confianza
β_0	62.405	0.256	[-45.330, 170.141]
β_1	1.551	0.008	[0.406, 2.696]
β_2	0.510	0.369	[-0.603, 1.623]
β_3	0.102	0.863	[-1.058, 1.262]
β_4	-0-144	0.796	[-1.234, 0.946]

Tabla 2.6. Test de Wald e intervalo de confianza al 0.95

El segundo modelo a considerar es, al igual que en el ejemplo 2.3, un modelo considerando solo las variables x_1 y x_2. Los resultados de los tests se pueden revisar en las tablas 2.7 y 2.8.

Test	Significación
Omnibus	< 0.001

Tabla 2.7. Prueba de Omnibus

Tanto en el primer modelo como en el segundo, el test de Omnibus sugieren la hipótesis alternativa de que hay variables con coeficiente distinto de cero.

Variable	Valor estimado	Significación del test de Wald	Intervalo de Confianza
β_0	52.577	< 0.001	[48.647, 56.507]
β_1	1.468	< 0.001	[1.260, 1.677]
β_2	0.662	< 0.001	[0.583, 0.741]

Tabla 2.8. Test de Wald e intervalo de confianza al 0.95

Sin embargo, en el primer modelo, solo el coeficiente β_1 llega a ser significativamente distinto de cero, con nivel de confianza del 0.95, mientras que en el segundo modelo, todos los coeficientes tomados son significativamente distintos de cero.

Observando también los intervalos de confianza del Wald se puede concluir resultados parecidos a los del test, los intervalos de confianza del primer modelo son mucho más largos y muchos de ellos contienen al 0, mientras que los del segundo modelo son bastante pequeños en comparación y no contienen al cero.

3

Diagnóstico de los Modelos Lineales Generalizados

En este capítulo se van a mostrar diferentes herramientas para poder verificar las hipótesis de los Modelos Lineales Generalizados y detectar valores anómalos dentro de los datos.

3.1. Residuales

Los residuales son una herramienta fundamental para revisar cuánto de adecuado es el modelo. La definición clásica de residual es la diferencia entre los valores reales y los valores predichos del modelo, es decir, $y_i - \hat{\mu}_i$. Esta definición se utiliza en los modelos de regresión lineal, donde se puede comprobar si estos siguen distribución normal. Sin embargo, en el caso de GLM, es necesario definir otro tipo de residuos que se puedan aplicar a todas las distribuciones de la familia exponencial, pues al poder tomar variables con distribuciones no normales, los residuos anteriores no tienen por qué seguir la distribución normal.

Idealmente se buscan residuos que se comporten de forma similar a los de la regresión lineal, es decir, con media cero y varianza constante, y si es posible, con distribución normal. Pero los residuos de los Modelos Lineales Generalizados pueden no tener varianza contante o no seguir una distribución gaussiana.

3.1.1. Residuos de Pearson

La forma mas directa de tratar con la varianza no constante es dividir por ella. De esta premisa, se puede definir los residuos de Pearson:

$$r_P = \frac{y - \mu}{\sqrt{V(\mu)}}$$

donde $V(\cdot)$ es la función de varianza.

Este residuo cumple con la propiedad de $\sum r_P^2 = \chi^2$, siendo χ^2 el estimador de Pearson.

Ejemplo 3.1. Para una distribución de Poisson de parámetro μ, los residuos de Pearson son: $r_P = \frac{y-\mu}{\sqrt{\mu}}$.

Ejemplo 3.2. Para una distribución de Binomial de parámetros n y p, los residuos de Pearson son: $r_P = \frac{y-np}{\sqrt{np(1-p)}}$.

3.1.2. Residuos de Anscombe

Un problema que tiene el residuo de Pearson es que para distribuciones no normales, la distribución de los residuos puede presentar asimetría, haciendo que no se distribuya como una Normal.

Una alternativa para resolver este problema es utilizar el residual de Anscombre. Estos se definen una función $A(y)$, que es aquella que permite que la distribución de $A(Y)$ sea lo más normal posible. En este tipo de modelos, la función viene definida de la siguiente forma:

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

Esta transformación permite que los residuos sean más normales. Sin embargo, hay que dividir estos residuos por la raíz de la varianza de $A(Y)$, es decir, por $A'(\mu)\sqrt{V(\mu)}$.

Ejemplo 3.3. Para una distribución de Poisson de parámetro μ, la función $A(\cdot)$ es:

$$A(\mu) = \int \frac{d\mu}{\mu^{\frac{1}{3}}} = \frac{3}{2}\mu^{\frac{2}{3}}$$

Luego, aplicando la función a cada uno de los sumandos del residuo $y - \mu$, tenemos $\frac{3}{2}(y^{\frac{2}{3}} - \mu^{\frac{2}{3}})$. Calculamos $A'(\mu)\sqrt{V(\mu)} = \mu^{\frac{1}{6}}$ y concluimos con la expresión del residuo de Anscombe para la Poisson:

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

A pesar de las buenas cualidades, no suele ser utilizado por su dificultad de cálculo (véase [20]).

3.1.3. Residuos de Devianza

Utiliza una medida basada en la Devianza llamada *la unidad de devianza* $d(y, \mu)$, que se define como:

$$d(y, \mu) = 2(t(y, y) - t(y, \mu))$$

con $t(y, \mu) = y\theta - b(\theta)$.

Nótese que, como se ha demostrado que θ se puede poner en función de la esperanza μ, luego la función $t(y, \mu)$ está definida correctamente.

Con ella, se define el residuo de devianza:

$$r_D = sign(y - \mu)\sqrt{d(y, \mu)}$$

donde $sign(x)$ es el signo de esa resta (-1 si es negativa o 1 si es positiva).

Ejemplo 3.4. Los residuos de devianza cuando se sigue una distribución de Poission son:

$$r_D = sign(y - \mu)\sqrt{2(yln(y/\mu) - y + \mu)}$$

3.2. Diagnóstico de Hipótesis del modelo

En la construcción de los Modelos Lineales Generalizados, se precisan de ciertos requerimientos para que el modelo funcione correctamente. Aunque los GLM reduzcan el número de requisitos que necesita, este aún sigue teniendo varios aspectos a tener en cuenta. Las condiciones del modelo son las siguientes:

- Independencia de los datos. Consiste en garantizar que los datos y_i provienen de variables independientes.
- Homocedasticidad. Las varianzas de los errores deben ser constantes y no depender de las variables explicativas.
- Componente aleatoria. Hay que verificar si la distribución elegida es la adecuada en función de los datos y tomados.
- Función de enlace. Es importante que se utilice una función de enlace correcta, para poder relacionar el predictor lineal con la componente aleatoria.
- Regresor lineal. Se tiene que determinar si hay que añadir o eliminar variables del predictor lineal. Dichas variables deben tener una escala correcta, para que el modelo no dé mayor importancia a unas variables que a otras.
- Datos atípicos. Al igual que pasa en los modelos de regresión lineal, pueden haber algunos datos que tomen valores muy elevados o muy pequeños comparados con el resto de la muestra. Este tipo de observaciones afectan a la estimación del modelo y deben de ser detectados y tratados para reducir su efecto.

3.2.1. Independencia de los datos

Es fundamental garantizar que los datos son independientes entre sí, es decir, el valor de los datos tomados en un instante del proceso de muestreo no dependen del valor de los datos tomados anteriormente. Si se da esta situación, los datos presentan un problema de multicolinealidad, que afecta a las estimaciones del modelo.

Existen varias medidas basadas en los valores singulares de la matriz modelo X (véase [27]), o se aplican otros métodos basados en la regresión Ridge o Lasso (véase [15]).

Una forma visual que permite ver la dependencia respecto del tiempo es realizar una gráfica de contraste entre los residuos en el instante i y los residuos en el instante anterior $i-1$. Esto permite buscar posibles patrones dentro de los datos, asumiendo que los datos están ordenados cronológicamente. Si se distinguen patrones, los datos pueden no ser independientes. Este criterio es meramente indicativo, pues al ser una representación gráfica, no se puede saber significativamente si este comportamiento es o no el deseado.

3.2.2. Homocedasticidad

Una forma sencilla de visualizar si los datos son homocedásticos, es realizar un gráfico con el valor del predictor y los residuos de la devianza. Este nos permite identificar si las varianzas tienen algún tipo de patrón creciente. Si lo presentan, hay heterocedasticidad, que se puede prevenir utilizando estimadores robustos (ver [25]).

3.2.3. Componente aleatoria

Hay varias formas sencillas de verificar que la distribución es correcta. Se pueden utilizar tests de hipótesis, como los de normalidad, aplicados a la variable dependiente o se pueden realizar gráficas como el Q-Q plot, que compara los cuantiles estimados con los de la distribución.

También se pueden revisar si los residuos cumplen condiciones similares a los de regresión lineal, como esperanza nula, varianza constante y normalidad.

3.2.4. Función de enlace

En la elección de la mejor función de enlace, se suelen tomar distintas funciones y elegir aquella que da mejores resultados en las métricas. Dado que se busca que la relación $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ sea lineal, se pueden realizar gráficas entre los residuos de devianza y los valores estimados del predictor lineal.

Además, Existen otros métodos en la literatura, como tomar la función de enlace como desconocida y estimarla (véase [2]) o métodos de comparación

de funciones de enlace cuando ambas dependen de un parámetro común (véase [23]).

3.2.5. Regresor lineal

Se necesita tratar los problemas de escala, pues el modelo puede priorizar variables no explicativas frente a otras verdaderamente importantes para la predicción de la variable dependiente. La solución más usual es normalizar las variables para evitar estos comportamientos no deseados, aunque se pueden realizar otras transformaciones de las variables como: elevarla a un exponente X_j^t o aplicar logaritmos $ln(X_j)$.

3.3. Detección de valores atípicos

En el análisis de los datos, existen individuos en la muestra que influyen más o menos en el cálculo y estimación del modelo. Con el término influir, se pretende decir que, si eliminamos o modificamos los datos con mayor influencia, esto cambia el modelo completamente.

Una medida para calcular la importancia de las variables es mediante la matriz $\mathbf{H}$, que proviene de comparar los valores estimados $\hat{\boldsymbol{\mu}}$ del modelo con los valores observados $\mathbf{y}$, teniendo esta relación:

$$\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$$

Ejemplo 3.5. En regresión lineal, $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Luego, la matriz en este caso particular es: $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Se define la importancia de la observación i como el elemento h_{ii} de la matriz $\mathbf{H}$.

Estudiar la importancia de los datos es de especial interés en la presencia de datos atípicos. Este tipo de observaciones son inusuales o extremas, que difieren bastante del resto de datos del conjunto. Pueden surgir por errores en la medición, en el procesamiento o incluso en el muestreo de datos. Estos no son necesariamente erróneos, pueden representar eventos poco comunes de la variable que se esté estudiando.

Un dato como este puede generar problemas en la estimación del modelo si llegara a ser influyente. Por tanto, es necesario realizar una fase de análisis para detectar si el conjunto de datos presenta este tipo de situaciones.

Para ello, se utiliza la distancia de Cook, una medida que compara las estimaciones de los parámetros $\boldsymbol{\beta}$, con y sin esa observación. Esta se puede definir de la siguiente forma:

$$C_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'W\mathbf{X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p \cdot \phi}$$

donde $\hat{\boldsymbol{\beta}}_{(i)}$ son los estimadores obtenidos sin la observación i, p es el número de variables y ϕ es el parámetro de dispersión.

Una forma de estimar este valor es mediante la siguiente aproximación:

$$C_i \approx \frac{h_{ii}(r_P)^2}{\phi p(1 - h_{ii})}$$

Una forma común de detectar la presencia de valores atípicos influyentes es mediante la representación gráfica de la medida de influencia y de la distancia de Cook. Los valores que se encuentren más arriba a la derecha de la gráfica son las observaciones anómalas influyentes.

4

Aplicación de Modelos Lineales Generalizados en el análisis de comportamiento de reptiles

Dada la versatilidad de aplicación en modelización estadística y computación de los Modelos Lineales Generalizados, y habiendo tenido la posibilidad de realizar una investigación matemática, se aborda el problema como aplicación del grupo de ecología de la Universidad de La Laguna.

En este capítulo se trata de aplicar los conceptos teóricos vistos en las secciones anteriores, para estudiar la agresividad de dos subespecies de lagartos en la isla de Tenerife.

4.1. Antecedentes del problema

Los datos de estudio provienen del artículo [3], donde se busca analizar el comportamiento de los reptiles, en base a factores morfológicos, de comportamiento o de reflectancia de la luz en la piel de los lagartos.

A lo largo del documento, se hace énfasis en las manchas de la piel de los reptiles, y en como la reflectancia de luz ultravioleta-visible es un buen indicador de habilidad de combate o de dominio en los individuos. Esto da la iniciativa de intentar relacionar este tipo de aspectos a la agresividad y a la supervivencia de los reptiles.

Para comprobar estas relaciones se llevan a cabo dos experimentos:

- Un primer experimento, donde se consideran todas las cualidades morfológicas, de comportamiento y de reflectancia del reptil, y se busca determinar aquellas influyentes en los resultados de peleas entre los reptiles. Se enfrentan parejas de lagartos en un entorno controlado hasta que uno de ellos de indicaciones de derrota, como evitando el conflicto con el otro reptil o intentando salir del recinto controlado.
- Un segundo experimento, que quiere determinar si la luz ultravioleta-visible azul de las manchas en la piel del reptil afectan en las victorias o derrotas de los lagartos. Para ello, en cada encuentro entre los reptiles, se toma uno al

azar y se le aplica una crema solar para reducir la reflectancia de las manchas de la piel.

Las muestras para cada una de las pruebas son de individuos de dos subespecies de lagartos, que fueron tomadas en dos localidades de Tenerife: El Pris y el Malpaís de Güímar, entre 2005 y 2006. El tamaño muestral de cada experimento es de 108 lagartos, tomados a partes iguales entre las dos subespecies.

4.2. Objetivos del estudio y variables

En el análisis propuesto en el artículo, se utilizan Modelos Lineales Generalizados para predecir si los individuos ganaban o no las peleas en cada uno de los experimentos.

Por tanto, en este capítulo, se va a tratar de la misma forma los datos, para intentar llegar a resultados similares a los del artículo. El objetivo es revisar los métodos desarrollados en la memoria y garantizar la reproducibilidad de los resultados del artículo.

Para la primera parte del estudio, se toman las siguientes variables:

- La variable de estudio Y, que indica si el lagarto ganó o perdió la pelea.

$$Y = \begin{cases} 1, & \text{si ganó la pelea} \\ 2, & \text{si perdió la pelea} \end{cases}$$

- La localización de los individuos, que denotamos con la variable binaria L.

$$L = \begin{cases} 1, & \text{si es del Malpaís de Güímar} \\ 2, & \text{si es del Pris de Tacoronte} \end{cases}$$

- Unas variables continuas morfológicas, que se denotan en el estudio como M_j, $\forall j \in \{1, 2, 3, 4, 5, 6\}$. En la tabla 4.1 se puede ver el significado de cada una.

Variable	Significado
M_1	Longitud hocico-cloaca
M_2	Masa corporal
M_3	Ancho de la cabeza
M_4	Longitud de la extremidad anterior
M_5	Longitud de la extremidad posterior
M_6	raíz cuadrada del área total de ocelos

Tabla 4.1. Variables Morfológicas

Variable	Significado
C_1	Frecuencia relativa de hinchar gola
C_2	Frecuencia relativa de sacar lengua
C_3	Frecuencia relativa de mordidas
C_4	Frecuencia relativa de sacudir cola

Tabla 4.2. Variables de Comportamiento

- Unas variables continuas de comportamiento, denotadas como C_j, $\forall j \in \{1, 2, 3, 4\}$. En la tabla 4.2 se puede observar una breve descripción de cada una.
- Unas variables continuas de reflectancia, que las modelizamos como R_j, $\forall j \in \{1, 2, 3, 4, 5\}$. En la tabla 4.3 se pueden ver las medidas tratadas.

Variable	Significado
R_1	Arco seno raíz cuadrada de la reflectancia entre 300 y 495 nm
R_2	Arco seno raíz cuadrada de la reflectancia entre 495 y 700 nm
R_3	Arco seno raíz cuadrada de la reflectancia entre 300 y 700 nm
R_4	Arco seno raíz cuadrada de reflectancia en el rango ultravioleta
R_5	Longitud de onda pico

Tabla 4.3. Variables Morfológicas

En general, los objetivos de la primera parte del análisis se pueden resumir de la siguiente forma: Ver la influencia de las variables predictoras sobre la categoría de ganador-perdedor (variable dependiente Y) e identificar el conjunto de variables que más influyen en la victoria o derrota de los reptiles.

La segunda parte del estudio se centra en la importancia de la aplicación de la crema en los reptiles. Las variables que se eligieron para esta segunda parte del estudio son:

- Tres variables binarias, que indican el lugar L, si el lagarto tiene o no crema C y si ganó o perdió la pelea P. Los lugares son los mismos que en la parte 1 del análisis, mientras que C e Y toman los siguientes valores:

$$C = \begin{cases} 0, & \text{si no tiene crema} \\ 1, & \text{si sí tiene crema} \end{cases}$$

$$Y = \begin{cases} 0, & \text{si ganó} \\ 1, & \text{si perdió} \end{cases}$$

- Se introducen 6 variables continuas morfológicas, que describen cualidades similares a las de la primera parte. Las denotamos como M_j, $\forall j \in \{1, 2, 3, 4, 5, 6\}$ y su descripción se puede ver en la tabla 4.4.

Variable	Significado
M_1	Longitud hocico-cloaca
M_2	Masa corporal
M_3	Ancho de la cabeza
M_4	Altura de la cabeza
M_5	Longitud de la extremidad anterior
M_6	Longitud de la extremidad posterior

Tabla 4.4. Variables Morfológicas

- Por último, también se consideran 3 variables de comportamiento C_j, $\forall j \in \{1, 2, 3\}$ parecidas a las de la primera parte. En la tabla 4.5 se puede observar el significado de cada variable.

Variable	Significado
C_1	Arco seno de frecuencia relativa de hinchar gola
C_2	Arco seno de frecuencia relativa de sacar lengua
C_3	Arco seno de frecuencia relativa de mordidas

Tabla 4.5. Variables de Comportamiento

Los objetivos de la segunda parte se resumen en: ver la influencia de las variables predictoras (solo morfológicas y de comportamiento) sobre la categoría de ganador-perdedor y tener especial consideración si el individuo tenía o no crema solar aplicada.

4.3. Modelo

Para llevar a cabo el análisis, se van a aplicar Modelos Lineales Generalizados a cada una de las partes del estudio. Dado que en ambos casos la variable de estudio es binaria, se opta por la versión logística de los GLM.

Para definir correctamente un GLM, hacen falta la componente aleatoria, la componente sistemática y la función de enlace, vistas en 1.

- La componente aleatoria viene definida por la categoría ganador-perdedor Y en ambas partes. Al solo tomar dos valores, se puede considerar una Bernoulli de parámetro P. Cabe destacar que la distribución Bernoulli pertenece a la familia exponencial, pues observando el ejemplo 1.4, basta con tomar $\theta = \ln[\frac{p}{1-p}]$, $a(\phi) = 1$, $b(\theta) = \ln[1 + \exp\{\theta\}]$, $\phi = 1$ y $c(y, \phi) = 0$ dentro de la fórmula (1.2) para obtener la distribución.
- La componente sistemática representa una combinación lineal de las variables que se quieren estudiar. Dado que se realizan dos análisis con distintas variables, los predictores lineales tendrán distinta forma.

El predictor lineal de la primera parte es el siguiente:

$$\eta_1 = \beta_0 + \beta_1 L + \sum_{j=1}^{6} \beta_{j+1} M_j + \sum_{j=1}^{4} \beta_{j+7} C_j + \sum_{j=1}^{5} \beta_{j+11} R_j$$

El predictor lineal de la segunda parte es el siguiente:

$$\eta_2 = \beta_0 + \beta_1 L + \beta_2 C + \sum_{j=1}^{6} \beta_{j+2} M_j + \sum_{j=1}^{3} \beta_{j+8} C_j$$

Cabe destacar que estos son los predictores lineales para los modelos completos, es decir, los que tienen en cuenta todas las variables de estudio. Si se quiere el predictor lineal teórico de un subconjunto determinado de variables, basta con tomar $\beta_j = 0$ en aquellas que no se van a utilizar.

- Por último, la función de enlace tomada en ambos casos es la función de enlace logit, descrita como la función de enlace canónica de la distribución binomial:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = ln\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right)$$

Esto hace que si estimamos los β_j del predictor lineal, se puede despejar μ en la fórmula y se obtenga la siguiente expresión:

$$\boldsymbol{\mu} = \frac{e^{\boldsymbol{\eta}}}{1+e^{\boldsymbol{\eta}}}$$

Donde se puede sustituir directamente el valor del predictor lineal estimado.

4.4. Análisis

En esta sección se van a realizar diferentes pruebas en ambas partes del estudio. Se introducen dos secciones, una con cada parte del análisis, para tratar los datos de manera más precisa y encontrar el modelo que mejor explique los datos y cumpla con los objetivos descritos.

Para el desarrollo de los cálculos de los diferentes parámetros y métricas, se utilizó la licencia de SPSS facilitada por la Universidad de La Laguna en los escritorios virtuales.

Variable	Datos completos	Mínimo	Máximo	Media	Desviación Típica
Y	108	1	2	1.50	0.502
L	108	1	2	1.46	0.501
M_1	105	1.9542	2.1303	2.0603	0.0285
M_2	95	1.4456	1.8949	1.7025	0.0919
M_3	95	1.0577	1.4370	1.2395	0.8049
M_4	95	1.4565	1.7710	1.5880	0.3923
M_5	94	1.6601	1.8697	1.7762	0.0363
M_6	94	0.6987	1.6461	1.2505	0.1881
C_1	98	0.0000	0.2527	0.0635	0.0555
C_2	100	0.0000	0.3741	0.0966	0.0885
C_3	100	0.0000	0.2355	0.0307	0.5463
C_4	100	0.0000	0.2175	0.0216	0.0395
R_1	78	0.2721	0.6923	0.4768	0.9820
R_2	78	0.1919	0.5460	0.3350	0.0670
R_3	78	0.3368	0.9664	0.6042	0.1245
R_4	78	0.7546	1.1024	0.9478	0.0686
R_5	78	2.5476	2.6283	2.5785	0.0140

Tabla 4.6. Estadísticos básicos

4.4.1. Parte I

Estadística Descriptiva

Introducimos el análisis de la primera parte con algunos datos de estadística descriptiva, pues esto permite resumir los datos de estudio.

La tabla 3.1 muestra los estadísticos descriptivos como la media, la desviación típica, el valor máximo y el valor mínimo. Cabe destacar también que de los 108 datos que posee la muestra, hay algunos registros incompletos, haciendo que solo se pueda trabajar con 75 reptiles de los 108 muestreados.

Previo a realizar cualquier estimación, normalizamos todas las variables continuas, es decir, para cada variable distinta de Y y de L y para cada dato, se resta el valor de la media y se divide por la desviación típica de cada variable.

Comparación de Modelos

Para intentar encontrar el mejor modelo, se realizan comparaciones entre modelos con diferentes variables. Como criterio de decisión se utiliza el AIC_C, pues el número de datos completos es $n = 75$ y el número de variables predictoras es $p = 16$, por tanto, $\frac{75}{16} = 4.6875 < 40$, indicando que se debe utilizar AIC_C como medida comparativa para obtener mejores resultados.

Comparar todos los modelos posibles es una tarea complicada, pues el conjunto de datos presenta un alto número de variables de estudio (hay $2^{16} = 65536$ modelos posibles). En este tipo se situaciones, es recomendable tomar algunos subconjuntos de variables que puedan ser de interés estudiar, agrupándolas por

grupos o revisando la significación. Se toman las siguientes agrupaciones de variables:

- *Nulo*: El modelo sin ninguna variable.
- *Completo*: El modelo con todas las variables.
- *Solo Sig.*: El modelo con solo las variables significativas del modelo completo.
- *Experto*: El modelo con las variables sugeridas por un experto en el tema y uno de los autores del artículo [3].
- *Hosmer-Lemeshow*: En la literatura, una propuesta que recomiendan estos autores es tomar solo las variables del modelo completo con significación menor que 0.25.
- *Solo Morfológicas*: El modelo con solo las variables morfológicas.
- *Solo Comportamiento*: El modelo con solo las variables de comportamiento.
- *Solo Reflectancia*: El modelo con solo las variables de Reflectancia.

Los resultados de las comparativas utilizando AIC_C y Δ_i se pueden observar en la tabla 4.7.

Modelo	AIC_C	Δ_i
Nulo	105.693	33.079
Completo	72.614	17.951
Solo Sig.	57.449	2.390
Experto	96.866	39.702
Hosmer-Lemeshow	55.059	0
Solo Morfológicas	133.917	77.088
Solo Comportamiento	90.422	35.234
Solo Reflectancia	113.329	56.273

Tabla 4.7. Alguna de las métricas

Analizando la tabla, los modelos a considerar son el modelo con solo las variables significativas del modelo completo y aquel con las variables con significación menor a 0.25, utilizando el criterio de Hosmer y Lemeshow. De entre ellos, se va a elegir este último, pues es que que menor valor de AIC_C tiene.

El mejor subconjunto de variables

Este modelo, como se indicó en la sección anterior, toma las variables del modelo con significación menor a 0.25. Estas variables son: M_1, M_2, M_3, M_4, C_2, C_3, C_4 y R_5.

En la tabla 3.3 se pueden ver los resultados de las métricas, en la tabla 3.4 la prueba Omnibus y en la tabla 3.5 los valores estimados y la significación y los intervalos de confianza de Wald al nivel de confianza de 0.95.

Métrica	Valor
Devianza	36.460
Devianza escalada	66.776
Chi-cuadrado de Pearson	36.036
Max Log-Verosimilitud	-18.230
AIC	54.460
AIC_C	57.229
BIC	75.317

Tabla 4.8. Métricas de bondad de ajuste

Test	Significación
Omnibus	< 0.001

Tabla 4.9. Prueba de Omnibus

Variable	Valor estimado	Significación del test de Wald	Intervalo de Confianza
β_0	-0.383	0.368	[-1.218, 0.452]
β_2	0.133	0.841	[-1.169, 1.435]
β_3	1.773	0.011	[0.398, 3.148]
β_4	0.487	0.255	[-0.352, 1.326]
β_5	-1.302	0.016	[-2.362, -0.243]
β_9	1.714	0.003	[0.573, 2.855]
β_{10}	1.773	0.007	[0.599, 3.779]
β_{11}	0.487	< 0.000	[-7.441, -2.711]
β_{16}	-1.302	0.020	[0.184, 2.110]

Tabla 4.10. Test de Wald e intervalo de confianza al 0.95

Observando sobretodo la última tabla 3.5 y tomando solo las variables significativas, se puede concluir que el mejor modelo estimado obtenido para medir estos datos es el siguiente:

$$\eta_1 = \beta_3 M_2 + \beta_5 M_4 + \beta_9 C_2 + \beta_{10} C_3 + \beta_{11} C_4 + \beta_{16} R_5$$

Sustituyendo los valores de las estimaciones, se tiene el siguiente modelo:

$$\eta_1 = 1.773 M_2 - 1.302 M_4 + 1.714 C_2 + 1.773 C_3 + 0.487 C_4 - 1.302 R_5$$

$$\hat{\mu} = \frac{e^{\eta_1}}{1 + e^{\eta_1}}$$

Diagnóstico de Hipótesis

Por último, se va a estudiar si el modelo cumple con las hipótesis de los Modelos Lineales Generalizados, para ver los posibles defectos que puede tener el modelo.

La distribución es correcta, pues al ser una variable binaria, la distribución no puede ser otra que una Bernoulli. También se han solucionado los problemas de escala que pueden tener las covariables normalizándolas al principio del análisis.

Para analizar la independencia, se realiza una gráfica con los residuos de Pearson del dato i como eje x y los del dato $i-1$ como eje y.

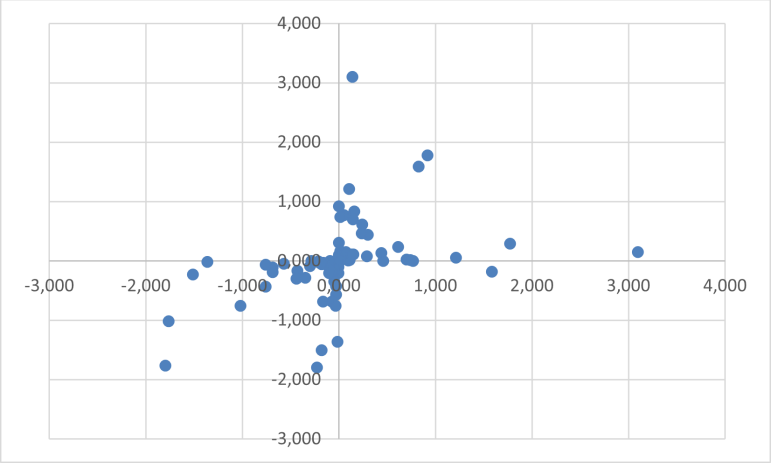

Figura 4.1. Gráfico de Independencia

En la figura 4.1 se puede observar que no presenta ningún tipo de patrón, por tanto, se puede asumir que se cumple la independencia.

Para comprobar la homocedasticidad y la función de enlace se contrastan los residuos de devianza con los predictores lineales. En la figura 4.2 se pueden observar los resultados.

Cuanto más cerca estén las curvas con mayor densidad de puntos, mejor es el modelo. En esta ocasión están suficientemente cerca para considerar el modelo bastante bueno.

Para detectar valores atípicos, se realiza una gráfica con los valores de importancia de los datos y sus respectivas distancias de Cook. En la figura 4.3 está representado.

Los valores anómalos son los que se encuentran en la parte de arriba a la derecha de la gráfica. Se puede ver que hay un valor atípico con bastante importancia. Si se intenta eliminar surgen otros valores anómalos con mayor distancia de Cook. Luego, no se van a realizar modificaciones en los datos y se va a asumir está deficiencia del modelo.

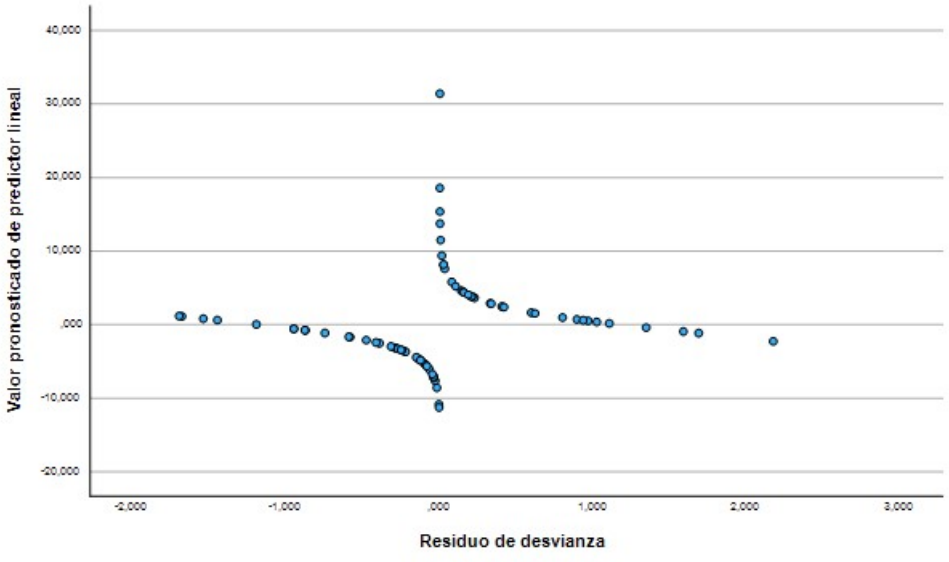

Figura 4.2. Residuos de devianza en función de los predictores lineales

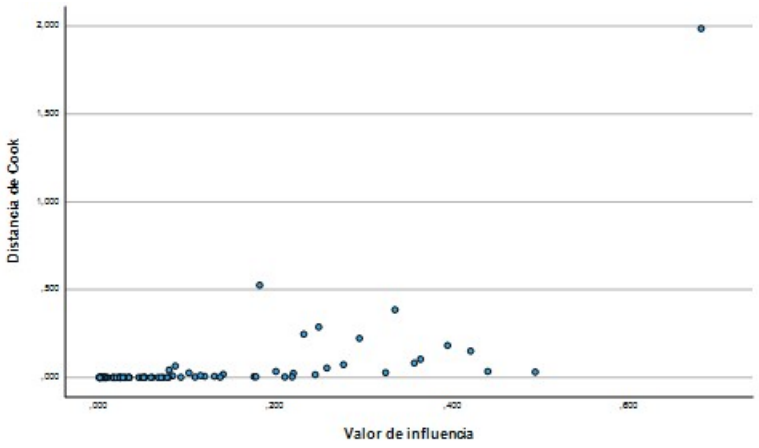

Figura 4.3. Importancia o influencia de los datos frente a la distancia de Cook

4.4.2. Parte II

Estadística Descriptiva

Al igual que en la primera parte, se inicia el estudio de los datos de la segunda con estadística descriptiva, con objetivos de realizar un resumen de los datos, así como intentar buscar datos atípicos.

En la tabla 3.6 se pueden observar los valores de los estadísticos, para cada una de las variables. En esta ocasión, de los 108 individuos tomados, solo 2 de ellos tienen datos incompletos, por tanto, se toman en la estimación 106

Variable	Datos completos	Mínimo	Máximo	Media	Desviación Típica
Y	108	0	1	0.50	0.502
L	108	1	2	1.46	0.501
C	108	0	1	0.50	0.502
M_1	107	1.9000	2.1300	2.0640	0.0371
M_2	108	1.4000	1.8800	1.6813	0.1047
M_3	108	1.1900	1.3600	1.2972	0.0351
M_4	108	1.1600	1.3400	1.2672	0.0378
M_5	107	1.5200	1.6800	1.6011	0.0347
M_6	108	1.5200	1.8400	1.7032	0.0900
C_1	108	0.0000	0.3190	0.0568	0.7206
C_2	108	0.0000	0.8861	0.3657	0.2377
C_3	108	0.0000	1.1071	0.1733	0.2624

Tabla 4.11. Estadísticos básicos

lagartos. Al igual que en la primera parte, tampoco se observan valores altos de la desviación típica, lo que puede indicar que los datos no poseen valores atípicos

La tabla 3.1 muestra los estadísticos descriptivos como la media, la desviación típica, el valor máximo y el valor mínimo. Cabe destacar también que de los 108 datos que posee la muestra, hay algunos registros incompletos, haciendo que solo se pueda trabajar con 75 reptiles de los 108 muestreados. Tampoco se observan altos valores de la desviación típica, lo que puede indicar que no poseen datos extremos.

También se realiza una normalización de las variables continuas (en esta ocasión, de todas menos Y, L y C) como en la primera parte.

Comparación de Modelos

Se adopta una estrategia similar a la primera parte, dado que el número de datos completos es $n = 106$ y el número de variables predictoras es $p = 11$, $\frac{106}{11} = 9.6364 < 40$, se toma la métrica de AIC_C como medida comparativa para encontrar el mejor modelo.

Ocurre una situación similar a la primera parte, donde el número de modelos posibles (es $2^{11} = 2048$) sigue siendo bastante elevado para compararlos uno a uno. Sin embargo, como interesa ver si la variable crema influye o no en el estudio, se realizan las agrupaciones una variante con la variable crema y otra sin. Las agrupaciones tratadas son las siguientes:

- *Nulo*: El modelo sin ninguna variable.
- *Completo*: El modelo con todas las variables.
- *Experto*: El modelo con las variables sugeridas por un experto en el tema y uno de los autores del artículo [3].
- *Experto + Crema*: El modelo con las variables sugeridas por el experto junto con la variable crema.

- *Hosmer-Lemeshow*: En la literatura, una propuesta que recomiendan estos autores es tomar solo las variables del modelo completo con significación menor que 0.25.
- *Morfológicas*: El modelo con solo las variables morfológicas.
- *Morfológicas + Crema*: El modelo con las variables morfológicas y la variable crema.
- *Comportamiento*: El modelo con solo las variables de comportamiento.
- *Comportamiento + crema*: El modelo con las variables de comportamiento y la variable crema.

Los resultados de las comparativas utilizando AIC_C y Δ_i se pueden observar en la tabla 3.7.

Modelo	AIC_C	Δ_i
Nulo	148.948	22.349
Completo	142.563	15.964
Experto	142.519	15.920
Experto + Crema	137.457	10.858
Hosmer-Lemeshow	126.599	0
Morfológicas	159.412	32.813
Morfológicas + crema	148.953	22.354
Comportamiento	134.665	8.066
Comportamiento + crema	128.451	1.852

Tabla 4.12. Alguna de las métricas

Analizando la tabla 3.7, se puede revisar que muchos de los modelos tienen mejor AIC_C, si se tiene en cuenta la variable Crema. Además, no se incluyó el conjunto de las variables solo significativas pues la única que lo era era la variable Crema, dando más indicaciones de que esta puede ser relevante en el análisis de la categoría de ganador-perdedor.

Por otra parte, el modelo con mejor valor de AIC_C es aquel con las variables con significación menor a 0.25, usando de nuevo la recomendación de Hosmer y Lemeshow.

El mejor modelo

El conjunto de variables consideradas para este modelo son aquellas con significación menor a 0.25 dentro del modelo completo. Estas variables son C (variable crema) y C_3.

En la tabla se pueden ver los resultados de las métrica, en la tabla la prueba Omnibus y en la tabla los valores estimados y la significación y los intervalos de confianza de Wald a nivel de confianza de 0.95.

Métrica	Valor
Devianza	13.073
Devianza escalada	19.143
Chi-cuadrado de Pearson	12.975
Max Log-Verosimilitud	-16.301
AIC	38.602
AIC_C	38.833
BIC	46.648

Tabla 4.13. Métricas de bondad de ajuste

Test	Significación
Omnibus	< 0.001

Tabla 4.14. Prueba de Omnibus

Variable	Valor estimado	Significación del test de Wald	Intervalo de Confianza
β_0	-0.699	0.007	[-1.206, -0.191]
β_2	1.234	0.001	[0.532, 1.936]
β_{11}	-0.911	0.000	[-1.391, -0.431]

Tabla 4.15. Test de Wald e intervalo de confianza al 0.95

De estas tablas se puede concluir que el mejor modelo estimado para la segunda parte del estudio es:

$$\eta_2 = \beta_0 + \beta_2 C + \beta_{11} C_3$$

Sustituyendo los valores de las estimaciones, se tiene el siguiente modelo estimado:

$$\eta_2 = -0.699 + 1.234C + -0.911C_3 \tag{3.2}$$

$$\hat{\mu} = \frac{e^{\eta_2}}{1 + e^{\eta_2}}$$

Diagnóstico de Hipótesis

La última parte del estudio es la verificación de las hipótesis de los Modelos Lineales Generalizados, para determinar posibles fallas en el modelo.

La distribución es correcta, pues al ser una variable binaria, la distribución no puede ser otra que una Bernoulli y se han normalizado las covariables para solucionar los posibles problemas de escala.

Vemos la independencia con la gráfica de los residuos de Pearson del dato i como eje x y los del dato $i - 1$ como eje y.

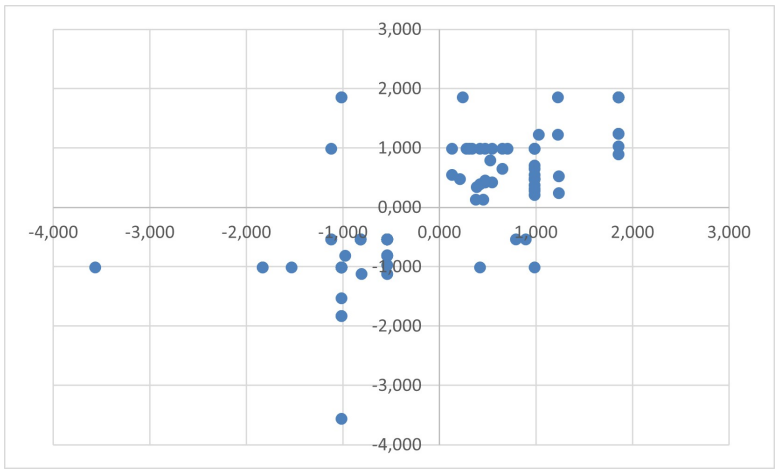

Figura 4.4. Gráfico de Independencia

En la figura 4.4 se puede observar una leve tendencia de los datos hacia arriba a la izquierda. Esto indica problemas en la independencia y se requiere de la aplicación de herramientas como las vistas en 3.2.1.

A continuación, se contrastan los residuos de devianza y los predictores lineares estimados para verificar que los datos son homocedásticos y que la función de enlace es correcta. En la figura 4.5 se pueden observar los resultados.

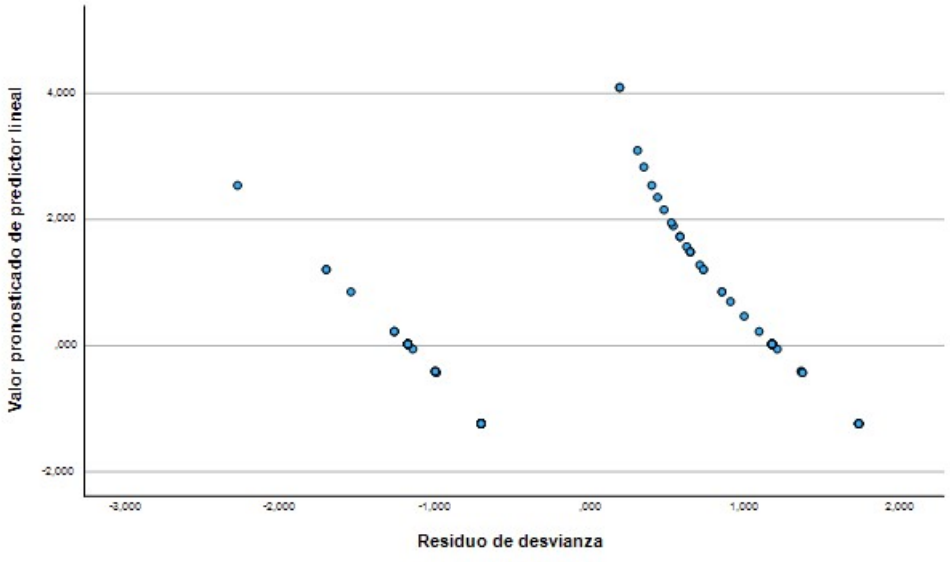

Figura 4.5. Residuos de devianza en función de los predictores lineales

En esta ocasión, en la figura 4.5 las curvas son casi paralelas, por tanto, pueden haber o problemas con la homocedasticidad o con la función de enlace.

Por último, se detectan los valores atípicos comparando la distancia de Cook con la influencia de las observaciones. En la figura 4.3 está representado.

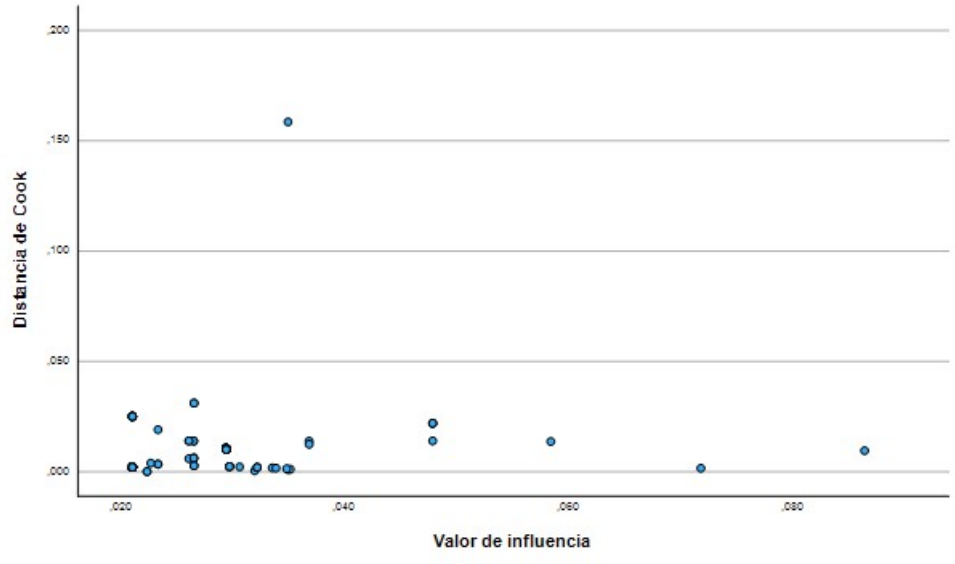

Figura 4.6. Importancia o influencia de los datos frente a la distancia de Cook

Sin embargo, se puede ver que no presenta datos anómalos. Aunque haya un dato con Distancia de Cook bastante elevada, no influye lo suficiente como para ser una observación atípica.

4.5. Resultados

De los análisis realizados en las secciones anteriores, se pueden concluir varios aspectos:

- Respecto de la parte I, se ha definido un modelo que verifica las hipótesis de los GLM y tiene buenas métricas de bondad de ajuste. Por tanto, el mejor conjunto de variables encontrado, que afecta significativamente a la variable Y es: M_2, M_4, C_2, C_3, C_4 y R_5. Utilizando las definiciones del inicio del capítulo, se puede decir que la victoria del reptil viene condicionada por la morfología (por la masa corporal y por la longitud de la extremidad anterior), por el comportamiento (por las frecuencias relativas de sacar la lengua, de mordidas y de sacudir la cola), y por la reflectancia (con la longitud de onda máxima).

- La segunda parte del estudio, el modelo con el mejor subconjunto de variables toma buenos valores en sus medidas de bondad de ajuste. Sin embargo, muchas de las hipótesis del modelo no las cumple. A pesar de no poder definir un modelo suficientemente bueno, viendo los modelos de predicción, se puede decir que al añadir la variable de la crema solar, las métricas mejoran considerablemente. Por tanto, esta variable condiciona bastante las victorias de los lagartos.

Conclusiones

En esta memoria de trabajo de fin de máster se ha llevado a cabo una aproximación a los Modelos Lineales Generalizados. Se han mostrado aspectos teóricos, como la estructura tripartita del modelo, las particularidades de cada componente, propiedades y funciones básicas, y un algoritmo que permite el cálculo aproximado de los estimadores de máxima verosimilitud de los parámetros.

Además, se han desarrollado diversas métricas que permiten la comparación entre modelos. Estas medidas son de diferentes tipos, como la Devianza que se basa en la log-verosimilitud, o el criterio de información de Akaike inspirado en la teoría de la información. En esta sección, también se indicaron algunos tests de hipótesis e intervalos de confianza útiles para poder verificar la significación de los parámetros del modelo.

Tras la elección del mejor conjunto de variables, se consiguió indicar algunos criterios y gráficas que permiten ver de manera intuitiva la veracidad de las hipótesis del modelo, indicando algunos artículos de la bibliografía para los lectores que busquen criterios más avanzados como tests de hipótesis y otras métricas que permitan un diagnóstico significativo.

Por último, se aplicaron con éxito los conceptos explicados a datos reales provenientes de un estudio importante del comportamiento de los reptiles. Era de especial interés identificar qué variables eran las que más influía en las victorias de los del mostrando los diferentes objetivos a cumplir, un modelo para cada objetivo y la aplicación de las metodologías y las conclusiones correspondientes.

En esta memoria, es muy difícil capturar todas las particularidades de los Modelos Lineales Generalizados (GLM). Actualmente, existen una gran variedad de algoritmos y de procedimientos para estimar los coeficientes del modelo y realizar inferencia en las estimaciones. Otras limitaciones en el estudio han sido: no tener en cuenta la presencia de sobredispersión en el conjunto de datos (suceso que altera los resultados de las estimaciones), tampoco se desarrollaron métodos de estimación del parámetro de dispersión del GLM (siempre se ha

asumido que se conoce) y sería esencial profundizar en un análisis teórico que explore la influencia del sobreajuste y el subajuste el los resultados del modelo.

A pesar de estas limitaciones, se ha conseguido mostrar una teoría bastante completa y aplicable en muchos de los casos reales donde se puedan utilizar este tipo de modelos.

Bibliografía

[1] Agresti, A. (2015). *Foundations of linear and generalized linear models*, John Wiley & Sons, Incorporated.

[2] Bani K. Mallick, Alan E. Gelfand (1994). *Generalized linear models with unknown link functions*, Biometrika, 81(2), pp 237–245. https://doi.org/10.1093/biomet/81.2.237

[3] Bohórquez Alonso, M.L., Mesa Avila, G., Suárez Rancel, M., Font, E. & Molina Borja, M. (2018). *Predictors of contest outcome in males of two subspecies of Gallotia galloti (Squamata: Lacertidae)*, Behavioral Ecology and Sociobiology, 72(3). https://doi.org/10.1007/s00265-018-2480-z

[4] Kenneth P. Burnham, David R. Anderson. (1998). *Model Selection and Inference*, Springer. https://doi.org/10.1007/978-1-4757-2917-7

[5] Coolen, A. C. C., Sheikh, M., Mozeika, A., Aguirre-Lopez, F., & Antenucci, F. (2020). *Replica analysis of overfitting in generalized linear regression models*, Journal of Physics A: Mathematical and Theoretical, 53(36). https://doi.org/10.48550/arXiv.2004.06329

[6] Diamond, I., Cox, D.R., & Snell, E.J. (1990). *Analysis of Binary Data*, Journal of the Royal Statistical Society. Series C (Applied Statistics), 39(2), pp 260–261. https://doi.org/10.2307/2347766

[7] Dey, D. K., Gelfand, A. E., & Peng, F. (1997). *Overdispersed generalized linear models*, Journal of Statistical Planning and Inference, 64(1), pp 93–107. https://doi.org/10.1016/s0378-3758(96)00207-8

[8] Dunn P. K., Smyth G. K. (2018). *Generalized Linear Models with Examples in R*, Springer Texts in Statistics, Springer Science Business Media, LLC. https://doi.org/10.1007/978-1-4419-0118-7_5

[9] Fang, K., & Zhang, Y. (1990). *Generalized multivariate analysis*, Science Press.

[10] Fuentes Fernández, M. D., Suárez Rancel, M., Quintana Gómez, P. D., & Molina-Borja, M. (2023). *Season, Body Condition, and Sex Variation of Ectoparasite Abundance in Tarentola delalandii (Squamata: Phyllodactylidae) from Two Ecologically Contrasting Populations of Tenerife (Ca-*

nary Islands), South American Journal of Herpetology, 26(1), pp 21-28. https://doi.org/10.2994/SAJH-D-20-00025.1

[11] Hardin, J.W., & Hilbe, J. (2001). *Generalized Linear Models and Extensions*, Stata Press.

[12] Huang, C. C. L., Jou, Y. J., & Cho, H. J. (2016). *A new multicollinearity diagnostic for generalized linear models*, Journal of Applied Statistics, 43(11), pp 2029–2043. https://doi.org/10.1080/02664763.2015.1126239

[13] IBM Corporation. (2022). *IBM SPSS Statistics*. Armonk, NY: IBM Corporation. https://www.ibm.com/es-es/spss.

[14] Liu, X. W., & Lu, D. G. (2018). *Survival analysis of fatigue data: Application of generalized linear models and hierarchical Bayesian model*, International Journal of Fatigue, 117, pp 39–46. https://doi.org/10.1016/j.ijfatigue.2018.07.027

[15] Mackinnon, M.J., & Puterman, M.L. (1989). *Collinearity in generalized linear models*, Communications in Statistics-theory and Methods, 18, pp 3463-3472. https://api.semanticscholar.org/CorpusID:120111535

[16] McCullagh, P. (1989). *Generalized Linear Models*, Routledge. 2nd ed. https://doi.org/10.1201/9780203753736

[17] Müller, M. (2012). *Generalized Linear Models*, In: Gentle, J., Härdle, W., Mori, Y. (eds) Handbook of Computational Statistics. Springer Handbooks of Computational Statistics. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-21551-3_24

[18] Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2010). *Generalized linear models : With applications in engineering and the sciences*, John Wiley & Sons, Incorporated.

[19] Nelder, J. A., & Wedderburn, R. W. M. (1972). *Generalized Linear Models*, Journal of the Royal Statistical Society. Series A (General), 135(3), pp 370–384. https://doi.org/10.2307/2344614

[20] Olsson, U. (2002). *Generalized Linear Models: An Applied Approach*, Studentlitteratur AB.

[21] Pedregosa F. et al. (2011). *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 12, pp 2825-2830 https://scikit-learn.org/

[22] Pregibon, D. (1980). *Goodness of Link Tests for Generalized Linear Models*, Journal of the Royal Statistical Society. Series C (Applied Statistics), 29(1), pp 15–14. https://doi.org/10.2307/2346405

[23] Puntanen, S. (2009). *Goodness of Link Tests for Generalized Linear Models*, 2nd ed., International Statistical Review. https://doi.org/10.1111/j.1751-5823.2009.00085_12.x

[24] R Core Team. (2019) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[25] Stark, P. (2006). *On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors"*, The American Statistician. https://www.academia.edu/58923132/On_The_So_Called_Huber_Sandwich_Estimator_and_Robust_Standard_Errors_?f_ri=85362

[26] Stephen M. Stigler. (1981). *"Gauss and the Invention of Least Squares"*, Ann. Statist. 9 (3), pp 465 - 474 https://doi.org/10.1214/aos/1176345451

[27] Weissfeld, L. A., & Sereika, S. M. (1991). *A multicollinearity diagnostic for generalized linear models*, Communications in Statistics - Theory and Methods, 20(4), pp 1183–1198. https://doi.org/10.1080/03610929108830558

[28] Fong, Y., Rue, H. & Wakefield, J. (2010). *Bayesian inference for generalized linear mixed models*, Biostatistics, 11 (3), pp 397–412. https://doi.org/10.1093/biostatistics/kxp053