Grado ...

TRABAJO FIN DE ESTUDIOS

Introducción a la Estadística Aplicada con ayuda de R

Marta García Moreno

Sevilla, Octubre de 2017

Índice general

Prólogo

Escrito colocado al comienzo de una obra en el que se hacen comentarios sobre la obra o su autor, o se introduce en su lectura; a menudo está realizado por una persona distinta del autor.

También se podrían incluir aquí los agradecimientos.

Resumen

Resumen. . .

Abstract

Abstract…

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El modelado estadístico es una herramienta fundamental para la investigación científica y el análisis de datos, la cual intenta aproximar la realidad a partir de la implementación de modelos matemáticos que tienen en cuenta la incertidumbre. Estos tipos de modelos son capaces de abarcar distintos problemas como pueden ser: la descripción de relaciones entre variables, la predicción de nuevos datos o la comprobación de hipótesis.

Hoy en día existen muchos métodos y técnicas para proceder a resolver los problemas antes mencionados, pero en este trabajo nos centraremos en el desarrollo de los Modelos Aditivos Generalizados (MAG). Sin embargo, hasta llegar a ellos, pasaremos por la descripción de los Modelos Lineales, los Modelos Lineales Generalizados (MLG) y los Modelos Aditivos. Esto se debe a que los MAG no son más que una extensión de los anteriores, así que haremos un transcurso desde modelos simples hasta un Modelo Aditivo Generalizado completo.

En el primer capítulo hablaremos de los Modelos Lineales y los Modelos Lineales Generalizados.El concepto de regresión lineal surgió a partir de una necesidad de estudiar la relación entre variables, de las cuales se conocen ciertos datos, mediante formalizaciones matemáticas. En concreto lo introdujo Francis Galton y luego fue desarrollado por el estadista y matemático Karl Pearson a finales del siglo XIX. Sin embargo, ya en el 1805 Legendre proponía la primera forma del método de mínimos cuadrados, por lo que estamoss hablando de técnicas que ya llevan más de tres siglos entre nosotros. A pesar de ello, no se desarrollan las nociones de los MLG hasta el 1970, estos modelos relajan las hipótesis que deben asumir los Modelos Lineales y permiten un primer acercamiento a que los modelos tengan un grado de no linealidad.

Comenzaremos el siguiente capítulo introduciendo los Modelos Aditivos y a partir de ellos nos adentraremos en los MAG y en los resultados necesarios para sus futuras aplicaciones, estudios y comparación. En particular, añadiremos una sección sobre las funciones de suavizado ('smoothers') y los tensores ('tensors') que aplicaremos luego en la práctica y también expondremos resultados sobre la comparación de estos tipos de modelos y sobre comprobación de hipótesis.

Los Modelos Aditivos Generalizados mejoran la metodología de los MLG incorporando la flexibilidad que aporta la regresión no paramétrica y mantienen la interpretabilidad de los datos del análisis de regresión con múltiples variables predictoras pues se modelan como una suma de términos 'suaves'. Actualmente, los GAM son un punto de partida magnífico para el modelado, un GAM bien ajustado debería funcionar de manera conveniente, incluso comparado con métodos de “boosting” o de “deep learning”. Además, el MAG tiene una base para una mejor interpretabilidad y métricas de incertidumbre más sencillas, por lo que en muchos casos del análisis de datos los Modelos Aditivos Generalizados son una buena opción.

Capítulo 2

Modelos lineales generalizados

2.1. Introducción

Como bien introdujimos antes, los modelos estadísticos pretenden explicar la relación entre dos o mas variables, en particular, tratan de describir el comportamiento de una variable respuesta (o dependiente), que se suele denotar por Y, mediante la información que otorgan las variables predictoras (o independientes), que se suelen denotar como $X_1, ..., X_p$.

La forma más general de expresar matemáticamente la intención de los modelos estadísticos es la siguiente:

$$Y = f(X_1, .., X_p) + \epsilon$$

Donde f es una función desconocida cuyo propósito es el de representar de la "mejor"[1] manera posible la relación entre las variables $X_1, ..., X_p$ e Y, y ϵ es un error aleatorio independiente de las variables predictoras que deberá cumplir ciertas condiciones según el tipo de modelo que estemos tratando.

Este trabajo tiene como finalidad el definir un conjunto de estrategias y técnicas que proporcionan un ajuste óptimo de la función f, es decir, que asemeje lo mejor posible la relación de las variables al fenómeno que se esté estudiando. Además, se incluye una posterior aplicación práctica de dichos resultados en el ámbito del cambio climático.

En lo que a este capítulo respecta, partiremos definiendo los modelos lineales de una forma breve y más general, ya que se ve de manera más extensa en varias asignaturas durante el grado. Tras ello daremos varios resultados básicos para el entendimiento y desarrollo de los Modelos Lineales Generalizados.

[1] Escribimos mejor entre comillas pues hay diversas formas de evaluar los modelos.

2.2. Modelos lineales

El modelo lineal ocupa un lugar clave en el manual de herramientas de todo estadístico aplicado. Esto se debe a su simple estructura, a la fácil interpretación de sus resultados y al sencillo desarrolo de la teoría de mínimos cuadrados. Sin embargo, a la hora de dar su definición se deben tener en cuenta ciertas restricciones que deben cumplir las variables y los errores del modelo. Estas condiciones hacen que el modelo no sea capaz de adaptarse bien a todos los fenómenos que uno se propone describir pero, a cambio, otorga esa sencillez de visualización antes mencionada.

Definición 2.2.1 (Modelo Lineal)**.** Sean $X_1, \ldots, X_p$ un conjunto de p vectores aleatorios de n componentes (con $p \leqslant n$) e $Y = (Y_1, \ldots, Y_n)^T$ un vector aleatorio de n componentes tal que $E[Y] = \mu$. Entonces, se entiende por modelo lineal (multivariante) aquel que determina la relación entre los vectores aleatorios mediante una combinación lineal de parámetros de la siguiente forma:

$$\mu = X\beta$$

$$Y = \mu + \epsilon$$

O vectorialmente como:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{(1,1)} & \cdots & x_{(1,j)} & \cdots & x_{(1,p)} \\ \cdots & \ddots & \vdots & \ddots & \cdots \\ x_{(i,1)} & \cdots & x_{(i,j)} & \cdots & x_{(i,p)} \\ \cdots & \ddots & \vdots & \ddots & \cdots \\ x_{(n,1)} & \cdots & x_{(n,j)} & \cdots & x_{(n,p)} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Donde:

- β es un vector de β_i parámetros a determinar que reflejan la magnitud del efecto lineal (constante) de los incrementos unitarios en las variables explicativas X_i sobre la variable explicada Y.
- ϵ es un vector aleatorio tal que los ϵ_i son variables aleatorias independientes e identicamente distribuidas por una distribución normal de esperanza nula y varianza σ^2 ($\epsilon_i \backsim N(0, \sigma^2)$). Representa el término de error del modelo y corresponde con $\epsilon = Y - E[Y]$.

Observación 2.2.1. Gracias a lo notado en el párrafo anterior podemos ver que: $Y \backsim N(\mu, \sigma^2 I_n)$ ya que:

$$E[Y] = E[X\beta + \epsilon] = X\beta + E[\epsilon] = X\beta = \mu$$

$$Cov(Y) = Cov(X\beta + \epsilon) = Cov(\epsilon) = \sigma^2 I_n$$

Veamos ahora cómo se pueden obtener los valores de los parámetros β.

Definición 2.2.2 (Estimador por mínimos cuadrados)**.** Elegiremos los valores de β que minimicen la suma de cuadrados:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \mu_i)^2 = \sum_{i=1}^{n} (Y_i - (X\beta)_i)^2 \tag{2.1}$$

Es decir:

$$\hat{\beta} = arg_{\beta \in \mathbb{R}^p} min ||Y - X\beta||^2$$

Donde $|| \cdot ||$ denota la norma euclídea. Para encontrar tal mínimo se razona derivando respecto de cada β_i y luego igualando a 0. De este modo los parámetros β_i vienen dados por la solución del siguiente sistema:

$$\begin{cases} \frac{\partial S}{\partial \beta_1} = 0 \\ \dots \\ \frac{\partial S}{\partial \beta_p} = 0 \end{cases}$$

Desarrollando este sistema de escuaciones llegamos a que es equivalente a $X^T X \hat{\beta} = X^T Y$, por lo que el estimador por mínimos cuadrados del vector de parámetros β viene dado por:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Observación 2.2.2. Notemos que esta expresión tiene sentido pues el producto $X^T X$ resulta en una matriz cuadrada de orden p y rango máximo.

Proposición 2.2.1 (Distribución del estimador por mínimos cuadrados)**.** *El estimador por mínimos cuadrados del vector de parámetros β, $\hat{\beta}$, sigue una distribución del tipo normal p-variante de esperanza β y matriz de covarianzas $V_{\hat{\beta}} = \sigma^2 (X^T X)^{-1}$. Es decir: $\hat{\beta} \backsim N_p(\beta, \sigma^2 (X^T X)^{-1})$*

Demostración. Partiremos viendo que el estimador $\hat{\beta}$ es insesgado:

$$E[\hat{\beta}] = E[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T E[Y] = (X^T X)^{-1} X^T X \beta = \beta$$

Por otro lado, calculemos la matriz de covarianzas de $\hat{\beta}$, para ello primero debemos notar que como los errores aleatorios ϵ_i son independientes e identicamentes distribuidos con esperanza nula y varianza σ^2, $\forall i \neq j$:

$$E[\epsilon_i \epsilon_j] = E[\epsilon_i] + E[\epsilon_j] + Cov(\epsilon_i, \epsilon_j) = 0$$

y, por tanto:

$$E[\epsilon \epsilon^T] = E \begin{bmatrix} \epsilon_1^2 & \epsilon_1\epsilon_2 & \cdots & \epsilon_1\epsilon_n \\ \epsilon_1\epsilon_2 & \epsilon_2^2 & \cdots & \epsilon_2\epsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_1\epsilon_n & \cdots & \cdots & \epsilon_n^2 \end{bmatrix} = \begin{pmatrix} E[\epsilon_1^2] & E[\epsilon_1\epsilon_2] & \cdots & E[\epsilon_1\epsilon_n] \\ E[\epsilon_1\epsilon_2] & E[\epsilon_2^2] & \cdots & E[\epsilon_2\epsilon_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[\epsilon_1\epsilon_n] & \cdots & \cdots & E[\epsilon_n^2] \end{pmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{bmatrix}$$

Luego, utilizando que $\hat{\beta} = \beta + (X^TX)^{-1}X\epsilon$, se tiene que:

$$\begin{aligned} Cov(\hat{\beta}) &= E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^T] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = \\ &= E[((X^TX)^{-1}X^T\epsilon)((X^TX)^{-1}X^T\epsilon)^T] = E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-T}] = \\ &= (X^TX)^{-1}X^TE[\epsilon\epsilon^T]X(X^TX)^{-T} = (X^TX)^{-1}X^T\sigma^2I_nX(X^TX)^{-T} = \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-T} = \sigma^2(X^TX)^{-1} \end{aligned}$$

Acabamos la prueba recordando que $\hat{\beta}$ no es más que una combinación lineal de variables aleatorias normales, concretamente de las Y_i, para decir que en efecto sigue una distribución normal p-variante como indica el enunciado. ■

Observación 2.2.3. Ahora bien, generalmente el valor de σ^2 es desconocido así que también sería preciso dar una estimación del mismo para que así los resultados anteriores fueran de alguna utilidad.

Definición 2.2.3 (Estimador de σ^2)**.** La varianza σ^2 admite un estimador insesgado que se basa en la suma de cuadrados:

$$\hat{\sigma}^2 = \frac{S}{n-p} = \frac{\sum_{i=1}^{n}(Y_i - (X\beta)_i)^2}{n-p}$$

Además se tiene que $\hat{\sigma}^2$ sigue una distribución:

$$\hat{\sigma}^2 \backsim \frac{\sigma^2}{n-p}\chi^2_{n-p}$$

Observación 2.2.4. No daremos la obtención de este estimador pero sí indicaremos que su distribución se obtiene directamente de que como S es la suma de normales $N(0, \sigma^2)$ y $\frac{\epsilon_i}{\sigma} \backsim N(0,1)$, entonces: $\frac{1}{\sigma^2}S \backsim \chi^2_{n-p}$. Finalmente, multiplicando y dividiento esta expresión por (n-p) y sustituyendo la definición del estimador $\hat{\sigma}^2$ obtenemos el resultado.

2.3. Modelos lineales generalizados

Nos adentramos ya con esta sección en la primera extensión de los modelos lineales de las que se tratan durante el trabajo. Los Modelos Lineales Generalizados (MLG) fueron originalmente formulados por John Nelder y Robert Wedderburn (1972), quienes tenían como propósito unificar varios modelos estadísticos como la regresión lineal, la logística y la de Poisson en un mismo modelo. Este tipo de modelos relajan algunas de las condiciones de los modelos lineales, como que los errores ya no deben seguir ninguna distribución específica, y además añade nuevos elementos como la función de enlace, que interviene en la relación entre las medias y la forma lineal del modelo. También se deberá tener en cuenta que seguirá la hipótesis de que las variables de respuestas siguen distribuciones de tipo exponencial. Más tarde nos adentraremos en cada uno de estos aspectos, de momento demos la estructura básica de un MLG.

Definición 2.3.1 (Estructura básica de un MLG).

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{pmatrix} = E[Y]$$

$$g(\mu_i) = X_i\beta, \forall i = 1, \ldots, n$$

Donde:

- X es la matriz modelo de dimensión $n \times p$ con $p \leqslant n$, que contiene a las variables predictoras, cada columna representa una variable predictora X_i.
- $\beta = (\beta_1, \ldots, \beta_p)^T$ es el vector de parámetros desconocidos como en el caso de los modelos lineales. A $\eta = X\beta$ se le conoce como predictor lineal.
- $g : \mathbb{R} \to \mathbb{R}$ es la función de enlace, que debe ser diferenciable y monótona. Representa la relación entre la media de las variables de respuesta y el predictor lineal.
- $Y = (Y_1, \ldots, Y_n)^T$ es un vector aleatorio, se suele suponer que las Y_i son variables aleatorias independientes y que siguen una distribución de tipo exponencial.

Observación 2.3.1. Desde esta formulación podemos ver fácilmente el por qué decimos que los MLG son una generalización de los modelos lineales, ya que basta tomar a la identidad como la función de enlace y suponer que la distribución considerada sea de tipo normal para encontrarnos ante la forma general de un modelo lineal como vimos en la sección anterior.

2.3.1. Familia de distribuciones exponenciales

Como hemos mencionado antes, la variable de respuesta de los modelos lineales generalizados deben seguir una distribución de tipo exponencial, en esta sección veremos qué significa eso y qué implicaciones tiene. Uno de los motivos más importantes por los que se supone que las variables de respuesta Y_i siguen distribuciones de esta familia se debe a que en los modelos lineales los cambios constantes en las variables predictoras implicaban cambios constantes en la variable de respuesta, pero ahora se quiere permitir que dichos cambios constantes de entrada puedan implicar también variaciones geométricas.

Definición 2.3.2 (Distribución de tipo exponencial). Una distribución se dice que es de tipo exponencial si su función de densidad es de la forma:

$$f_\theta(y) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$$

Donde:

- a, b y c son funciones arbitrarias.
- ϕ es conocido como parámetro de escala.

- θ es conocido como parámetro canónico de la distribución. Más adelante veremos que depende completamente de los parámetros del modelo β.

Ejemplo 2.3.1. La distribución normal es de tipo exponencial pues su función de densidad es:

$$f_\mu(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = e^{\frac{-y^2+2y\mu-mu^2}{2\sigma^2} - log(\sigma\sqrt{2\pi})} = e^{\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - log(\sigma\sqrt{2\pi})}$$

Y por tanto toma los siguientes parámetros de la familia exponencial:

- $\theta = \mu$
- $b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$
- $a(\phi) = \phi = \sigma^2$
- $c(\phi, y) = \frac{-y^2}{2\phi} - log(\sqrt{\phi 2\pi}) = \frac{-y^2}{2\sigma} - log(\sigma\sqrt{2\pi})$

Es posible dar una forma general para la esperanza y la varianza de las variables de tipo exponencial dependiendo de los parámetros de su función de densidad. Lo vemos en el siguiente resultado.

Proposición 2.3.1. *Sea Y una variable de tipo exponencial, entonces verifica:*

- $E[Y] = b'(\theta)$
- $Var(Y) = b''(\theta)a(\phi)$

Demostración. Partimos considerando la función de verosimilitud logáritmica para θ:

$$l(\theta) = log(f_\theta(y)) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

Y su derivada respecto de theta:

$$\frac{\partial l}{\partial \theta}(\theta) = \frac{y - b'(\theta)}{a(\phi)}$$

Ahora bien, si cambiamos la observación y por la variable Y, podemos evaluar la esperanza de esta derivada, la cual será 0 por propiedades de la función de verosimilitud logáritmica.

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

Y de aquí se obtiene directamente que $E[Y] = b'(\theta)$. Seguimos derivando para obtener que:

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

Y utilizando que: $E[\frac{\partial^2 l}{\partial \theta^2}] = -E[(\frac{\partial l}{\partial \theta})^2]$, nos queda que:

$$\frac{b''(\theta)}{a(\phi)} = \frac{E[(Y - b'(\theta))^2]}{a(\phi)^2} = \frac{E[(Y - E[Y])^2]}{a(\phi)^2} = \frac{Var(Y)}{a(\phi)^2}$$

De donde se obtiene que: $Var(Y) = b''(\theta)a(\phi)$. ■

Observación 2.3.2. Cuando ϕ es conocido el manejo de la función a no tiene dificultad, pero en muchos casos ϕ suele ser desconocido, así que para agilizar los resultados escribiremos $a(\phi) = \frac{\phi}{\omega}$, donde ω es una constante conocida. De hecho, todos los casos prácticos de interés se podrán expresar así y la mayoría con $\omega = 1$. De este modo nos queda: $Var(Y) = b''(\theta)\frac{\phi}{\omega}$.

Podemos recoger las carácterísticas de las principales distribuciones de tipo exponencial en la siguiente tabla:

	Binomial $Bi(n,p)$	Normal $N(\mu,\sigma^2)$	Poisson $Po(\lambda)$	Gamma $Ga(p,\lambda)$
$\theta(\mu)$	$log(\frac{\mu}{n-\mu}$	μ	$log(\mu)$	$-\frac{1}{\mu}$
ϕ	1	σ^2	1	$\frac{1}{p}$
$a(\phi)$	1	σ^2	1	$\frac{1}{p}$
$b(\theta)$	$nlog(1+e^\theta)$	$\frac{\theta^2}{2}$	e^θ	$-log(-\theta)$
$c(y,\phi)$	$log(\binom{n}{y})$	$-\frac{1}{2}(\frac{y^2}{\phi} + log(2\pi\phi))$	$-log(y!)$	$plog(py) - log(y\Gamma(p))$

2.3.2. Ajuste de los modelos lineales generalizados

La estimación de parámetros e inferencia con modelos aditivos generalizados se basa en la estimación de máxima verosimilitud, ya que, gracias a la hipótesis de que las Y_i pertenecezcan a la familia de distribuciones exponenciales, siempre se dispondrá de funciones de densidad. Aún así, como veremos más adelante, se deberá recurrir a un algoritmo basado en mínimos cuadrados para hallar tal máximo. En esta sección estudiaremos resultados basados en la teoría de verosimilitud para los modelos lineales generalizados.

Partiremos considerando un MLG como el de la definición 2.3.1

Capítulo 3

Título del Capítulo

3.1. Primera sección

Capítulo 4

Título del Capítulo

4.1. Primera sección

Apéndice A

Apéndice: Título del Apéndice

A.1. Primera sección

Apéndice B

Apéndice: Título del Apéndice

B.1. Primera sección

Bibliografía

JJ Allaire, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2023. URL https://CRAN.R-project.org/package=rmarkdown. R package version 2.21.

Pedro L. Luque-Calvo. *Escribir un Trabajo Fin de Estudios con R Markdown*, 2017.

Pedro L. Luque-Calvo. *Cómo crear Tablas de información en R Markdown*, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

RStudio Team. *RStudio: Integrated Development Environment for R.* RStudio, Inc., Boston, MA, 2015. URL http://www.rstudio.com/.

Techopedia. "definition - what does business intelligence (bi) mean?". Disponible en https://www.techopedia.com/definition/345/business-intelligence-bi, 2017.

Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2023a. URL https://CRAN.R-project.org/package=ggplot2. R package version 3.4.4.

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023b. URL https://CRAN.R-project.org/package=dplyr. R package version 1.1.1.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2023. URL https://yihui.org/knitr/. R package version 1.42.