UNIVERSIDAD DE EXTREMADURA

FACULTAD DE CIENCIAS

Grado en Estadística

Memoria de trabajo fin de grado

MODELOS LINEALES GENERALIZADOS Y ADITIVOS GENERALIZADOS

Francisco José Clemente García

Junio, 2022

D. Miguel González Velasco, profesor del Departamento de Matemáticas de la Universidad de Extremadura, y

D. Manuel Mota Medina, profesor del Departamento de Matemáticas de la Universidad de Extremadura,

INFORMAN:

Que D. Francisco José Clemente García ha realizado bajo su dirección el Trabajo Fin de Grado y que la memoria reúne los requisitos necesarios para su evaluación.

En Badajoz, 9 de junio de 2022

Fdo.: *Miguel González Velasco* y *Manuel Mota Medina*

Índice general

1. Introducción **11**

2. Modelo lineal **13**

2.1. Definiciones 13

2.1.1. Modelo lineal normal 14

2.2. Modelo lineal de rango completo 15

2.2.1. Estimación 15

2.2.2. Contraste de hipótesis 19

2.3. Modelo de regresión lineal normal 20

2.4. Ejemplo 21

3. Modelo lineal generalizado **29**

3.1. La familia exponencial de distribuciones 29

3.1.1. Definiciones y resultados 30

3.1.2. Distribuciones con parámetros de localización y dispersión 32

3.2. Modelo lineal generalizado 36

3.3. Estimación 37

3.3.1. Métodos numéricos 40

3.3.2. Comportamiento asintótico 40

3.4. Contraste de hipótesis 42

3.4.1. Bondad de ajuste 42

3.4.2. Contraste de hipótesis sobre β 44

3.5. Modelos de regresión 45

3.5.1. Regresión normal (varianza común conocida) 46

3.5.2. Regresión logística binaria 46

3.5.3. Regresión de Poisson 47

3.5.4. Regresión exponencial 48

3.6. Ejemplo 49

4. Modelo aditivo generalizado **53**

4.1. Modelo aditivo 53

4.1.1. Suavizado 53

4.1.2. Definición 60

4.1.3. Estimación 60

4.1.4. Ejemplo 61

4.2. Modelo aditivo generalizado 67

4.2.1. Definición 67

4.2.2. Estimación 68

4.2.3. Ejemplo 69

A. Otros resultados **73**

A.1. Resultados de Análisis Matemático 73

A.1.1. Método de Newton-Raphson 73

A.1.2. Diferenciación bajo el signo integral 74

A.2. Conceptos de Álgebra Lineal 75

A.3. Distribuciones de probabilidad 75

A.3.1. Distribuciones absolutamente continuas 75

A.3.2. Distribuciones discretas 78

A.4. Resultados de estadística 80

A.4.1. Estimación por máxima verosimilitud 80

A.5. Suavizado 82

A.5.1. *Splines* cúbicos naturales 82

A.5.2. Regresión local: LOESS 83

A.6. *Software* `R` 84

Bibliografía **85**

Resumen

Este trabajo describe una serie de técnicas para el estudio de modelos estadísticos que no satisfacen las hipótesis del modelo lineal normal. El modelo lineal generalizado (MLG) nos permite estimar y contrastar hipótesis en modelos cuya variable respuesta tiene distribución de probabilidad perteneciente a la familia exponencial de distribuciones. La esperanza de dicha variable respuesta está ligada a un predictor lineal, función matricial del parámetro del modelo, mediante una función de enlace que satisface unas condiciones de regularidad adecuadas. El modelo aditivo generalizado da flexibilidad al MLG incluyendo técnicas de regresión no paramétrica.

Abstract

This essay describes an array of techniques to understand statistical models asside of normal linear model hypotheses. Generalized linear model (GLM) allows to estimate and test hypothesis in models whose response variable follows some exponential family distribution. Expectation in responses are linked to a linear predictor, a matrix transformation of the model parameter, via some link function satisfaying enough regularity contions . Generalized additive model gives flexibility to GLM adding nonparametric regression techniques.

Capítulo 1

Introducción

Uno de los pilares de la Estadística como disciplina es la modelización de una variable respuesta (univariante o multivariante) a partir de un conjunto de datos prefijados u observados aleatoriamente de otras variables conmumente denominadas *predictoras*.

Históricamente, dada su sencillez y las posibilidades de cálculo del momento surge el modelo lineal (o modelo lineal general) asentado sobre nociones conocidas del álgebra lineal y de teoría de la probabilidad y la estadística. Las hipótesis del modelo, independencia en las observaciones y relación de tipo lineal entre la variable respuesta y las predictoras, son poco exigentes y generalmente han venido acompañadas de una distribución normal en la respuesta. La distribución normal es una de las distribuciones de probabilidad más estudiadas desde hace siglos, observada en multitud de fenómenos de tipo biológico, variables antropométricas, teoría de errores en experimentación física, concentración de compuestos bioquímicos en sangre, etc. Esta hipótesis adicional da lugar al modelo lineal normal. El trabajo de R. Fisher en la agricultura profundizó el conocimiento estadístico sobre esta familia de modelos.

Con el tiempo, la teoría de la probabilidad avanzó en el estudio sobre distintas familias de distribuciones, particularmente en la que se ha dado a conocer como familia exponencial, al tiempo que se aceleran las posibilidades de calculo numérico gracias a la computación. En este contexto, se pretende hallar modelos que ensanchen los límites del modelo lineal normal, abriéndose a otras posibles distribuciones en la variable respuesta. Nelder y Wedderburn presentan en 1976 [10] el modelo lineal generalizado, que toma como hipótesis una buena adecuación de la respuesta a la familia de distribuciones exponenciales, proporcionando así un modelo que recoge perfectamente al modelo lineal normal

pero que también son válidos procesos de contaje, tiempos entre eventos o variables de respuesta binaria. La estimación, por máxima verosimilitud, no es inmediata como lo era en el modelo normal, pero se introducen en el modelo conceptos como el de devianza y métodos iterativos de búsqueda de estimadores máximo verosímiles para contrastar la buena adecuación de los datos y la optimalidad de las estimaciones.

Paralelamente, todo un estudio sobre regresión no paramétrica y suavizado concluye en los modelos aditivos como intento de superar la linealidad anteriormente citada. Para ello, se articula toda una teoría sobre funciones de suavizado, que escapa al temario del Grado en Estadística, y a la que recurriremos auxiliarmente para llegar al punto final de este trabajo: el modelo aditivo generalizado formulado por Hastie y Tibshirani en 1986 [6]. Este modelo abarcará al lineal generalizado así como este lo hizo con el lineal normal, apoyado sobre la aditividad entre funciones de suavizado, que operan, cada una de ellas, sobre las distintas variables explicativas del modelo.

Teniendo en cuenta todo lo anterior, este trabajo partirá de algunas ideas presentadas no del todo conexas entre sí a lo largo del Grado en Estadística (modelo lineal normal, teoría asintótica de distribuciones, familia exponencial, interpolación polinómica o suavizado mediante *splines* cúbicos naturales) para construir una teoría de modelización estadística muy moderna y flexible, de notable importancia en nuestros días, cuyo objetivo fundamental es el análisis de datos procedentes de de muy diversas fuentes (ciencias ambientales, economía financiera, salud, etc.) que no pueden ser ajustados bajo hipótesis restrictivas como normalidad y linealidad.

Notación. A fin de facilitar la lectura de este trabajo, consideraremos todos los vectores serán, en su misma notación, matrices columna. De manera general, si $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ y A es una matriz de $m \times n$, escribiremos un sistema lineal de ecuaciones $Ax = b$ y si e y v son dos vectores de $\mathbb{R}^n$, podremos escribir $e^{\,\mathrm{t}}\, v = \sum_{i=1}^{n} e_i v_i$.

Sobre este trabajo. Como nota final, destacamos que este documento ha sido redactado en su totalidad con LaTeX, combinado en los ejemplos con el *software* estadístico R mediante el paquete `kntir` (ver [17]). Los gráficos mostrados son generados por R a través del paquete `ggplot2` (ver [15]).

Capítulo 2

Modelo lineal

El modelo lineal con el que iniciamos este trabajo es una de las formas más generales de modelización estadística. Tiene como únicos requisitos la independencia de las observaciones (componentes de un vector respuesta) y la existencia de un vector de menor dimensión al número de observaciones que ajusta linealmente (matricialmente) a las medias del vector respuesta. Como se verá, la generalidad de tener un modelo donde las variables no siguen ninguna distribución de probabilidad específica nos privará de desarrollar el problema de estimación con intervalos de confianza y el de contraste de hipótesis, carencia que se verá resuelta con la imposición de la distribución normal. Todos los resultados aquí presentados (y más) han sido ya estudiados con detalle en la asignatura Modelos Lineales del Grado en Estadística. Por esta razón, este capítulo será más esquemático y breve que los dos siguientes, teniendo por objetivo recordar los resultados más relevantes del modelo y dotar de sentido a los siguientes capítulos. Además del modelo de regresión lineal, que veremos hacia el final del capítulo, el modelo lineal contiene toda la teoría de modelos de diseño de experimentos (completamente aleatorizados, cuadrados latinos, modelos bifactoriales, etc.) que, por las limitaciones de este trabajo hemos optado por no desarrollar.

2.1. Definiciones

Definición 2.1 (Modelo lineal). Sean $Y = (Y_1, \ldots, Y_n)^{\mathrm{t}}$ un vector aleatorio de dimensión n, X una matriz de orden $n \times p$, con $p \leq n$, de constantes reales conocidas y $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{t}} \in \mathbb{R}^p$, un vector de parámetros desconocidos. Diremos que Y verifica las

condiciones de un modelo lineal si para cada coordenada Y_i, $1 \leq i \leq n$:

1. Las Y_i son variables aleatorias, incorreladas y homocedásticas con varianza común σ^2, con $\sigma^2 > 0$.

2. $\mathrm{E}[Y_i] = \mu_i = {x_i}^{\mathrm{t}}\beta$, siendo ${x_i}^{\mathrm{t}}$ la i-ésima fila de la matriz X, $1 \leq i \leq n$.

Por tanto, podremos escribir el modelo lineal, dada la condición 2 de su definición del siguiente modo:

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,j} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix} \tag{2.1}$$

Podemos definir un vector aleatorio $\varepsilon = Y - \mathrm{E}[Y]$ no observable (al no conocer $\mathrm{E}[Y]$) de dimensión n y media nula, $\mathrm{E}[\varepsilon] = \mathrm{E}[Y - \mathrm{E}[Y]] = 0_n \in \mathbb{R}^n$, de modo que el modelo queda expresado como $Y = X\beta + \varepsilon$.

La expresión $Y = X\beta + \varepsilon$ del modelo hace evidente que este consta de una componente determinística, $X\beta$, y otra aleatoria, ε. Se tiene por tanto que el vector de medias y la matriz de covarianzas de Y, a partir de esta formulación del modelo, son $\mathrm{E}[Y] = \mathrm{E}[X\beta + \varepsilon] = X\beta + \mathrm{E}[\varepsilon] = X\beta$ y $\mathrm{Cov}[Y] = \mathrm{Cov}[X\beta + \varepsilon] = \mathrm{Cov}[\varepsilon] = \sigma^2 I_n$. La incorrelación y la homocedasticidad de las variables aleatorias ε_i e Y_i queda reflejada en la expresión diagonal de la matriz de covarianzas $\sigma^2 I_n$; $\mathrm{cov}[\varepsilon_i, \varepsilon_l] = \mathrm{cov}[Y_i, Y_l] = \sigma^2 \delta_{i,l}$, $1 \leq i, l \leq n$.

Definición 2.2 (Modelo de rango completo e incompleto). Supongamos que Y se ajusta a un modelo lineal con $\mathrm{E}[Y] = X\beta$ según la Definición 2.1. Si $\mathrm{rg}(X) = p$, diremos que el modelo es de rango completo. En caso $\mathrm{rg}(X) < p$, es de rango incompleto.

2.1.1. Modelo lineal normal

Un caso particular de modelo lineal es el modelo lineal normal, en el que la variable respuesta sigue una distribución normal multivariante, la cual denotamos $Y \sim \mathrm{N}_n(X\beta, \sigma^2 I_n)$ (ver Definición A.7 del Apéndice). La incorrelación entre las variables aleatorias coordenadas de Y se traduce en independencia y su homocedasticidad nos conducen a hablar de distribución multivariante normal esférica, $\mathrm{Cov}[Y] = \sigma^2 I_n$.

DEFINICIÓN 2.3 (Modelo lineal normal). Diremos que Y se ajusta a un modelo lineal normal si $Y_1, \ldots, Y_n$ son independientes y cada $Y_i \sim \mathrm{N}\left(x_i{}^{\mathrm{t}}\,\beta, \sigma^2\right)$ siendo $x_i{}^{\mathrm{t}}$ la i-ésima fila de X.

Luego el vector aleatorio Y puede ser expresado como $Y = X\beta + \varepsilon$ con $\varepsilon \sim \mathrm{N}_n\left(0, \sigma^2 I_n\right)$.

La condición de homocedasticidad puede no estar garantizada en muchos casos, siendo cada coordenada de Y tal que $Y_i \sim \mathrm{N}\left(x_i{}^{\mathrm{t}}\,\beta, \sigma_i^2\right)$. Mediante las transformaciones $Y_i^* = \sigma_i^{-1} Y_i$ se tiene el vector aleatorio

$$Y^* = \left(Y_1^*, \ldots, Y_n^*\right)^{\mathrm{t}} \sim \mathrm{N}_n\left(\Sigma^{-\frac{1}{2}} X\beta, I_n\right) \tag{2.2}$$

siendo $\Sigma^{-\frac{1}{2}}$ la matriz diagonal de σ_i^{-1}, $1 \leq i \leq n$, que sí es un modelo lineal normal como el de la Definición 2.3.

2.2. Modelo lineal de rango completo

2.2.1. Estimación

Definido el modelo lineal, queremos encontrar estimadores del vector de parámetros β verificando ciertas condiciones que nos garanticen que los estimadores tienen buenas propiedades en cuanto a linealidad, insesgadez o varianza mínima.

Si no asumimos ningún modelo de probabilidad concreto para ε, podemos abordar la estimación de β mediante el método de mínimos cuadrados ordinarios. Este consistirá en encontrar un vector $\hat{\beta}$ que minimice la distancia entre Y y su vector de medias:

$$\hat{\beta} = \arg_{\beta \in \mathbb{R}^p} \min \left|\left|Y - X\beta\right|\right|_2^2, \tag{2.3}$$

donde $||x||_2^2 = \sum_{i=1}^n x_i^2$ para cualquier $x \in \mathbb{R}^n$.

Para un estudio más detallado de la geometría del modelo lineal en el espacio vectorial $\mathbb{R}^n$ y qué representa su solución mínimo cuadrática, se remite al lector a Montanero (2008) [9].

Para encontrar el vector donde se encuentra el mínimo de $||Y - X\beta||_2^2$, definimos la función $S : \beta \in \mathbb{R}^p \to S(\beta) = ||Y - X\beta||_2^2 = \sum_{i=1}^n \varepsilon_i^2 \in \mathbb{R}$ y resolvemos el sistema de

ecuaciones

$$\begin{cases} \frac{\partial S(\hat{\beta})}{\partial \beta_1} = & 0 \\ \dots & \\ \frac{\partial S(\hat{\beta})}{\partial \beta_p} = & 0 \end{cases} \tag{2.4}$$

Es fácil probar que el sistema de ecuaciones (2.4) es equivalente al sistema de ecuaciones lineales $X^{\,t}X\hat{\beta} = X^{\,t}Y$, conocido como sistema de ecuaciones normales del modelo. Teniendo en cuenta que $X^{\,t}X$ es una matriz cuadrada de orden p y que $\operatorname{rg}(X^{\,t}X) = p$ (y por tanto invertible), es inmediato obtener que

$$\hat{\beta} = (X^{\,t}X)^{-1}X^{\,t}Y \tag{2.5}$$

es la única solución de la ecuación (2.4). Así, $\hat{\beta}$ es el estimador de mínimos cuadrados de β.

Para probar que $S(\beta)$ alcanza el mínimo en $\hat{\beta}$ así definido, bastará tener en cuenta que para todo $\beta \in \mathbb{R}^p$, se verifica $||Y - X\beta||_2^2 = \left|\left|Y - X\hat{\beta}\right|\right|_2^2 + \left|\left|X\hat{\beta} - X\beta\right|\right|_2^2$.

Observación (Notación). En lo que sigue de este capítulo, escribiremos abreviadamente en ocasiones $S = X^{\,t}X$ y $X^+ = (X^{\,t}X)^{-1}X^{\,t}$, de modo que

$$\hat{\beta} = (X^{\,t}X)^{-1}X^{\,t}Y = S^{-1}X^{\,t}Y = X^+Y \tag{2.6}$$

Notemos que $\hat{\beta}$ es un estimador lineal al ser una transformación lineal de Y. Además es insesgado, pues $\mathrm{E}\left[\hat{\beta}\right] = \beta$. Finalmente, señalamos que $\mathrm{Cov}\left[\hat{\beta}\right] = \sigma^2(X^{\,t}X)^{-1}$.

Definición 2.4 (Vector de predicciones). Sea $Y = X\beta + \varepsilon$ un modelo lineal de rango completo y sea $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^{\,t} \in \mathbb{R}^p$ una estimación de β, llamamos vector de predicciones a $\hat{Y} = X\hat{\beta}$.

Definición 2.5 (Vector de residuos). Definimos $e = Y - \hat{Y}$ como el vector de residuos del modelo.

En general, la varianza, σ^2, del modelo será desconocida y tendremos que obtener un estimador suyo. Se tiene que

$$\tilde{\sigma}^2 = \frac{1}{n-p}\left|\left|Y - X\hat{\beta}\right|\right|_2^2 \tag{2.7}$$

es un estimador insesgado de σ^2.

TEOREMA 2.1 (Gauss-Markov). *Supongamos que Y se ajusta a un modelo lineal de rango completo,* $\mathrm{rg}\,(X) = p$. *El mejor estimador (en el sentido de los mínimos cuadrados) lineal insesgado de β es $\hat{\beta} = X^{+}Y$. Además, cada $\hat{\beta}_j$ es el estimador lineal insesgado de mínima varianza de β_j, $1 \leq j \leq p$.*

TEOREMA 2.2. *Supongamos que Y se ajusta a un modelo lineal de rango completo,* $\mathrm{rg}\,(X) = p$. *Sea $\lambda \in \mathbb{R}^p$ un vector de constantes conocidas. Entonces el estimador lineal insesgado de mínima varianza de $\lambda^{\mathrm{t}}\,\beta$ es $\lambda^{\mathrm{t}}\,X^{+}Y$.*

Modelo lineal normal

El modelo lineal normal es un caso particular del modelo lineal sobre el que hemos desarrollado la teoría de estimación mediante mínimos cuadrados al estimar β. En consecuencia, $\hat{\beta}$ y $\tilde{\sigma}^2$ son estimadores de estos parámetros en el modelo lineal normal. Es inmediato cuestionarse si la estimación por el método de máxima verosimilitud en el caso del modelo lineal normal conduce a estos mismos estimadores para β y σ^2.

Si Y verifica las condiciones de un modelo lineal normal de media $\mu = \mathrm{E}\,[Y] = X\beta$, su distribución será una normal multivariante, $Y \sim \mathrm{N}_n\,(\mu, \sigma^2 I_n) \equiv \mathrm{N}_n\,(X\beta, \sigma^2 I_n)$, luego la función de verosimilitud[1] de los parámetros dado el vector de observaciones $y \in \mathbb{R}^n$, $L(\mu, \sigma^2; y)$ puede reescribirse en términos de β y σ^2:

$$L(\mu, \sigma^2; y) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \mathrm{e}^{-\frac{||y-\mu||_2^2}{2\sigma^2}} = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \mathrm{e}^{-\frac{||y-X\beta||_2^2}{2\sigma^2}} = L(\beta, \sigma^2; y). \tag{2.8}$$

Mediante el método de estimación por máxima verosimilitud (ver Apéndice A.4.1) se tiene que

$$\begin{aligned} \hat{\beta} &= X^{+}Y \text{ y} \\ \hat{\sigma}^2 &= \frac{1}{n}\left|\left|Y - X\hat{\beta}\right|\right|_2^2. \end{aligned} \tag{2.9}$$

son los estimadores máximo verosímiles de β y σ^2, y la función de verosimilitud alcanza su valor máximo en

$$L(\hat{\beta}, \hat{\sigma}^2; y) = (2\pi \mathrm{e}\hat{\sigma}^2)^{-\frac{n}{2}}. \tag{2.10}$$

Vemos que el estimador máximo verosímil de β coincide con el mínimo cuadrático, en consecuencia hereda todas las propiedades ya vistas. En cuanto al estimador de la varianza,

[1] Se trata de un abuso de notación, L representa siempre la función de verosimilitud de cualquier parámetro, cualquiera que sea la expresión funcional concreta de esta.

$\tilde{\sigma}^2 \neq \hat{\sigma}^2$, pero sí tenemos la relación

$$\hat{\sigma}^2 = \frac{n-p}{n}\tilde{\sigma}^2 = \left(1 - \frac{p}{n}\right)\tilde{\sigma}^2 \tag{2.11}$$

y, en consecuencia, $\mathrm{E}\left[\hat{\sigma}^2\right] = (1 - \frac{p}{n})\sigma^2 \neq \sigma^2$, *i.e.*, $\hat{\sigma}^2$ es sesgado.

OBSERVACIÓN. A continuación trabajamos con distintas distribuciones de probabilidad continuas que se deducen de relaciones a partir de las distribuciones normal y normal multivariante. Pueden consultarse con detalle sus definiciones y algunos resultados propios de la teoría de la Probabilidad en el Apéndice A.3.1.

TEOREMA 2.3. *Sea* $Y \sim \mathrm{N}_n\left(X\beta, \sigma^2 I_n\right)$, *i.e., se ajusta a un modelo lineal normal de rango completo con* $\mathrm{rg}\,(X) = p$. *Los estimadores* $\hat{\beta}$ *y* $\tilde{\sigma}^2$ *verifican las siguientes propiedades:*

1. *Ambos son insesgados.*
2. $\hat{\beta} \sim \mathrm{N}_p\left(\beta, \sigma^2 S^{-1}\right)$.
3. $\frac{(n-p)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$.
4. $\hat{\beta}$ *y* $\tilde{\sigma}^2$ *son independientes.*

COROLARIO 2.1. *Sean* $Y \sim \mathrm{N}_n\left(X\beta, \sigma^2 I_n\right)$, *i.e., se ajusta a un modelo lineal normal de rango completo con* $\mathrm{rg}\,(X) = p$, $h(x)$ *la función de densidad de probabilidad de la distribución* $\chi^2(n-p)$, *y* α_0 *y* α_1 *escalares tales que* $\int_{\alpha_0}^{\alpha_1} h(x)\,\mathrm{d}\,x = 1 - \alpha$ $(\alpha \in]0,1[)$. *Un intervalo de confianza para* σ^2 *a un nivel* $1-\alpha$ *es*

$$I_{1-\alpha}(\sigma^2) = [(n-p)\alpha_1^{-1}\tilde{\sigma}^2, (n-p)\alpha_0^{-1}\tilde{\sigma}^2] \tag{2.12}$$

PROPOSICIÓN 2.1. *Sean* $Y \sim \mathrm{N}_n\left(X\beta, \sigma^2 I_n\right)$, *i.e., se ajusta a un modelo lineal normal de rango completo con* $\mathrm{rg}\,(X) = p$, $h(x)$ *la función de densidad de probabilidad de la distribución* $t(n-p)$ *y* $t_{\alpha/2}(n-p)$ *un valor tal que* $\int_{t_{\alpha/2}(n-p)}^{\infty} h(x)\,\mathrm{d}\,x = \frac{\alpha}{2}$ $(\alpha \in]0,1[)$. *Se verifica:*

1. *Si* $(X\,{}^{\mathrm{t}}\,X)^{-1} = (c_{j,k})_{1 \leq j,k \leq p}$, *entonces*

$$I_{1-\alpha}(\beta_j) = [\hat{\beta}_j - t_{\alpha/2}(n-p)\sqrt{c_{j,j}\tilde{\sigma}^2}, \hat{\beta}_j + t_{\alpha/2}(n-p)\sqrt{c_{j,j}\tilde{\sigma}^2}] \tag{2.13}$$

es un intervalo de confianza para β_j *a un nivel de confianza* $1-\alpha$, $1 \leq j \leq p$.

2. *Si $\lambda \in \mathbb{R}^p$ un vector de constantes conocidas, entonces el estimador de máxima verosimilitud de $\lambda^{\mathrm{t}}\beta$ es $\lambda^{\mathrm{t}}\hat{\beta} \sim \mathrm{N}(\lambda^{\mathrm{t}}\beta, \sigma^2\lambda^{\mathrm{t}}S^{-1}\lambda)$. Además*

$$I_{1-\alpha}(\lambda^{\mathrm{t}}\beta) = [\lambda^{\mathrm{t}}\hat{\beta} - t_{\alpha/2}(n-p)\sqrt{\tilde{\sigma}^2\lambda^{\mathrm{t}}S^{-1}\lambda}, \lambda^{\mathrm{t}}\hat{\beta} + t_{\alpha/2}(n-p)\sqrt{\tilde{\sigma}^2\lambda^{\mathrm{t}}S^{-1}\lambda}] \quad (2.14)$$

es un intervalo de confianza para $\lambda^{\mathrm{t}}\beta$ a un nivel $1-\alpha$.

2.2.2. Contraste de hipótesis

Sea $Y \sim \mathrm{N}_n(X\beta, \sigma^2 I_n)$, es decir, se ajusta a un modelo lineal normal de rango completo, $\mathrm{rg}(X) = p$, podemos plantear distintos contrastes de hipótesis, todos ellos resueltos a partir del método de razón de verosimilitudes, conocidas las distintas distribuciones de probabilidad que se deducen de la normalidad del modelo:

Contraste $\mathrm{H}_0 : \beta = \beta^* \in \mathbb{R}^p$ **conocido.** Sea $h(x)$, la función de densidad de probabilidad de la distribución $F(p, n-p)$ (ver Definición A.10) y $F_\alpha(p, n-p)$ tal que $\int_{F_\alpha(p,n-p)}^{\infty} h(x)\,\mathrm{d}\,x = \alpha$ $(\alpha \in]0,1[)$. El método de razón de verosimilitudes nos conduce a rechazar H_0 para el nivel de significación $\alpha \in]0,1[$ si $F \geq F_\alpha(p, n-p)$, donde

$$F = \frac{n-p}{p}\frac{Q_1}{Q_0}, \quad (2.15)$$

siendo $Q = Q_0 + Q_1 = (Y - X\beta^*)^{\mathrm{t}}(Y - X\beta^*)$, $Q_0 = (Y - X\hat{\beta})^{\mathrm{t}}(Y - X\hat{\beta})$. Además $F \sim F(p, n-p, \nu)$ (ver Definición A.11) con $\nu = \frac{1}{2\sigma^2}(\beta - \beta^*)^{\mathrm{t}}S(\beta - \beta^*)$.

Contraste $\mathrm{H}_0 : \lambda^{\mathrm{t}}\beta = \lambda^*$**,** $\lambda \in \mathbb{R}^p$ **y** $\lambda^* \in \mathbb{R}$**, ambos conocidos.** Sea $t_{\alpha/2}(n-p)$ tal que $\int_{t_\alpha(n-p)}^{\infty} h(x)\,\mathrm{d}\,x = \frac{\alpha}{2}$ $(\alpha \in]0,1[)$, siendo $h(x)$ la función de densidad de probabilidad de la distribución $t(n-p)$ (ver Definición A.12). El método de razón de verosimilitudes nos conduce a rechazar H_0 para el nivel de significación $\alpha \in]0,1[$ si

$$\frac{|\lambda^{\mathrm{t}}\hat{\beta} - \lambda^*|}{\sqrt{\tilde{\sigma}^2\lambda^{\mathrm{t}}S^{-1}\lambda}} \geq t_{\alpha/2}(n-p). \quad (2.16)$$

Contraste $\mathrm{H}_0 : \gamma_1 = \gamma_1^* \in \mathbb{R}^r$ **conocido.** Sea el entero positivo $r \in]0,p[$. Sean $\gamma_1 \in \mathbb{R}^r$ y $\gamma_2 \in \mathbb{R}^{p-r}$ vectores tales que $\beta = (\gamma_1|\gamma_2)$ y, en consecuencia, X_1 y X_2 son dos matrices de dimensiones $n \times r$ y $n \times (p-r)$, de modo que $X\beta = X_1\gamma_1 + X_2\gamma_2$. Sea $F_\alpha(r, n-p)$ tal que $\int_{F_\alpha(r,n-p)}^{\infty} h(x)\,\mathrm{d}\,x = \alpha$ $(\alpha \in]0,1[)$, siendo $h(x)$ la función de densidad de probabilidad de la distribución $F(r, n-p)$ (ver Definición A.10). El método de razón de verosimilitudes

para contrastar la hipótesis nula nos conduce a rechazar H_0 para el nivel de significación $\alpha \in]0,1[$ si $F \geq F_\alpha(r, n-p)$, donde

$$F = \frac{n-p}{r}\frac{Q_1}{Q_0}, \tag{2.17}$$

siendo

- $Q_0 = (Y - X\hat{\beta})^{\mathrm{t}}(Y - X\hat{\beta}) = Y^{\mathrm{t}}(I_n - XX^+)Y$
- $Q_1 = (Y - X_1\gamma_1^*)^{\mathrm{t}}(XX^+ - X_2(X_2{}^{\mathrm{t}}X_2)^{-1}X_2{}^{\mathrm{t}})(Y - X_1\gamma_1^*)$

Además, $F \sim F(r, n-p, \nu)$ (ver Definición A.11), con $\nu = \frac{1}{2\sigma^2}(\gamma_1 - \gamma_1^*)^{\mathrm{t}} B(\gamma_1 - \gamma_1^*)$ y $B = X_1{}^{\mathrm{t}}X_1 - X_1{}^{\mathrm{t}}X_2(X_2{}^{\mathrm{t}}X_2)^{-1}X_2{}^{\mathrm{t}}X_1$.

2.3. Modelo de regresión lineal normal

En el modelo lineal, la matriz de modelo, X, era una matriz de constantes conocidas fijadas por el experimentador. Por contra, en el modelo de regresión lineal, cada columna de X serán las observaciones de una variable aleatoria medida en n unidades experimentales independientes. En los casos en los que las filas $x_i{}^{\mathrm{t}}$ no vienen prefijadas, sino que resultan aleatorias, tendremos que $(y_i, x_i{}^{\mathrm{t}})$, $1 \leq i \leq n$, son n observaciones de vectores aleatorios de dimensión $p+1$, $(Y, X_1, \ldots, X_p)^{\mathrm{t}}$.

Diremos que $(Y, X_1, \ldots, X_p)^{\mathrm{t}}$ es un modelo de regresión lineal normal en n observaciones si

$$Y|(X_1 = x_1, \ldots, X_p = x_p) \sim \mathrm{N}\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j, \sigma^2\right) \tag{2.18}$$

y en consecuencia se tiene

1. $\mathrm{E}[Y|(X_1 = x_1, \ldots, X_p = x_p)] = \beta_0 + \sum_{j=1}^{p}\beta_j x_j$; siendo x_j los n-vectores resultado de observar la variable aleatoria X_j.
2. $\mathrm{cov}[X_j, X_k] = 0$,
3. $\mathrm{Var}[Y|(X_1 = x_1, \ldots, X_p = x_p)] = \sigma^2 < \infty$ y

Consideremos las variables aleatorias independientes $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$. El vector cuyas coordenadas son $Y_1, \ldots, Y_n$ se ajusta a un modelo lineal normal con

$$\mu_i = \mathrm{E}[Y_i] = x_i{}^{\mathrm{t}}\beta = \beta_0 + \sum_{j=1}^{p}\beta_j x_{i,j}, \tag{2.19}$$

siendo la matriz X de orden $n \times (p+1)$, $p+1 \leq n$ y cada fila $x_i{}^{\mathrm{t}} = [1, x_{i,1}, \ldots, x_{i,p}]$.

DEFINICIÓN 2.6 (Coeficiente de determinación). Supongamos que Y se ajusta a un modelo linea. Sea $y = (y_1, \ldots, y_n)^{\mathrm{t}}$ el vector de valores observados de Y e $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)^{\mathrm{t}} = X\hat{\beta}$. Sea $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$. Llamamos coeficiente de determinación y lo denotamos R^2 al valor

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \in [0, 1] \tag{2.20}$$

que representa la proporción de varianza total de Y que es explicada por el modelo.

2.4. Ejemplo

El siguiente problema viene recogido por Dobson (2002) [3]. Tras seis meses de dieta alta en carbohidratos en veinte ($n = 20$) hombres enfermos de diabetes (dependientes de insulina) se recoge la información acerca de su edad (en años), el peso corporal relativo, *i.e.*, comparado con el peso *ideal teórico* dada la altura, y dos variables asociadas a la dieta tales como el porcentaje de calorías obtenidas de hidratos de carbono complejos y el porcentaje de calorías como proteínas.

Los datos vienen recogidos en `carbohydrate` dentro del paquete `dobson` en `R`.

```
library(dobson); data(carbohydrate)
```

Mostramos un resumen de estos

```
summary(carbohydrate)
```

```
##   carbohydrate        age            weight         protein
##  Min.   :24.00   Min.   :23.00   Min.   : 85.0   Min.   :12.0
##  1st Qu.:32.25   1st Qu.:34.50   1st Qu.:100.0   1st Qu.:14.0
##  Median :37.00   Median :47.50   Median :106.0   Median :15.0
##  Mean   :37.60   Mean   :46.15   Mean   :110.7   Mean   :15.9
##  3rd Qu.:42.25   3rd Qu.:57.25   3rd Qu.:120.2   3rd Qu.:18.0
##  Max.   :51.00   Max.   :64.00   Max.   :144.0   Max.   :20.0
```

Modelo. En el estudio se busca establecer un modelo de regresión lineal, siendo el porcentaje de carbohidratos (`carbohydrate`) la variable respuesta, Y, que predecir a partir de las demás. Tenemos el siguiente modelo de regresión lineal:

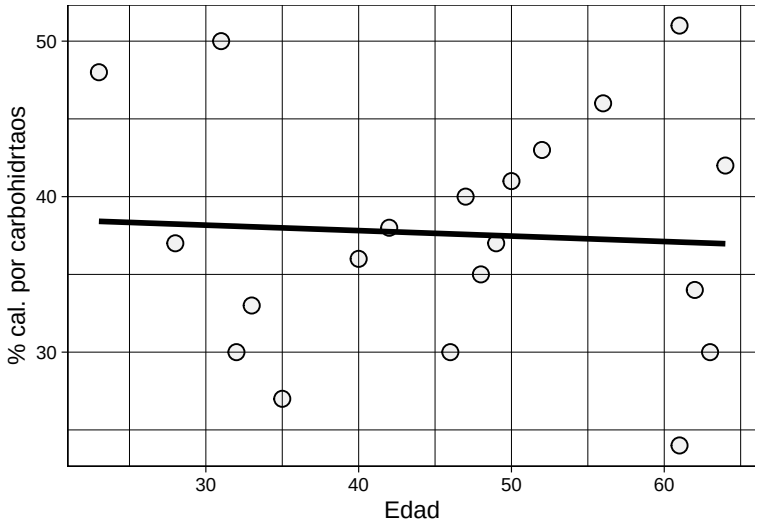
$$\mathrm{E}\,[Y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3},\ 1 \leq i \leq n = 20, \tag{2.21}$$

que guardamos en `R` como

```
modelo <- lm(carbohydrate~age+weight+protein, data=carbohydrate)
```

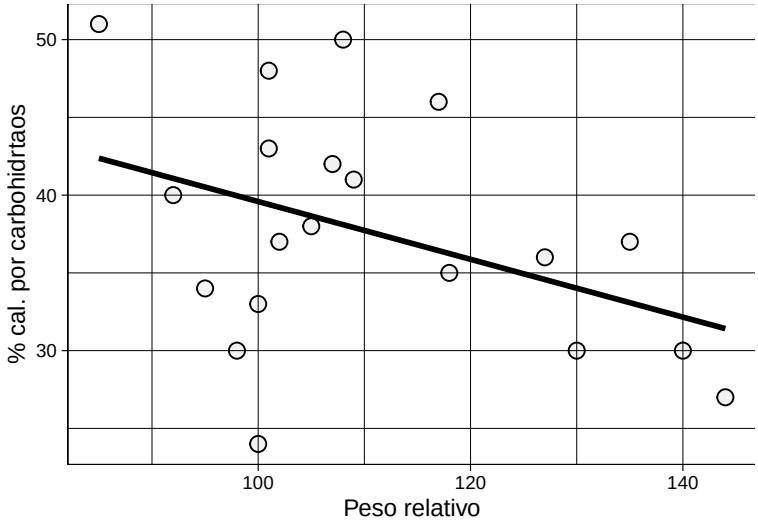
Además, veamos la representación mediante `ggplot2` de la variable respuesta Y como función lineal de cada una de las variables regresoras por separado:

- Y frente a X_1 (edad):



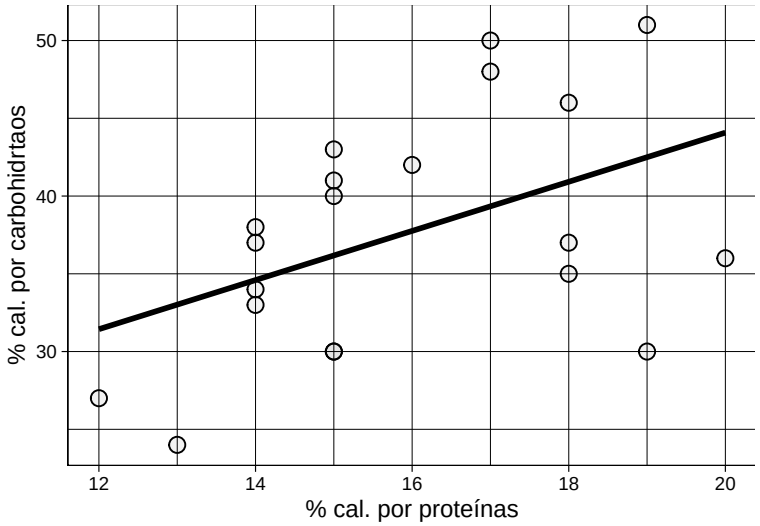
  La recta ajustada para modelo de regresión simple es casi horizontal, habiendo bastante dispersión entre los datos. Cabe suponer que esta variable sea rechazada una vez ajustado el modelo lineal.

- Y frente a X_2 (peso relativo):



  Los datos describen una relación lineal negativa entre `weight` y `carbohydrate`. No parece haber mucha dispersión.

- Y frente a X_3 (% cal. por proteínas):



Tampoco parece haber distancias notables entre la línea ajustada y la nube de puntos de las observaciones, esta vez con una pendiente positiva (*i.e.*, relación lineal positiva entre las variables).

Matricialmente, se expresa

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_{20} \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i,1} & x_{i,2} & x_{i,3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{20,1} & x_{20,2} & x_{20,3} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \tag{2.22}$$

Fijado el modelo, comprobemos que se verifican las hipótesis del modelo lineal normal a partir del estudio del vector de residuos, $e = Y - \hat{Y}$:

Normalidad. Debemos contrastar si la variable respuesta se ajusta a un modelo normal multivariante. Para ello, realizamos el contraste de normalidad de Shapiro-Wilks para los residuos:

```
shapiro.test(rstandard(modelo))
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(modelo)
## W = 0.93186, p-value = 0.1677
```

Nos lleva a asumir la hipótesis nula de normalidad en la variable respuesta.

Homocedasticidad. Sea el contrate de hipótesis

$$\begin{cases} \mathrm{H}_0: & \mathrm{Var}\left[Y_1\right] = \ldots = \mathrm{Var}\left[Y_n\right] (= \sigma^2). \\ \mathrm{H}_1: & \text{Existen } i \text{ y } l,\ 1 \leq i, l \leq n, \text{ tales que } \mathrm{Var}\left[Y_i\right] \neq \mathrm{Var}\left[Y_l\right]. \end{cases} \tag{2.23}$$

Nos servimos del estadístico de contraste de Breusch-Pagan, dado por la función `bptest()` del paquete `lmtest`:

```
library(lmtest)
bptest(modelo)
##
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 0.50666, df = 3, p-value = 0.9174
```

El contraste devuelve un p-valor muy superior a la significación habitual, $p = 0.9174 > 0.05 = \alpha$, luego no hay razones significativas para rechazar la hipótesis nula de homocedasticidad, *i.e.*, existe una varianza σ^2 común.

Incorrelación Debemos además verificar que las coordenadas del vector de residuos e son incorreladas. Sea el contraste

$$\begin{cases} \mathrm{H}_0: & \mathrm{cov}\left[e_i, e_l\right] = 0, 1 \leq i, l \leq n, i \neq l. \\ \mathrm{H}_1: & \text{Existen índices } i \text{ y } l,\ i \neq l \text{ tales que } \mathrm{cov}\left[e_i, e_l\right] \neq 0. \end{cases} \tag{2.24}$$

Del mismo paquete, usamos la función `dwtest()` para ejecutar el contraste de Durbin-Watson que evalúa la incorrelación de los residuos.

```
dwtest(modelo)
##
##  Durbin-Watson test
##
## data:  modelo
## DW = 1.8752, p-value = 0.3314
## alternative hypothesis: true autocorrelation is greater than 0
```

El p-valor obtenido es superior al nivel de significación usual, $p = 0.3314 > 0.05$, luego no hay razones estadísticas para rechazar la hipótesis nula de incorrelación de los errores.

Estimación y contraste de hipótesis. Todo lo que sigue, es una aplicación de la teoría vista en el Capítulo 2.2.1.

Verificada la hipótesis de homocedasticidad, lo primero será estimar σ^2 (que ya hemos señalado que es única mediante el contraste de homocedasticidad) a partir de un estimador insesgado $\tilde{\sigma}^2$. Ejecutamos

```
summary(modelo)
##
## Call:
## lm(formula = carbohydrate ~ age + weight + protein, data = carbohydrate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3424  -4.8203   0.9897   3.8553   7.9087
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.96006   13.07128   2.828  0.01213 *
## age         -0.11368    0.10933  -1.040  0.31389
## weight      -0.22802    0.08329  -2.738  0.01460 *
## protein      1.95771    0.63489   3.084  0.00712 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.956 on 16 degrees of freedom
## Multiple R-squared:  0.4805,Adjusted R-squared:  0.3831
## F-statistic: 4.934 on 3 and 16 DF,  p-value: 0.01297
```

La función `lm()` de `R` opera para modelos lineales cualesquiera, sin imponer ninguna distribución de probabilidad sobre Y, por tanto las estimaciones $\hat{\beta}$ son las mínimas cuadráticas. En nuestro caso ya hemos contrastado y aceptado la normalidad, luego estos estimadores serán además los máximo verosímiles (ver ecuación (2.9)). La columna `Estimate` recoge estimador de β,

$$\hat{\beta} = (36.96006, -0.11368, -0.22802, 1.95771)^{\mathrm{t}} \in \mathbb{R}^4. \tag{2.25}$$

Operando con $n = 20$ y $p = 4$, se tiene que $\tilde{\sigma}^2 = 35.479$. Luego, podemos construir intervalos de confianza para cada β_j a partir de $I_{1-\alpha}(\beta_j) = [\hat{\beta}_j - t_{\alpha/2}(n-p)\sqrt{c_{j,j}\tilde{\sigma}^2}, \hat{\beta}_j +$

$t_{\alpha/2}(n-p)\sqrt{c_{j,j}\tilde{\sigma}^2}]$, $0 \leq j \leq 3$. La columna `Std. Error` son los valores $\sqrt{c_{j,j}\tilde{\sigma}^2}$ en cada j, y $t_{0.975}(16) = 2.119905 \approx 2.12$. Construimos los intervalos de confianza como sigue:

- $I_{0.95}(\beta_0) = 36.9601 \mp 13.0713 \times 2.12 = [9.250, 64.670] \not\ni 0$,
- $I_{0.95}(\beta_1) = -0.1137 \mp 0.1093 \times 2.12 = [-0.345, 0.118] \ni 0$,
- $I_{0.95}(\beta_2) = -0.2280 \mp 0.0833 \times 2.12 = [-0.405, -0.051] \not\ni 0$ e
- $I_{0.95}(\beta_3) = 1.9577 \mp 0.6349 \times 2.12 = [0.612, 3.304] \not\ni 0$,

siendo la variable `age` no significativa para una significación $\alpha = 0.05$. Podemos reformular el modelo excluyéndola como sigue

```
modelo2 <- lm(carbohydrate~weight+protein, data=carbohydrate)
summary(modelo2)
##
## Call:
## lm(formula = carbohydrate ~ weight + protein, data = carbohydrate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6812  -3.9135   0.9464   4.0880   9.7948
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.13032   12.57155   2.635  0.01736 *
## weight      -0.22165    0.08326  -2.662  0.01642 *
## protein      1.82429    0.62327   2.927  0.00941 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.971 on 17 degrees of freedom
## Multiple R-squared:  0.4454,Adjusted R-squared:  0.3802
## F-statistic: 6.827 on 2 and 17 DF,  p-value: 0.006661
```

El nuevo vector de parámetros estimados es

$$\hat{\beta} = (33.13032, -0.22165, 1.82429)^{\mathrm{t}} \in \mathbb{R}^3. \tag{2.26}$$

Análogamente, tenemos uan nueva columna `Std. Error` para `modelo2` y en esta ocasión $t_{0.975}(17) = 2.109816 \approx 2.11$. los nuevos intervalos de confianza, para $\alpha = 0.05$ son

- $I_{0.95}(\beta_0) = 33.1303 \mp 12.5715 \times 2.11 = [6.607, 59.654] \not\ni 0$,
- $I_{0.95}(\beta_2) = -0.2216 \mp 0.0833 \times 2.11 = [-0.397, -0.046] \not\ni 0$ e
- $I_{0.95}(\beta_3) = 1.8243 \mp 0.6233 \times 2.11 = [0.509, 3.139] \not\ni 0$.

Bondad de ajuste. Para determinar el buen ajuste de los datos al modelo lineal normal propuesto,

$$\mathrm{E}[Y_i] = 33.1303 - 0.2216x_{i,2} + 1.8243x_{i,3},\ 1 \leq i \leq n = 20, \tag{2.27}$$

recurrimos al estadístico F, para `modelo2`.

Concluimos que el modelo es significativo puesto que el p-valor de significatividad conjunta de las variables es $p = 0.006661 < 0.05 = \alpha$, como suele ser habitual.

En cuanto al coeficiente de determinación $R^2 = 0.4454$, es decir, el modelo explica aproximadamente un $44.54\,\%$ de la variabilidad total de la variable respuesta. Hemos visto que las variables se ajustan a lo exigido por la definición de modelo lineal normal y su significatividad. Para un análisis más completo de la investigación realizada en el ejemplo cabría plantearse al observación de nuevas variables (con las que no contamos en el enunciado) mejorando la capacidad explicativa del modelo.

Capítulo 3

Modelo lineal generalizado

La primera extensión del modelo lineal que consideramos en este trabajo es el modelo lineal generalizado. La propuesta original de este modelo se debe a Nelder y Wedderburn (1972) [10]. Si bien en el modelo lineal no asumíamos ninguna distribución para la variable respuesta, en el modelo lineal normal, como su nombre indica, se asume la hipótesis de normalidad. El modelo lineal generalizado partirá de la hipótesis de que la variable objetivo sigue alguna distribución de las llamadas exponenciales (estudiadas de manera general en la asignatura Inferencia Estadística del Grado en Estadística), las cuales incluyen la distribución normal. Una segunda novedad de este modelo frente al modelo lineal normal es que la relación entre las medias y la función lineal de los parámetros del modelo, β, no es directa, sino que viene mediada por cierta función de enlace sobre la que impondremos las condiciones de monotonía y derivabilidad. Se tiene, por tanto, que el modelo lineal generalizado, como su propio nombre indica, viene a generalizar el modelo lineal normal recogiendo perfectamente a este como caso particular.

3.1. La familia exponencial de distribuciones

Muchas de las familias de distribuciones que se utilizan en la inferencia estadística se encuadran en la llamada familia exponencial. A continuación presentamos una definición formal y completa, proporcionada por Rohatgi (1976) [12]. En el Capítulo 3.1.2 ofrecemos una expresión de la densidad que nos será de más utilidad en el ámbito de los modelos lineales generalizados.

3.1.1. Definiciones y resultados

Sea Θ un intervalo abierto k-dimensional (posiblemente no acotado) de $\mathbb{R}^k$, y sea $\{p(\,\cdot\,;\theta) : \theta = (\theta_1, \ldots, \theta_k)^{\mathrm{t}} \in \Theta\}$ una familia de funciones de densidad definidas sobre $\mathbb{R}^n$. Suponemos que el conjunto $\{x \in \mathbb{R}^n : p(x;\theta) > 0\}$ es independiente del parámetro $\theta = (\theta_1, \ldots, \theta_k)^{\mathrm{t}}$.

Definición 3.1 (Familia exponencial). Diremos que la familia de funciones de densidad $\{p(\,\cdot\,;\theta) : \theta \in \Theta\}$ es una familia exponencial si existen funciones $Q_i : \Theta \to \mathbb{R}$, $1 \leq i \leq k$, y $D : \Theta \to \mathbb{R}$ y funciones medibles $T_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leq i \leq k$, y $S : \mathbb{R}^n \to \mathbb{R}$ verificando

$$p(x;\theta) = \exp\left(\sum_{i=1}^{k} Q_i(\theta)T_i(x) + D(\theta) + S(x)\right),\ x \in \mathbb{R}^n. \tag{3.1}$$

Observación. Llamamos familia exponencial uniparamétrica a la dada en la Definición 3.1 con $k = 1$, $p(x;\theta) = \exp\left(Q(\theta)T(x) + D(\theta) + S(x)\right)$, para $x \in \mathbb{R}^n$.

Definición 3.2 (Parámetro natural). Sea $\{p(\,\cdot\,;\theta) : \theta \in \Theta\}$ una familia exponencial uniparamétrica. Llamamos parámetro natural de $p(\,\cdot\,;\theta)$ al valor $Q(\theta)$.

Familia exponencial con $n = 1$ y $k = 1$

En lo que sigue, presentamos algunos resultados útiles cuando la variable aleatoria, X, es unidimensional y depende un parámetro escalar, θ, *i.e.*, su función de densidad de probabilidad es $p(x;\theta) = \exp\left(Q(\theta)T(x) + D(\theta) + S(x)\right)$, para $x \in \mathbb{R}$.

En este caso, podemos encontrar expresiones generales para calcular la esperanza y la varianza de $T(X)$. En efecto, tenemos que $\int_{-\infty}^{+\infty} p(x;\theta)\,\mathrm{d}\,x = 1$ por ser $p(\,\cdot\,;\theta)$ función de densidad de probabilidad. Si derivamos a ambos lados respecto a θ e intercambiamos la integración respecto a x y la derivación respecto a θ, según el Teorema A.2:

$$\int_{-\infty}^{+\infty} \frac{\partial p(x;\theta)}{\partial\theta}\,\mathrm{d}\,x = \frac{\partial \int_{-\infty}^{+\infty} p(x;\theta)\,\mathrm{d}\,x}{\partial\theta} = \frac{\partial 1}{\partial\theta} = 0. \tag{3.2}$$

Igualmente, si derivamos dos veces, tenemos

$$\int_{-\infty}^{+\infty} \frac{\partial^2 p(x;\theta)}{\partial\theta^2}\,\mathrm{d}\,x = 0. \tag{3.3}$$

Proposición 3.1. *Sea X una variable aleatoria con función de densidad de probabilidad $p(x;\theta)$ perteneciente a una familia exponencial. Se verifica:*

1. $\mathrm{E}\left[T(X)\right] = -\frac{D'(\theta)}{Q'(\theta)}$ *y*

2. $\operatorname{Var}[T(X)] = \frac{Q''(\theta)D'(\theta) - Q'(\theta)D''(\theta)}{(Q'(\theta))^3}$.

Demostración. 1. De la expresión exponencial de $p(\,\cdot\,;\theta)$ tenemos que

$$\frac{\partial p(x;\theta)}{\partial \theta} = (T(x)Q'(\theta) + D'(\theta))p(x;\theta). \tag{3.4}$$

Como $\int_{-\infty}^{+\infty} \frac{\partial p(x;\theta)}{\partial \theta}\,\mathrm{d}\,x = 0$, reordenando y por la definición de esperanza se tiene que

$$\operatorname{E}[T(X)] = -\frac{D'(\theta)}{Q'(\theta)}. \tag{3.5}$$

2. Si derivamos dos veces y sustituimos en el segundo sumando resultante, se tiene

$$\begin{aligned}\frac{\partial^2 p(x;\theta)}{\partial \theta^2} &= (T(x)Q''(\theta) + D''(\theta))p(x;\theta) + (T(x) + Q'(\theta) + D'(\theta))^2 p(x;\theta) \\ &= (T(x)Q''(\theta) + D''(\theta))p(x;\theta) + (Q'(\theta))^2 (T(x) - \operatorname{E}[T(X)])^2 p(x;\theta).\end{aligned} \tag{3.6}$$

La integral de la expresión anterior también es nula y sustituimos en el segundo sumando la definición de la varianza, llegamos a

$$\begin{aligned}0 &= \int_{-\infty}^{+\infty} \frac{\partial^2 p(x;\theta)}{\partial \theta^2}\,\mathrm{d}\,x \\ &= \operatorname{E}\left[T(X)Q''(\theta) + D''(\theta))p(x;\theta) + Q'(\theta)^2 (T(x) - \operatorname{E}[T(X)])^2\right] \\ &= Q''(\theta)\operatorname{E}[T(X)] + D''(\theta) + Q'(\theta)^2 \operatorname{Var}[T(X)] \\ &= Q''(\theta)\left(-\frac{D'(\theta)}{Q'(\theta)}\right) + D''(\theta) + Q'(\theta)^2 \operatorname{Var}[T(X)] \\ &= \frac{D''(\theta)Q'(\theta) - Q''(\theta)D'(\theta)}{Q'(\theta)} + Q'(\theta)^2 \operatorname{Var}[T(X)]\,.\end{aligned} \tag{3.7}$$

Luego

$$\operatorname{Var}[T(X)] = \frac{Q''(\theta)D'(\theta) - Q'(\theta)D''(\theta)}{Q'(\theta)^3}. \tag{3.8}$$

□

Definición 3.3 (Forma canónica). Sea X una variable aleatoria con función de densidad de probabilidad $p(x;\theta)$ perteneciente a una familia exponencial con $n = 1$ y $k = 1$. Diremos que $p(x;\theta)$ está en forma canónica si $T(x) = x$.

Cuando la familia de densidades está en forma canónica, la Proposición 3.1 nos proporciona $\operatorname{E}[X]$ y $\operatorname{Var}[X]$.

3.1.2. Distribuciones con parámetros de localización y dispersión

En este aprtado vamos a reescribir algunos resultados y definiciones sobre la familia exponencial en los casos $k = 1$ y $k = 2$, pues serán aquellas que atañen al modelo lineal generalizado. De hecho, Nelder y Wederburn (1976) [10], Faraway (2016) [4] y Wood (2007) [16] explican el modelo lineal generalizado a partir de la familia exponencial como haremos en la ecuación (3.9).

Definiciones y resultados

Es posible reparametrizar la familia exponencial en el caso $k = 1$ utilizando el parámetro natural $Q(\theta)$ (al que, sin pérdida de generalidad, denotamos en adelante θ), siendo este un parámetro relacionado con la localización. Además, las funciones $D(\theta)$ y $S(x)$ pueden depender de un segundo parámetro, ϕ, que sin ser el que describe la familia, sí aporta información relativa a la dispersión de las distribuciones de la familia exponencial.

Sean las funciones $b : \Theta \to \mathbb{R}$ y $a : \Phi \to \mathbb{R}$, y las funciones medibles $T : \mathbb{R} \to \mathbb{R}$ y $c : \mathbb{R} \times \Phi \to \mathbb{R}$. Asumiremos:

- $D(\theta) = -\frac{b(\theta)}{a(\phi)}$
- $S(x) = c(x, \phi)$

Tenemos entonces la función de densidad en términos del parámetro natural θ:

$$p(x; \theta, \phi) = \exp\left(\frac{T(x)\theta - b(\theta)}{a(\phi)} + c(x, \phi)\right),\ x \in \mathbb{R}. \tag{3.9}$$

En ocasiones contaremos con un único parámetro, θ, por lo que implícitamente estaremos asumiendo $a(\phi) = \phi = 1$. En tales casos, la función de densidad de probabilidad se simplifica y toma la expresión

$$p(x; \theta) = \exp(T(x)\theta - b(\theta) + c(x)),\ x \in \mathbb{R}. \tag{3.10}$$

Podemos reformular la Proposición 3.1 para la nueva ecuación (3.9) como sigue:

PROPOSICIÓN 3.2. *Sea X una variable aleatoria con función de densidad de probabilidad $p(x; \theta, \phi)$ perteneciente a una familia exponencial de localización y dispersión. Se verifica:*

1. $\mathrm{E}[T(X)] = b'(\theta)$
2. $\mathrm{Var}[T(X)] = a(\phi)b''(\theta)$.

Distribuciones de la familia exponencial y sus propiedades

A continuación veremos que las distribuciones normal, gamma, exponencial, binomial, geométrica y de Poisson pertenecen a la familia exponencial en forma canónica. Luego, podremos servirnos de su expresión como distribuciones de la familia exponencial y la Proposición 3.2 para calcular $\mathrm{E}[X]$ y $\mathrm{Var}[X]$. Además estaremos interesados en calcular la información de Fisher (ver Definición A.24)

$$\mathcal{I}(\theta) = \mathrm{Cov}\left[U_\theta(X)\right], \tag{3.11}$$

siendo $U_\theta(x)$ el *score* como se introduce en la Definición A.23, para el parámetro natural. Veremos que es de utilidad en la estimación de los modelos lineales generalizados.

Distribución normal. Sea la variable aleatoria $X \sim \mathrm{N}(\mu, \sigma^2)$ (ver Definición A.6) con función de densidad de probabilidad

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{(2\pi)\sigma^2}} \mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \ x \in \mathbb{R}. \tag{3.12}$$

Escribimos la función de log-verosimilitud como

$$\begin{aligned} \ell(\mu, \sigma^2; x) &= \log(2\pi\sigma^2) - \frac{1}{2}\frac{x^2 + \mu^2 - 2x\mu}{\sigma^2} \\ &= \frac{x\mu - \frac{1}{2}\mu^2}{\sigma^2} + \frac{1}{2}\left[\log(2\pi\sigma^2) - \frac{x^2}{\sigma^2}\right] \end{aligned} \tag{3.13}$$

luego, los parámetros son $\theta = \mu$ y $\phi = \sigma^2$, y las funciones son

$$\begin{cases} b(\theta) = \frac{\theta^2}{2} \\ a(\phi) = \phi \\ c(x, \phi) = \frac{1}{2}\left[\log(2\pi\sigma^2) - \frac{x^2}{\sigma^2}\right] \end{cases} \tag{3.14}$$

y podemos calcular la esperanza y la varianza de X:

$$\mathrm{E}[X] = b'(\theta) = \mu, \ \mathrm{Var}[X] = a(\phi)b''(\theta) = \sigma^2 \tag{3.15}$$

Se tiene que $\ell'(\theta; x, \phi) = U_\theta(x) = \frac{x-\theta}{\phi}$, luego la información de Fisher para θ es

$$\mathcal{I}(\theta) = \mathrm{E}\left[U_\theta(X)^2\right] = \frac{1}{\phi}. \tag{3.16}$$

Distribución gamma. Sea la variable aleatoria $X \sim \text{Gamma}\,(\alpha, \beta)$ (ver Definición A.13) con función de densidad de probabilidad

$$p(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-x\beta} I_{]0,\infty[}(x). \tag{3.17}$$

Su función de log-verosimilitud es

$$\begin{aligned} \ell(\alpha, \beta; x) &= \log(x^{\alpha-1}\beta^\alpha \mathrm{e}^{-\beta x}) - \log(\Gamma(\alpha)) \\ &= \frac{\frac{\beta}{\alpha}x - \log(\beta)}{-\frac{1}{\alpha}} + (\alpha - 1)\log(x) - \log(\Gamma(\alpha)), \end{aligned} \tag{3.18}$$

luego los parámetros son $\theta = \frac{\beta}{\alpha}$ y $\phi = \alpha^{-1}$, y las funciones son

$$\begin{cases} b(\theta) = \log(\theta) \\ a(\phi) = -\phi \\ c(x, \phi) = \frac{\log(\phi)}{\phi} + (\phi^{-1} - 1)\log(x) - \log(\Gamma(\phi^{-1})) \end{cases} \tag{3.19}$$

y podemos calcular la esperanza y la varianza de X:

$$\mathrm{E}\,[X] = b'(\theta) = \frac{\alpha}{\beta}, \ \mathrm{Var}\,[X] = a(\phi)b''(\theta) = \frac{\alpha}{\beta^2} \tag{3.20}$$

Se tiene que $\ell'(\theta; x, \phi) = U_\theta(x) = -\frac{1}{\phi}(x - \frac{1}{\theta}$, luego la información de Fisher para θ es

$$\mathcal{I}(\theta) = \mathrm{E}\left[U_\theta(X)^2\right] = \frac{1}{\theta^2\phi} \tag{3.21}$$

Distribución exponencial. Es un caso particular de la la distribución gamma cuando $\alpha = 1$ (ver Proposición A.11).

Distribución geométrica. Sea la variable aleatoria $X \sim \text{Geom}\,(\pi)$ (ver Definición A.17) con función masa de probabilidad

$$p(x; \pi) = \pi(1 - \pi)^x I_{\mathbb{Z}_0^+}(x). \tag{3.22}$$

Su función de log-verosimilitud es

$$\ell(\pi; x) = x\log(1 - \pi) + \log(\pi), \tag{3.23}$$

luego el parámetro es $\theta = \log(1 - \pi)$ y las funciones son

$$\begin{cases} b(\theta) = -\log(\mathrm{e}^\theta + 1) \\ c(x) = 0 \end{cases} \tag{3.24}$$

y podemos calcular la esperanza y la varianza de X:

$$\mathrm{E}\left[X\right] = b'(\theta) = \frac{1-\pi}{\pi},\ \mathrm{Var}\left[X\right] = b''(\theta) = \frac{1-\pi}{\pi^2} \tag{3.25}$$

Se tiene que $\ell'(\theta; x) = U_\theta(x) = x + \frac{\mathrm{e}^\theta}{\mathrm{e}^\theta+1}$, luego la información de Fisher para θ es

$$\mathcal{I}(\theta) = \mathrm{E}\left[U_\theta(X)^2\right] = -\frac{\mathrm{e}^\theta}{(\mathrm{e}^\theta+1)^2}. \tag{3.26}$$

Distribución binomial. Sea la variable aleatoria $X \sim \mathrm{b}_n\left(\pi\right)$ (ver Definición A.18) con función masa de probabilidad

$$p(x;\pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x}I_{\{0,\dots,n\}}(x). \tag{3.27}$$

Su función de log-verosimilitud es

$$\begin{aligned}\ell(\pi; x) &= \log\binom{n}{x} + x\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi)\\ &= x\log\left(\frac{\pi}{1-\pi}\right) - (-n\log(1-\pi)) + \log\binom{n}{x},\end{aligned} \tag{3.28}$$

luego el parámetro es $\theta = \log(\frac{\pi}{1-\pi})$ y las funciones son

$$\begin{cases} b(\theta) = n\log(1+\mathrm{e}^\theta) \\ c(x) = \log\binom{n}{x} \end{cases} \tag{3.29}$$

y podemos calcular la esperanza y la varianza de X:

$$\mathrm{E}\left[X\right] = b'(\theta) = n\pi,\ \mathrm{Var}\left[X\right] = b''(\theta) = n\pi(1-\pi) \tag{3.30}$$

Se tiene que $\ell'(\theta; x) = U_\theta(x) = x - n\mathrm{e}^\theta(1+\mathrm{e}^\theta)^{-1}$, luego la información de Fisher para π es

$$\mathcal{I}(\theta) = \mathrm{E}\left[U_\theta(X)^2\right] = n\frac{\mathrm{e}^\theta}{(1+\mathrm{e}^\theta)^2}. \tag{3.31}$$

Distribución de Poisson. Sea la variable aleatoria $X \sim \mathrm{Poisson}\left(\lambda\right)$ (ver Definición A.19) con función masa de probabilidad

$$p(x;\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}I_{\mathbb{Z}_0^+}(x). \tag{3.32}$$

Escribimos la función de log-verosimilitud

$$\ell(\lambda; x) = -\lambda + x\log(\lambda) - \log(x!) = x\log(\lambda) - \lambda - \log(x!), \tag{3.33}$$

luego el parámetro es $\theta = \log(\lambda)$ y las funciones son

$$\begin{cases} b(\theta) = \exp(\theta) \\ c(x) = -\log(x!) \end{cases} \tag{3.34}$$

y podemos calcular la esperanza y la varianza de X:

$$\mathrm{E}[X] = b'(\theta) = \lambda, \ \mathrm{Var}[X] = b''(\theta) = \lambda, \tag{3.35}$$

Se tiene que $\ell'(\theta; x) = U_\theta(x) = x - \mathrm{e}^\theta$, luego la información de Fisher para θ es

$$\mathcal{I}(\theta) = \mathrm{E}\left[U_\theta(X)^2\right] = \mathrm{e}^\theta. \tag{3.36}$$

3.2. Modelo lineal generalizado

Definición 3.4 (Modelo lineal generalizado). Sean $Y = (Y_1, \ldots, Y_n)^{\mathrm{t}}$ un vector aleatorio de dimensión n y media $\mathrm{E}[Y] \in \mathbb{R}^n$, X una matriz de dimensión $n \times p$, con $p \leq n$, de constantes conocidas, que escribiremos

$$X = \begin{bmatrix} {x_1}^{\mathrm{t}} \\ \vdots \\ {x_n}^{\mathrm{t}} \end{bmatrix}, \tag{3.37}$$

siendo el vector ${x_i}^{\mathrm{t}}$ la i-ésima fila de X, y $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{t}} \in \mathbb{R}^p$ un vector de parámetros desconocidos. Además, sea $g : \mathbb{R} \to \mathbb{R}$ una función monótona y diferenciable. Diremos que Y se ajusta a un modelo lineal generalizado si verifica:

1. Las coordenadas de Y son variables aleatorias independientes con funciones de densidad $p(y_i; \theta_i, \phi_i)$, $1 \leq i \leq n$, pertenecientes todas ellas a la misma familia exponencial en forma canónica, $T \equiv \mathrm{id}$.

2. $g(\mu_i) = {x_i}^{\mathrm{t}} \beta$, siendo ${x_i}^{\mathrm{t}}$ la i-ésima fila de X, $1 \leq i \leq n$.

Definición 3.5 (Predictor lineal). Llamamos predictor lineal de un modelo lineal generalizado a $\eta = (\eta_1, \ldots, \eta_n)^{\mathrm{t}} = X\beta$, *i.e.*, $\eta_i = {x_i}^{\mathrm{t}} \beta$, $1 \leq i \leq n$.

Observación. Tengamos en cuenta:

1. En el contexto de los modelos lineales generalizados, $g : z \in \mathbb{R} \to g(z) \in \mathbb{R}$ recibe el nombre de función de enlace. Si consideramos la función vectorial $G : z \in \mathbb{R}^n \to G(z) = (g(z_1), \ldots, g(z_n))^{\mathrm{t}} \in \mathbb{R}^n$, entonces la condición 2 de la Definición 3.4 se puede escribir como $G(\mathrm{E}[Y]) = X\beta$ y diremos que G es la función de enlace vectorial.

2. En la condición 1 de la Definición 3.4 hemos indicado que cada Y_i sigue una distribución de la misma familia exponencial, luego $\mathrm{E}[Y_i] = \mu_i = b'(\theta_i)$, $1 \leq i \leq n$. Esta relación viene dada por la expresión de la esperanza de las Y_i, que están en forma canónica y hemos desarrollado en la Proposición 3.2: $\mathrm{E}[T(Y_i)] = \mathrm{E}[Y_i] = b'(\theta_i)$, $1 \leq i \leq n$.

En ocasiones, escribiremos el modelo a partir del predictor lineal, como $g(\mu_i) = \eta_i = x_i{}^{\mathrm{t}}\beta$. Alternativamente, por ser g invertible, podemos escribir $\mu_i = g^{-1}(\eta_i)$, $1 \leq i \leq n$.

Faraway (2016) [4] propone asumir $a(\phi_i) = a_i(\phi) = \phi/w_i$ para ciertos valores $\phi \in \mathbb{R}$ y $w_i \neq 0$, $1 \leq i \leq n$. Con ello[1] la función de densidad de probabilidad puede ser reescrita como

$$p(y_i; \theta_i, \phi_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right] = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi)\right] = p(y_i; \theta_i, \phi). \tag{3.38}$$

Ejemplo 3.1. El modelo lineal normal de la Definición 2.3 es un modelo lineal generalizado en el que la familia de distribuciones exponenciales es la normal con el parámetro asociado a la distribución, $\theta_i = \mu_i$ y cuya función de enlace es $g \equiv \mathrm{id}$; luego $\theta_i = \mu_i = \eta_i$ y la varianza común es $\sigma^2 = a(\phi) = \phi = \phi_i$, $1 \leq i \leq n$. Si hubiera heterocedasticidad, procedemos según lo señalado en el Capítulo 2.1.1 (página 15).

3.3. Estimación

Consideremos un modelo lineal generalizado como el dado en la Definición 3.4. Siguiendo lo propuesto por Dobson (2002) [3], los parámetros de un modelo lineal generalizado serán estimados por máxima verosimilitud (ver Apéndice A.4.1) y no por mínimos cuadrados, ya que la hipótesis de pertenecer a una familia de distribuciones exponencial nos garantiza que disponemos de las densidades $p(y_i; \theta_i, \phi)$, $1 \leq i \leq n$. La log-verosimilitud

[1] Abusamos de notación en el uso de la función $c(y, \cdot)$ en adelante.

para un vector de observaciones de Y, $y = (y_1, \ldots, y_n)^{\mathrm{t}}$, de una familia exponencial en forma canónica y con parámetros $\theta = (\theta_1, \ldots, \theta_n)^{\mathrm{t}}$ será:

$$\begin{aligned} \ell(\theta; y, \phi) &= \sum_{i=1}^{n} \ell(\theta_i; y_i, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + \sum_{i=1}^{n} c(y_i, \phi) \\ &= \frac{1}{\phi} \sum_{i=1}^{n} (y_i\theta_i - b(\theta_i)) w_i + \sum_{i=1}^{n} c(y_i, \phi), \end{aligned} \tag{3.39}$$

donde $\mu_i = b'(\theta_i)$ y $\eta_i = g(\mu_i) = {x_i}^{\mathrm{t}}\beta$, $1 \leq i \leq n$. Luego, asumiendo b' invertible, $\theta_i = (g \circ b')^{-1}(\eta_i) = (g \circ b')^{-1}({x_i}^{\mathrm{t}}\beta)$ y escribimos la función de log-verosimilitud[2] en términos de β del siguiente modo:

$$\begin{aligned} \ell(\theta; y, \phi) &= \frac{1}{\phi} \sum_{i=1}^{n} (y_i\theta_i - b(\theta_i)) w_i + \sum_{i=1}^{n} c(y_i, \phi) \\ &= \frac{1}{\phi} \sum_{i=1}^{n} [y_i (g \circ b')^{-1}({x_i}^{\mathrm{t}}\beta) - b((g \circ b')^{-1}({x_i}^{\mathrm{t}}\beta)] w_i + \sum_{i=1}^{n} c(y_i, \phi) \\ &= \ell(\beta; y). \end{aligned} \tag{3.40}$$

Una propiedad de la familia exponencial de distribuciones es que satisface suficientes condiciones de regularidad para asegurar que el máximo global de la función de log-verosimilitud se obtiene de manera única como solución del sistema de p ecuaciones en p incógnitas (β_j, $1 \leq j \leq p$):

$$\begin{cases} U_{\beta,1}(y) = \frac{\partial \ell(\beta; y)}{\partial \beta_1} = 0 \\ \vdots \\ U_{\beta,p}(y) = \frac{\partial \ell(\beta; y)}{\partial \beta_p} = 0 \end{cases} \tag{3.41}$$

PROPOSICIÓN 3.3. *Sea Y un vector aleatorio n-dimensional verificando las condiciones de un modelo lineal generalizado. Se prueba que*

$$U_{\beta,j}(y) = \frac{\partial \ell(\beta; y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{i,j}}{\operatorname{Var}[Y_i]\, g'(\mu_i)},\ 1 \leq j \leq p, \tag{3.42}$$

siendo $x_{i,j}$ la j-ésima coordenada de ${x_i}^{\mathrm{t}}$, $1 \leq i \leq n$.

Demostración. Derivando parcialmente y aplicando la regla de la cadena en cada coorde-

[2]Se trata de un abuso de notación, ℓ representa siempre la función de log-verosimilitud de cualquier parámetro, cualquiera que sea la expresión funcional concreta de esta.

nada, se tiene:

$$\begin{aligned}
U_{\beta,j}(y) &= \frac{\partial \ell(\beta;y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell(\theta_i;y_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} w_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b'(\theta_i)\frac{\partial \theta_i}{\partial \beta_j}\right) \\
&= \frac{1}{\phi}\sum_{i=1}^{n} w_i (y_i - b'(\theta_i))\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} w_i (y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} \\
&= \frac{1}{\phi}\sum_{i=1}^{n} w_i (y_i - \mu_i)\frac{1}{\frac{\partial^2 b(\theta_i)}{\partial \theta_i{}^2}}\frac{\partial \mu_i}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} \frac{w_i(y_i-\mu_i)}{\frac{w_i \operatorname{Var}[Y_i]}{\phi}}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\operatorname{Var}[Y_i]}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\operatorname{Var}[Y_i]}\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} x_{i,j} \\
&= \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{i,j}}{\operatorname{Var}[Y_i]\, g'(\mu_i)}.
\end{aligned} \tag{3.43}$$

□

Definición 3.6 (Función de varianza). Sea un modelo lineal generalizado con vector de medias μ. Llamamos función de varianza, y lo denotamos $V(\mu_i)$ a

$$V(\mu_i) = \frac{b''(\theta_i)}{w_i} = \frac{\operatorname{Var}[Y_i]}{\phi}. \tag{3.44}$$

En consecuencia el resultado de la Proposición 3.3 se reformula

$$U_{\beta,j}(y) = \frac{1}{\phi}\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{i,j}}{V(\mu_i)g'(\mu_i)} \tag{3.45}$$

Para el cálculo de la matriz de información $\mathcal{I}(\beta)$, de dimensión $p \times p$, acudimos a su definición (ver Definición A.24) y tenemos que el coeficiente (j,k) de dicha matriz es

$$\begin{aligned}
(\mathcal{I}(\beta))_{j,k} &= \mathrm{E}\left[(U_{\beta,j}(y) - \mathrm{E}\left[U_{\beta,j}(y)\right])(U_{\beta,k}(y) - \mathrm{E}\left[U_{\beta,k}(y)\right])\right] = \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \frac{(y_i-\mu_i)x_{i,j}}{\operatorname{Var}[Y_i]\, g'(\mu_i)} \sum_{l=1}^{n} \frac{(y_l-\mu_l)x_{l,k}}{\operatorname{Var}[Y_l]\, g'(\mu_l)}\right] = \sum_{i=1}^{n} \frac{x_{i,j}x_{i,k}\operatorname{Var}[Y_i]}{\operatorname{Var}[Y_i]^2\, g'(\mu_i)^2},\ 1 \le j,k \le p
\end{aligned} \tag{3.46}$$

porque $\mathrm{E}\left[(Y_i - \mu_i)(Y_l - \mu_l)\right] = 0$ cuando $i \neq l$ por la independencia de las Y_i; y como $\mathrm{E}\left[(Y_i - \mu_i)^2\right] = \operatorname{Var}[Y_i]$, $1 \le i \le n$, concluimos que

$$(\mathcal{I}(\beta))_{j,k} = \sum_{i=1}^{n} \frac{x_{i,j}x_{i,k}}{\operatorname{Var}[Y_i]\, g'(\mu_i)^2} = \frac{1}{\phi}\sum_{i=1}^{n} \frac{x_{i,j}x_{i,k}}{V(\mu_i)g'(\mu_i)^2},\ 1 \le j,k \le p \tag{3.47}$$

La matriz $\mathcal{I}(\beta)$ se puede obtener como

$$\mathcal{I}(\beta) = X^{\mathrm{t}} W X \tag{3.48}$$

donde W es la matriz diagonal de orden n

$$W = \operatorname{diag}\left\{ \frac{1}{\operatorname{Var}[Y_1]\, g'(\mu_1)^2}, \ldots, \frac{1}{\operatorname{Var}[Y_n]\, g'(\mu_n)^2} \right\} \tag{3.49}$$

3.3.1. Métodos numéricos

Método de Newton-Raphson

En general las ecuaciones $U_{\beta,j}(y) = 0$, $1 \leq j \leq p$, son no lineales y tienen que ser resueltas por métodos numéricos. Para resolverlas, nos apoyaremos en los métodos de Newton-Raphson presentados en el Apéndice A.1.1. El objetivo será obtener iterativamente aproximaciones $\hat{\beta}^{(m)}$ que convergen a una solución real, $\hat{\beta}$, según la ecuación

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} - \left(\frac{\partial^2 \ell(\beta; y)}{\partial \beta_j \partial \beta_k} \right)^{-1}_{\beta=\hat{\beta}^{(m-1)}} U^{(m-1)} \tag{3.50}$$

siendo $U^{(m-1)} = (U_{\beta,1}(y), \ldots, U_{\beta,p}(y))^{\mathrm{t}}{}_{\beta=\hat{\beta}^{(m-1)}}$.

Bajo ciertas condiciones el método es convergente y por tanto $\hat{\beta}^{(m)} \to \hat{\beta}$.

Método del tanteo

Podemos sustituir la matriz de segundas derivadas parciales por su esperanza, que aplicando la Proposición A.21 tenemos la igualdad

$$\mathrm{E}\left[\left(\frac{\partial^2 \ell(\beta; y)}{\partial \beta_j \partial \beta_k} \right) \right] = -\mathcal{I}(\beta). \tag{3.51}$$

Luego el método del tanteo queda parecido al de Newton-Raphson pero con la matriz de información

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + \mathcal{I}(\beta)^{-1}_{\beta=\hat{\beta}^{(m-1)}} U^{(m-1)} \tag{3.52}$$

siendo $U^{(m-1)} = (U_{\beta,1}(y), \ldots, U_{\beta,p}(y))^{\mathrm{t}}{}_{\beta=\hat{\beta}^{(m-1)}}$ y asumiendo que $\mathcal{I}(\beta)$ es una matriz invertible.

3.3.2. Comportamiento asintótico

Para obtener la distribución asociada al estimador máximo verosímil de β, juegan un papel muy importante las cantidades $U_{\beta,j}(y)$, $1 \leq j \leq p$.

La matriz de información $\mathcal{I}(\beta)$ es semidefinida positiva (ver Definición A.4) por ser matriz de covarianzas. Si además la suponemos no singular, sería definida positiva (ver Definición A.5) y podemos definir $\Psi^{\frac{1}{2}}$ tal que $\Psi^{\frac{1}{2}}\Psi^{\frac{1}{2}} = \Psi = \mathcal{I}(\beta)^{-1}$ y se verifica $\Psi U_\beta(y) \sim \mathrm{N}_p(0, I_p)$ (asintóticamente cuando $n \to \infty$) porque el vector $U_\beta(y)$ es siempre de media nula como se demuestra en la Proposición A.20 y su matriz de covarianzas es la matriz de información.

Luego por la relación entre las distribuciones normal y ji-cuadrado (ver Proposición A.4), podemos formular la siguiente proposición:

PROPOSICIÓN 3.4. *Asintóticamente cuando* $n \to \infty$, $U_\beta(y) \sim \mathrm{N}_p\,(0, \mathcal{I}(\beta))$ *y, por tanto,* $U_\beta(y)^{\,\mathrm{t}}\, \Psi U_\beta(y) \sim \chi^2(p)$.

COROLARIO 3.1. *Asintóticamente cuando* $n \to \infty$, $\hat{\beta} \sim \mathrm{N}_p\,(\beta, \mathcal{I}(\beta)^{-1})$ *y, por tanto,* $(\hat{\beta} - \beta)^{\,\mathrm{t}}\, \mathcal{I}(\beta)(\hat{\beta} - \beta) \sim \chi^2(p)$.

Demostración. La demostración completa contiene elementos de teoría asintótica multivariante no estudiada en el Grado en Estadística. Una introducción a esta se apoya en los siguientes resultados:

Que $\hat{\beta} \sim \mathrm{N}_p\,(\beta, \mathcal{I}(\beta)^{-1})$ puede verse como $\mathcal{I}^{\frac{1}{2}}(\beta)(\hat{\beta} - \beta) \sim \mathrm{N}_p\,(0, I_p)$.

Si $U_\beta(y) = (U_{\beta,1}(y), \ldots, U_{\beta,p}(y))^{\,\mathrm{t}}$ es el vector *scrore* para muestras y de tamaño n, se puede escribir el desarrollo de Taylor de $U_\beta(y)$ como

$$U_\beta(y) = 0 + \mathrm{H}_{\ell(\beta;y)}(\hat{\beta})(\beta - \hat{\beta}) + \frac{1}{2}(\beta - \hat{\beta})^{\,\mathrm{t}}\, f(\hat{\beta})(\beta - \hat{\beta}) \tag{3.53}$$

donde

$$\mathrm{H}_{\ell(\beta;y)} = \begin{bmatrix} \frac{\partial^2 \ell(\beta;y)}{\partial\beta_1\partial\beta_1} & \cdots & \frac{\partial^2 \ell(\beta;y)}{\partial\beta_1\partial\beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell(\beta;y)}{\partial\beta_p\partial\beta_1} & \cdots & \frac{\partial^2 \ell(\beta;y)}{\partial\beta_p\partial\beta_p} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 \log(\prod_{i=1}^n p(y_i;\beta))}{\partial\beta_1\partial\beta_1} & \cdots & \frac{\partial^2 \log(\prod_{i=1}^n p(y_i;\beta))}{\partial\beta_1\partial\beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \log(\prod_{i=1}^n p(y_i;\beta))}{\partial\beta_p\partial\beta_1} & \cdots & \frac{\partial^2 \log(\prod_{i=1}^n p(y_i;\beta))}{\partial\beta_p\partial\beta_p} \end{bmatrix}, \tag{3.54}$$

$\mathrm{H}_{\ell(\beta;y)}(\hat{\beta})$ es dicho hessiano evaluado en $\hat{\beta}$ y f es una función valorada en el espacio de las matrices cuadradas de orden p.

Como $\hat{\beta} \to \beta$ cuando $n \to \infty$, podemos asumir que $\frac{1}{2}(\beta - \hat{\beta})^{\,\mathrm{t}}\, f(\hat{\beta})(\beta - \hat{\beta}) \to 0$. Si además sustiuimos la matriz hessiano por su esperanza

$$\mathcal{I}^{-1}(\hat{\beta})U_\beta(y) = (\beta - \hat{\beta}) \tag{3.55}$$

Como $U_\beta(y) \sim \mathrm{N}_p\,(0_p, \mathcal{I}(\beta))$ cuando $n \to \infty$, se tiene que

$$\hat{\beta} - \beta \sim \mathrm{N}_p\left(0_p, \mathcal{I}^{-1}(\beta)\right) \tag{3.56}$$

y por tanto $\mathcal{I}^{\frac{1}{2}}(\beta)(\hat{\beta} - \beta) \sim \mathrm{N}_p\,(0, I_p)$ queda probado, luego es inmediato que $(\hat{\beta} - \beta)^{\,\mathrm{t}}\, \mathcal{I}(\beta)(\hat{\beta} - \beta) \sim \chi^2(p)$. □

DEFINICIÓN 3.7 (Estadístico de Wald). Llamamos estadístico de Wald a $(\hat{\beta} - \beta)^{\,\mathrm{t}}\, \mathcal{I}(\beta)(\hat{\beta} - \beta)$.

Observación. Para modelos lineales normales los resultados anteriores son exactos.

Observación. En la práctica no trabajaremos con $\mathcal{I}(\beta)$ sino con $\mathcal{I}(\hat{\beta})$.

Proposición 3.5. *Sean $h(x)$ la función de densidad de probabilidad de la distribución $\mathrm{N}(0,1)$ y $z_{\alpha/2}$ un valor tal que $\int_{z_{\alpha/2}}^{\infty} h(x)\,\mathrm{d}\,x = \frac{\alpha}{2}$. Sea la matriz $\Psi = \mathcal{I}(\beta)^{-1}$ cuyos coeficientes denotamos $\psi_{j,k}$, $1 \leq j,k \leq p$. Un intervalo de confianza para β_j a un nivel $1-\alpha$ es*

$$I_{1-\alpha}(\beta_j) = [\hat{\beta}_j - z_{\alpha/2}\sqrt{\psi_{j,j}}, \hat{\beta}_j + z_{\alpha/2}\sqrt{\psi_{j,j}}] \tag{3.57}$$

3.4. Contraste de hipótesis

3.4.1. Bondad de ajuste

Consideremos un modelo lineal generalizado como el dado en la Definición 3.4. Llamamos bondad de ajuste a la valoración de lo adecuado que es un modelo para describir un conjunto de datos. El primer contraste que consideraremos será el de bondad de ajuste siendo

$$\begin{cases} \mathrm{H}_0: & \text{el modelo presenta un buen ajuste.} \\ \mathrm{H}_1: & \text{el modelo no presenta un buen ajuste.} \end{cases} \tag{3.58}$$

Hasta el momento, hemos trabajado siempre con una cantidad de parámetros a estimar, p, a lo sumo igual al número de observaciones independientes n. Diremos que un modelo es saturado si contiene el número máximo de parámetros que podemos comparar, *i.e.*, $p = n$. Generalmente, trabajaremos con modelos donde $p < n$. En la siguiente sección estudiaremos la validez de estos modelos propuestos frente al modelo saturado mediante el estadístico de razón de verosimilitudes y el concepto de devianza[3].

Definición 3.8 (Estadístico de razón de verosimilitudes). Sea $\beta^{\,\mathrm{s}}$ el vector de parámetros para un modelo saturado y su estimación es $\hat{\beta}^{\,\mathrm{s}}$. Sea $\hat{\beta}$ la estimación de máxima verosimilitud en el modelo propuesto. Llamamos estadístico de razón de verosimilitudes a

$$\rho = \frac{L(\hat{\beta}^{\,\mathrm{s}}; y)}{L(\hat{\beta}; y)}. \tag{3.59}$$

Observación. Es más común trabajar con el logaritmo de ρ, al que llamaremos estadístico de diferencia de log-verosimilitudes, $\log(\rho) = \ell(\hat{\beta}^{\,\mathrm{s}}; y) - \ell(\hat{\beta}; y)$.

[3]Traducción del inglés *deviance*, literalmente desviación.

Podemos entender $\log(\rho)$ como una medida de desviación entre el modelo propuesto y el saturado. Valores altos de $\log(\rho)$ indican que el modelo en el que estamos trabajando se separa bastante del saturado y, por tanto, se rechaza H_0. Para trabajar estadísticamente esta idea, desarrollamos el concepto de devianza.

DEFINICIÓN 3.9 (Devianza). Consideremos un modelo lineal generalizado con vector de parámetros β y su estimador de máxima verosimilitud, $\hat{\beta}$. Sea $\beta^{\,s}$ el vector de parámetros de un modelo saturado y $\hat{\beta}^{\,s}$ su estimador de máxima verosimilitud. Llamamos devianza del modelo a $D = 2(\ell(\hat{\beta}^{\,s}; y) - \ell(\hat{\beta}; y))$.

Sumando y restando en D se tiene

$$\begin{aligned} D &= 2(\ell(\hat{\beta}^{\,s}; y) - \ell(\beta^{\,s}; y)) \\ &= 2[(\ell(\hat{\beta}^{\,s}; y) - \ell(\beta^{\,s}; y)) - (\ell(\hat{\beta}; y) - \ell(\beta; y)) + (\ell(\beta^{\,s}; y) - \ell(\beta; y))] \\ &= 2(\ell(\hat{\beta}^{\,s}; y) - \ell(\beta^{\,s}; y)) - 2(\ell(\hat{\beta}; y) - \ell(\beta; y)) + 2(\ell(\beta^{\,s}; y) - \ell(\beta; y)) \end{aligned} \tag{3.60}$$

Para obtener la distribución asintótica de D, basta tener en cuenta la distribución asintótica de los sumandos en el último término de la ecuación (3.60). En general, $2(\ell(\hat{\beta}; y) - \ell(\beta; y)) \sim \chi^2(p)$ asintóticamente cuando $n \to \infty$. En efecto, en nuestra justificación de la distribución de $\hat{\beta}$ obtuvimos que

$$\ell(\beta; y) - \ell(\hat{\beta}; y) = -\frac{1}{2}(\beta - \hat{\beta})^{\,t}\, \mathcal{I}(\hat{\beta})(\beta - \hat{\beta}), \tag{3.61}$$

luego $2(\ell(\hat{\beta}; y) - \ell(\beta; y)) \sim \chi^2(p)$ asintóticamente. Tenemos en cada sumando:

- $2(\ell(\hat{\beta}^{\,s}; y) - \ell(\beta^{\,s}; y)) \sim \chi^2(n)$ aproximadamente, siendo n el número de parámetros en el modelo saturado, es decir, el número de observaciones.
- $2(\ell(\hat{\beta}; y) - \ell(\beta; y)) \sim \chi^2(p)$ aproximadamente, siendo p el número de parámetros en el modelo de interés.
- $\nu = 2(\ell(\beta^{\,s}; y) - \ell(\beta; y))$ es una constante positiva que se aproxima a 0 cuando el modelo de interés se ajusta a los datos tan bien como el saturado.

Con esto, se tiene que, bajo ciertas condiciones que garanticen la independencia de los sumandos en la ecuación (3.60), si el modelo propuesto es correcto, $D \sim \chi^2(n - p)$ asintóticamente, de manera que, dada una significación $\alpha \in]0, 1[$, un valor de D superior a $\chi^2_\alpha(n - p)$ nos lleva a rechazar H_0, *i.e.*, el modelo de p no ajustaría bien los datos.

EJEMPLO 3.2. Sean $Y_i \sim \text{Poisson}\,(\lambda_i)$, $1 \leq i \leq n$, variables aleatorias independientes. Su función de log-verosimilitud en términos de $\lambda = (\lambda_1, \ldots, \lambda_n)^{\,\mathrm{t}}$ es

$$\ell(\lambda; y) = \sum_{i=1}^{n} y_i \log(\lambda_i) - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \log(y_i!) \tag{3.62}$$

En el caso de un modelo saturado, todos los λ_i son diferentes y el vector de parámetros del modelo es $\beta^{\,\mathrm{s}} = \lambda = (\lambda_1, \ldots, \lambda_n)^{\,\mathrm{t}}$. La estimación de máxima verosimilitud de cada λ_i es $\hat{\lambda}_i = y_i$, luego el valor máximo de la función de log-verosimilitud del modelo saturado se alcanza en

$$\begin{aligned} \ell(\hat{\beta}^{\,\mathrm{s}}; y) &= \sum_{i=1}^{n} y_i \log(\hat{\lambda}_i) - \sum_{i=1}^{n} \hat{\lambda}_i - \sum_{i=1}^{n} \log(y_i!) \\ &= \sum_{i=1}^{n} y_i \log(y_i) - \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!) \end{aligned} \tag{3.63}$$

Si el modelo de interés contiene p parámetros, $p < n$, el estimador de máxima verosimilitud, $\hat{\lambda} \in \mathbb{R}^p$, nos permite estimar los λ_i y obtener los valores ajustados, $\hat{\lambda}_i = g^{-1}(x_i{}^{\,\mathrm{t}}\,\beta)$, que coinciden con los valores estimados de los parámetros porque la estimación es insesgada: $\hat{y}_i = \hat{\lambda}_i$. Llegamos a

$$\ell(\hat{\beta}; y) = \sum_{i=1}^{n} y_i \log(\hat{y}_i) - \sum_{i=1}^{n} \hat{y}_i - \sum_{i=1}^{n} \log(y_i!) \tag{3.64}$$

y obtenemos la devianza

$$D = 2(\ell(\hat{\beta}^{\,\mathrm{s}}; y) - \ell(\hat{\beta}; y)) = 2(\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - \sum_{i=1}^{n} (y_i - \hat{y}_i)). \tag{3.65}$$

3.4.2. Contraste de hipótesis sobre β

El problema de estimación en un modelo lineal generalizado puede verse como un problema de reducción dimensional, de n a p parámetros. Esta reducción acarrea una perdida de calidad en nuestras estimaciones a medida que reducimos p.

Si consideramos un modelo lineal generalizado con $\beta \in \mathbb{R}^p$, dado $q < p$, podremos contrastar[4]

$$\begin{cases} \mathrm{H}_0 : \beta = \beta_{(0)} (= (\beta_1, \ldots, \beta_q, 0, \ldots, 0)^{\,\mathrm{t}}) \\ \mathrm{H}_1 : \beta = \beta_{(1)} (= (\beta_1, \ldots, \beta_q, \beta_{q+1} \ldots, \beta_p)^{\,\mathrm{t}}) \end{cases} \tag{3.66}$$

[4]No hemos señalado ningún orden necesario en las coordenadas de β, luego podemos plantear el modelo dejando como últimas las $p - q$ coordenadas objeto de contraste, por sencillez en la formulación.

lo que intuitivamente nos viene a decir que, supuesta H_0 cierta, bastarían q parámetros para ajustar los datos al modelo elegido. Aceptar H_1 nos diría que habernos desprendido de esos $p-q$ parámetros resulta estadísticamente significativo.

Equivalentemente podremos reformular:

$$\begin{cases} \mathrm{H}_0 : \beta_{q+1} = \ldots = \beta_p = 0 \\ \mathrm{H}_1 : \beta_j \neq 0 \text{ para algún } j \in \{q+1, \ldots, p\} \end{cases} \tag{3.67}$$

Se tiene, por tanto, que cada hipótesis indica un modelo, con distribuciones de la misma familia y con idéntica función de enlace, diferenciándose exclusivamente en el número de parámetros: el modelo original con p parámetros y el modelo reducido con q parámetros. Se comparan los estadísticos de bondad de ajuste para ambos modelos. Sean

- D_0: devianza asociada al modelo reducido por H_0.
- D_1: devianza asociada al modelo original.
- $\Delta D = D_0 - D_1 = 2(\ell(\hat{\beta}_{(1)}; y) - \ell(\hat{\beta}_{(0)}; y))$, que es una medida de discrepancia o desviación de un modelo frente a otro.

Si H_0 es cierta, entonces, asintóticamente, $D_{(0)} \sim \chi^2(n-q)$ y $D_{(1)} \sim \chi^2(n-p)$ según vimos en el Capítulo 3.4.1. Bajo ciertas condiciones de independencia, de lo anterior, se obtiene que, asintóticamente cuando $n \to \infty$, $\Delta D \sim \chi^2(p-q)$, supuesta H_0. Por tanto, se rechaza H_0 al nivel α ($\alpha \in]0,1[$) si $\Delta D > \chi_\alpha(p-q)$.

3.5. Modelos de regresión

El modelo lineal generalizado presenta notables ventajas a la hora de construir modelos de regresión en aquellos casos que escapan a las hipótesis del modelo lineal normal. En esta sección, presentaremos una serie de modelos de regresión, que son casos particulares del modelo lineal generalizado, entre ellos: regresión normal, regresión logística binaria, regresión de Poisson y regresión exponencial. Dobson (2002) [3] dedica sendos capítulos a desarrollar tanto desde un punto de vista teórico como aplicado cada uno.

Definición 3.10 (Devianza naíf). Sea $G^{\mathrm{s}}(\mathrm{E}[Y]) = X\beta^{\mathrm{s}}$ un modelo saturado ($p = n$) y sea $G^0(\mathrm{E}[Y]) = \beta^0 1_n = (\beta^0, \ldots, \beta^0)^{\mathrm{t}}$ el modelo naíf ($p = 1$). Llamamos devianza naíf a

$$D_0 = 2(\ell(\beta^{\mathrm{s}}; y) - \ell(\beta^0; y)). \tag{3.68}$$

donde $1_n = (1, \ldots, 1)^{\mathrm{t}}$ de dimensión n.

Definición 3.11 (Devianza explicada). Sea $G(\mathrm{E}[Y]) = X\beta$ un modelo lineal generalizado. Llamamos devianza explicada al valor

$$R^2 = 1 - \frac{D}{D_0}. \tag{3.69}$$

De hecho, se puede entender la devianza explicada como una generalización del coeficiente de determinación al modelo lineal generalizado y coincide con este en cuando $g \equiv \mathrm{id}$.

3.5.1. Regresión normal (varianza común conocida)

Consideramos las variables aleatorias Y_i, las cuales tienen función de densidad de probabilidad

$$p(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}}. \tag{3.70}$$

Coincide con el modelo de regresión lineal normal visto en el Capítulo 2.3. Resulta inmediato que es un modelo lineal generalizado visto el Ejemplo 3.1.

3.5.2. Regresión logística binaria

Consideremos las variables aleatorias independientes $Y_i \sim \mathrm{b}_1(\pi_i)$ cuyas funciones de masa de probabilidad serán

$$p(y_i; \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1 - y_i} \tag{3.71}$$

y donde $\mu_i = \pi_i$, $1 \leq i \leq n$. Sean $X_1 \ldots, X_p$, p variables aleatorias observadas n veces, de modo que el modelo de regresión logística binaria viene dado por la ecuación

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = g(\pi_i) = {x_i}^{\mathrm{t}}\beta,\ 1 \leq i \leq n, \tag{3.72}$$

donde ${x_i}^{\mathrm{t}} = [1, x_{i,1}, \cdots, x_{i,p}]$ es la i-ésima fila de la matriz X (compuesta por 1_n en su primera columna, y las n observaciones de las p variables regresoras) y la función de enlace es $g : x \in [0, 1] \to \log(x/(1 - x)) \in \mathbb{R}$.

Es más fácil entender el modelo como

$$\frac{\pi_i}{1 - \pi_i} = \mathrm{e}^{{x_i}^{\mathrm{t}}\beta},\ 1 \leq i \leq n, \tag{3.73}$$

o incluso mejor aún como

$$\pi_i = \frac{\mathrm{e}^{x_i{}^{\mathrm{t}}\beta}}{1+\mathrm{e}^{x_i{}^{\mathrm{t}}\beta}} = \frac{\mathrm{e}^{\eta_i}}{1+\mathrm{e}^{\eta_i}},\ 1 \leq i \leq n. \tag{3.74}$$

Definición 3.12 (*Odd*). Sea un cierto suceso A, con probabilidad de ocurrencia $P(A) = \pi \in]0,1[$. Llamamos *odd*[5] del suceso A al valor real positivo

$$\mathrm{odd}\,(A) = \frac{P(A)}{P(\bar{A})} = \frac{\pi}{1-\pi}. \tag{3.75}$$

A partir de esta definición, la ecuación (3.73) nos dice que a $\mathrm{odd}\,(\pi_i) = \exp(x_i{}^{\mathrm{t}}\beta)$, $1 \leq i \leq n$. El modelo puede reescribirse como $g(\pi_i) = \log(\mathrm{odd}\,(\pi_i)) = x_i{}^{\mathrm{t}}\beta$. Indica que si aumentamos en una unidad la variable X_j, el logaritmo de *odds*, aumenta en β_j.

3.5.3. Regresión de Poisson

Consideremos las variables aleatorias independientes $Y_i \sim \mathrm{Poisson}\,(\lambda_i)$, de modo que escribiremos la función de masa de probabilidad de cada una como

$$p(y_i;\lambda_i) = \mathrm{e}^{-\lambda_i}\frac{\lambda_i}{y_i!} \tag{3.76}$$

con $y_i \in \mathbb{Z}_0^+$ y $\lambda_i \in \mathbb{R}^+$, $1 \leq i \leq n$. Los modelos de regresión de Poisson aparecerán ante la necesidad de modelizar un proceso de contaje, por lo que la variable respuesta deberá tomar, necesariamente, valores enteros positivos.

Ejemplo 3.3. Tenemos una muestra de n tortugas hembra de las que se mide una serie de variables relacionadas con su entorno (temperatura del agua, concentración de oxígeno en agua, concentración de tóxicos, etc.) y el número de huevos puesto en la última temporada de desove. La variable Y_i mide el número de huevos puesto por la tortuga i-ésima. Podemos preguntarnos si dada una función de enlace g, es posible modelizar $g(\lambda_i) = x_i{}^{\mathrm{t}}\beta$.

Crecimiento lineal

Supongamos que la media de las variables Y_i mantiene una relación lineal con respecto a una variable aleatoria Z observada en $z = (z_1,\ldots,z_n)^{\mathrm{t}}$ y llamamos $\mathrm{E}\,[Y_i] = \lambda_i = \beta_0 + \beta_1 z_i$. Para este caso tendremos un vector de parámetros $\beta = (\beta_0,\beta_1) \in \mathbb{R}^2$, y una matriz de

[5] A veces traducido como ventaja.

constantes X de la forma

$$X = \begin{bmatrix} x_1{}^{\mathrm{t}} \\ \vdots \\ x_n{}^{\mathrm{t}} \end{bmatrix} = \begin{bmatrix} 1 & z_1 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix} \tag{3.77}$$

Obtenemos su función de log-verosimilitud, $\ell(\beta; y) = \sum_i (y_i \log(x_i{}^{\mathrm{t}}\,\beta) - x_i{}^{\mathrm{t}}\,\beta - \log(y_i!))$. El modelo anterior es un modelo lineal generalizado con función de enlace $g \equiv \mathrm{id}$.

Crecimiento exponencial

Supongamos que la media de las variables Y_i mantiene una relación exponencial con respecto a una variable real Z observada en $(z_1, \ldots, z_n)^{\mathrm{t}}$ y llamamos $\mathrm{E}[Y_i] = \lambda_i = \alpha z_i^{\beta_1}$, para cualquier $\alpha > 0$ y β real. El modelo de regresión de Poisson con crecimiento exponencial viene dado por

$$g(\lambda_i) = \log(\lambda_i) = \log(\alpha z_i^{\beta_1}) = \log(\alpha) + \beta_1 \log(z_i),\ 1 \leq i \leq n. \tag{3.78}$$

Si llamamos $\beta_0 = \log(\alpha)$, la matriz de modelo, X, será de la forma

$$X = \begin{bmatrix} 1 & \log(z_1) \\ \vdots & \vdots \\ 1 & \log(z_n) \end{bmatrix} \tag{3.79}$$

y el vector de parámetros será $\beta = (\beta_0, \beta_1) = (\log(\alpha), \beta_1)$.

Expresamos su función de log-verosimilitud como $\ell(\beta; y) = \sum_{i=1}^n [\lambda_i + y_i \log(\lambda_i) - \log(y_i!)] = \sum_{i=1}^n [\mathrm{e}^{\beta_0} z_i^{\beta_1} + y_i \beta_0 + y_i \beta_1 \log(z_i) - \log(y_i!)]$.

3.5.4. Regresión exponencial

Consideremos las variables aleatorias independientes $Y_i \sim \mathrm{Exp}(\lambda_i)$, de modo que escribiremos la función de masa de probabilidad de cada una como

$$p(y_i; \lambda_i) = \lambda_i \mathrm{e}^{-\lambda_i y_i} \tag{3.80}$$

con $y_i \geq 0$ y $\lambda_i > 0$, $1 \leq i \leq n$. Supongamos que la media de las variables Y_i mantiene una relación exponencial con respecto a la combinación lineal de p variables aleatorias X_j, $1 \leq j \leq p$. El modelo de regresión exponencial viene dado por

$$\log(\lambda_i^{-1}) = -\log(\lambda_i) = g(\mu_i) = x_i{}^{\mathrm{t}}\,\beta,\ 1 \leq i \leq n. \tag{3.81}$$

La regresión exponencial se emplea para el estudio de modelos de supervivencia, tiempos de respuesta a un fármaco entre pacientes, etc.

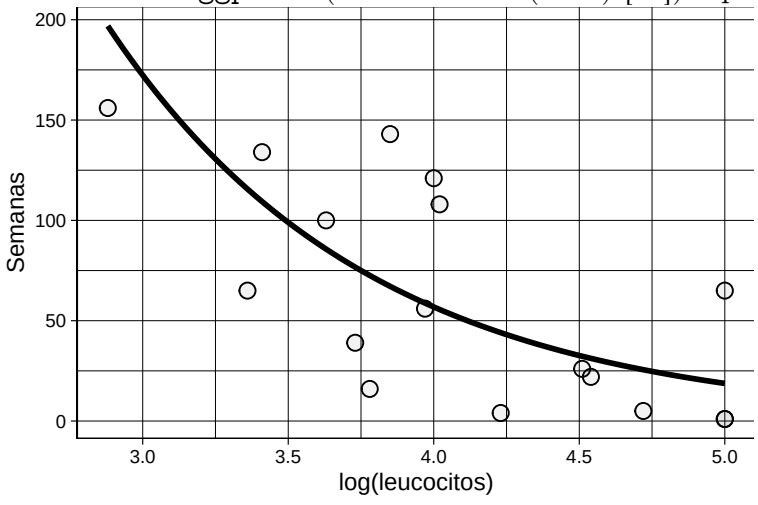
Ejemplo 3.4. Desarrollamos un ejemplo detalladamante en el Capítulo 3.6.

3.6. Ejemplo

El siguiente ejemplo viene dado por Cox y Snell (1981) [2] y es reproducido por Dobson (2002) [3]. Se pretende modelizar, para enfermos de leucemia, la relación existente entre el número de semanas transcurridas hasta la muerte de un paciente desde su diagnóstico (Y) y el logaritmo en base diez del número inicial de leucocitos que presentaba (X_1). Para ello se han tomado 17 pacientes con leucemia, cuyos datos recogemos en el siguiente código de `R`:

```
x <- c(3.36, 2.88, 3.63, 3.41, 3.78, 4.02, 4.00, 4.23, 3.73,
      3.85, 3.97, 4.51, 4.54, 5.00, 5.00, 4.72, 5.00)
y <- c(65, 156, 100, 134, 16, 108, 121, 4, 39,
      143, 56, 26, 22, 1, 1, 5, 65)
datos = data.frame(x,y)
summary(datos)
##        x               y
##  Min.   :2.880   Min.   :  1.00
##  1st Qu.:3.730   1st Qu.: 16.00
##  Median :4.000   Median : 56.00
##  Mean   :4.096   Mean   : 62.47
##  3rd Qu.:4.540   3rd Qu.:108.00
##  Max.   :5.000   Max.   :156.00
```

Con ayuda de la biblioteca `ggplot2` (ver Wickham (2016) [15]) representamos:



Las semanas indican tiempo entre diagnóstico y defunción.

La distribución exponencial es utilizada habitualmente para describir tiempos de supervivencia: $Y_i \sim \text{Exp}(\lambda_i)$, *i.e.*, $p(y_i;\lambda_i) = \lambda_i e^{-y_i\lambda_i}$, $1 \leq i \leq 17$.

Una posible especificación es $\text{E}[Y_i] = \exp(\beta_0 + \beta_1 x_i)$, $1 \leq i \leq 17$, que asegura que $\text{E}[Y_i]$ es no negativa para todos los valores de los parámetros y todos los valores de x. Para este caso, la función de enlace es $g \equiv \log$. Es inmediato ver que este ejemplo se corresponde con el descrito en el Capítulo 3.5.4 y el modelo de regresión queda

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i,1}. \tag{3.82}$$

Estimación de β. Fijado el modelo de regresión de la ecuación (3.82). La función `glm()` estima modelos lineales generalizados dada una familia de distribuciones de tipo exponencial.

```
regr <- glm(y~x,family=Gamma(link="log"),data=datos)
```

En nuestro caso `family=Gamma` porque la distribución exponencial es un caso particular de la distribución gamma (ver Proposición A.11) y señalamos la función de enlace logaritmo, `link=log`. Además el método de evaluación para $\hat{\beta}$ es mediante el método del tanteo (ver Capítulo 3.3.1). Sea pues,

```
summary(regr)
##
## Call:
## glm(formula = y ~ x, family = Gamma(link = "log"), data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9922  -1.2102  -0.2242   0.2102   1.5646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.4775     1.6034   5.287 9.13e-05 ***
## x            -1.1093     0.3872  -2.865   0.0118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.9388638)
##
##     Null deviance: 26.282  on 16  degrees of freedom
```

```
## Residual deviance: 19.457  on 15  degrees of freedom
## AIC: 173.97
##
## Number of Fisher Scoring iterations: 8
```

Llegamos a que $\hat{\beta} = (8.4775, -1.1093)^{\mathrm{t}}$ después de ocho iteraciones, luego un intervalo de confianza para $1 - \alpha = 0.95$ en cada parámetro es, de acuerdo con la Proposición 3.5, $I_{1-\alpha}(\beta_j) = [\beta_j - z_{\frac{\alpha}{2}}\sqrt{\psi_{j,j}}, \beta_j + z_{\frac{\alpha}{2}}\sqrt{\psi_{j,j}}]$:

- $I_{0.95}(\beta_0) = 8.4775 \mp 1.96 \times 1.6034 = [5.3348, 11.620]$ y
- $I_{0.95}(\beta_1) = -1.1093 \mp 1.96 \times 0.3872 = [-1.8682, -0.3504]$.

Bondad de ajuste. Para contrastar la buena adecuación de los datos al modelo,

$$\begin{cases} \mathrm{H}_0: & \text{Los datos se ajustan al modelo propuesto.} \\ \mathrm{H}_1: & \text{Los datos no se ajustan al modelo propuesto.} \end{cases} \tag{3.83}$$

nos servimos de la devianza (`residual deviance`), $D_{\exp} = 19.457$. Bajo la hipótesis nula de buena adecuación, $D \sim \chi^2(15)$.

```
qchisq(0.95,15)
## [1] 24.99579
```

luego $D_{\exp} < \chi^2_{1-\alpha} = 24.99579$ para $\alpha = 0.05$, luego asumimos H_0.

Contraste de hipótesis. Sea el modelo original dado por el parámetro $\beta_{(1)} = (\beta_0, \beta_1)^{\mathrm{t}}$ y el reducido dado por $\beta_{(0)} = \beta_0$ respectivamente[6]. Sea el contraste

$$\begin{cases} \mathrm{H}_0: \beta_{(0)} = \beta_{(1)} \\ \mathrm{H}_1: \beta_{(0)} \neq \beta_{(1)} \end{cases} \equiv \begin{cases} \mathrm{H}_0: \beta_1 = 0 \\ \mathrm{H}_1: \beta_1 \neq 0 \end{cases} \tag{3.84}$$

Formulamos este nuevo modelo en `R`:

```
naif <- glm(y~1,family=Gamma(link="log"), data=datos)
```

Vemos cómo la devianza residual (`residual deviance`) coincide con la naíf (`null deviance`).

[6] De hecho, el modelo reducido es el modelo naíf (ver Definición 3.10).

```
summary(naif)
##
## Call:
## glm(formula = y ~ 1, family = Gamma(link = "log"), data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5103  -1.1120  -0.1074   0.6023   1.0789
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.135      0.211   19.59 1.31e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.7570041)
##
##     Null deviance: 26.282  on 16  degrees of freedom
## Residual deviance: 26.282  on 16  degrees of freedom
## AIC: 178.09
##
## Number of Fisher Scoring iterations: 6
```

En este caso concreto no hubiera sido necesario formular este modelo, ya que el valor de la devianza naíf ya fue obtenido en el modelo `regr`. Se plantea para ver explícitamente cómo de calcularía para dos modelos, original vs. reducido, cualesquiera.

El estadístico de contraste es $\Delta D = D_\omega - D_\Omega = 26.282 - 19.457 = 6.825$. Bajo hipótesis nula, $\Delta D \sim \chi^2(1)$. $\Delta D = 6.825 > 3.841459 = \chi^2_\alpha(1)$ para $\alpha = 0.05$, siendo $\chi^2_\alpha(1)$ un valor tal que si $h(x)$ es la función de densidad de la distribución $\chi^2(1)$, entonces $\int_{\chi^2_\alpha(1)}^{\infty} h(x)\,\mathrm{d}\,x = \alpha$; luego concluimos que sí hay diferencias significativas entre el modelo de dos y un parámetro.

Capítulo 4

Modelo aditivo generalizado

Los modelos aditivos generalizados deben su nombre a que son la convergencia de los modelos aditivos y los modelos lineales generalizados. De hecho, estas dos familias de modelos son las dos maneras que tenemos de extender el modelo lineal. Si en los modelos lineales generalizados optamos por complejizar la relación entre el valor esperado de la respuesta y las variables regresoras añadiendo una función de enlace g y ampliando la familia de distribuciones admitidas por la variable respuesta; en los modelos aditivos consideraremos la variable objetivo sin ninguna distribución impuesta, que depende de las variables explicativas mediante funciones de suavizado (no lineales). La propuesta de los modelos aditivos generalizados son la síntesis de estos dos y es planteada inicialmente por Hastie y Tibshirani (1986) [6].

4.1. Modelo aditivo

4.1.1. Suavizado

Una variable

Sean Y un vector aleatorio n-dimensional de distribución de probabilidad desconocida, cuyas coordenadas Y_i, $1 \leq i \leq n$, son independientes, x un vector n dimensional de constantes conocidas, $f : \mathbb{R} \to \mathbb{R}$ una función desconocida que llamaremos de suavizado desconocida y ε un vector aleatorio de media nula y matriz de covarianzas $\sigma^2 I_n$. El problema de suavizado en una variable consistirá en hallar una estimación de f de modo

que

$$Y_i = f(x_i) + \varepsilon_i,\ 1 \leq i \leq n. \tag{4.1}$$

Bases de funciones. Sean $b_1, \ldots, b_k$ una base de funciones de modo que toda función f puede ser representada como una combinación lineal de los elementos de la base:

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j \tag{4.2}$$

donde los β_j son k parámetros desconocidos.

Ejemplo 4.1. Sea $\mathbb{P}_n[x]$ el espacio de los polinomios de orden n, el cual tiene dimensión $n+1$ y su base canónica es $x^n, x^{n-1}, \ldots, x, 1$. Si suponemos el caso, por ejemplo, de $n = 4$, se tiene que escribimos el modelo como $y_i = \beta_1 + x_i\beta_2 \ldots + x_i^4\beta_4 + \varepsilon_i$.

El problema de trabajar con bases de polinomios está en que solo nos dará ajustes puntuales, no del comportamiento general de un conjunto amplio de datos. Además, si extendemos el problema a varias variables, como haremos más adelante, y asumimos la posibilidad de interacción entre las variables, el número de sumandos podría llegar a aumentar incluso hasta superar el tamaño de la muestra, n (ver Faraway (2016) [4]).

Supongamos ahora que se tienen k nodos (números reales, partición de $\mathbb{R}$) ordenados de menor a mayor, $x_1^*, \ldots, x_k^*$, luego tenemos una partición del intervalo $[x_1^*, x_k^*]$ en $k-1$ tramos. Definimos la base de funciones lineales a tramos

$$b_j(x) = \begin{cases} \frac{x - x_{j-1}^*}{x_j^* - x_{j-1}^*}, & x \in]x_{j-1}^*, x_j^*] \\ \frac{x_{j+1}^* - x}{x_{j+1}^* - x_j^*}, & x \in]x_j^*, x_{j+1}^*[\\ 0, & \text{e.o.c. (en otro caso)} \end{cases} \tag{4.3}$$

si $2 \leq j \leq k-1$, y en los extremos se tiene

$$b_1(x) = \begin{cases} \frac{x_2^* - x}{x_2^* - x_1^*}, & x < x_2^* \\ 0, & \text{e.o.c.} \end{cases}, \quad b_k(x) = \begin{cases} \frac{x - x_{k-1}^*}{x_k^* - x_{k-1}^*}, & x > x_{k-1}^* \\ 0, & \text{e.o.c.} \end{cases} \tag{4.4}$$

Control del suavizado por penalización de la curvatura. Ya vimos en el Capítulo 2 que ante la ausencia de una distribución sobre Y, queríamos hallar la mejor solución $\hat{\beta}$ que minimizase $||Y - X\beta||_2^2$. No contaremos con esta expresión lineal $Y = X\beta\varepsilon$, aunque sí es un caso particular de la ecuación (4.1) con $x_{i,j} = b_j(x_i)$, y los mínimos cuadrados

vendrán penalizados, multiplicados por un factor λ que modula la influencia de esta corrección mediante penalizaciones

$$||Y - X\beta||_2^2 + \lambda \sum_{j=2}^{k-1} \left(f(x_{j-1}^*) - 2f(x_j^*) + f(x_{j+1}^*) \right)^2 . \tag{4.5}$$

El sumatorio denota una medida de *rugosidad* en tanto que es una suma cuadrática de las segundas diferencias de f en los nodos. De hecho, si la función es una recta (caso lineal) y los nodos son equiespaciados, el sumatorio se anula: en efecto, si $f(x) = T(x) + \alpha$, siendo T una aplicación lineal, α una constante y dos nodos consecutivos distan δ, se tiene

$$\begin{aligned} f(x_j^* - \delta) - 2f(x_j^*) + f(x_j^* + \delta) &= T(x_j^* - \delta) + \alpha - 2(T(x_j^*) + \alpha) + T(x_j^* + \delta) + \alpha \\ &= T(x_j^* - \delta) - 2T(x_j^*) + T(x_j^* + \delta) \\ &= T(x_j^* - \delta - 2x_j^* - 2\delta + x_j^* + \delta) \\ &= T(0) = 0 \end{aligned} \tag{4.6}$$

y tendríamos los mínimos cuadrados ordinarios.

Definición 4.1 (Parámetro de suavizado). Llamamos parámetro de suavizado[1] al valor real positivo $\lambda > 0$ que modula la influencia del sumatorio en el segundo sumando de la ecuación (4.5) en el ajuste.

La elección de λ nos llevará a modelos sobreajustados (más rugosos) o ajustes más lisos.

- Si $\lambda = 0$, no hay penalización de la curvatura y tendremos sobreajuste.
- Si $\lambda \to \infty$, hay una alta penalización de la curvatura y el ajuste será liso.

Sea D la matriz de dimensión $(k-2) \times k$

$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}, \tag{4.7}$$

podemos escribir, dado $\beta_j = f(x_j^*)$,

$$\sum_{j=2}^{k-1} \left(f(x_{j-1}^*) - 2f(x_j^*) + f(x_{j+1}^*) \right)^2 = \beta^{\mathrm{t}} D^{\mathrm{t}} D\beta = \beta^{\mathrm{t}} S\beta \tag{4.8}$$

[1]También llamado parámetro de ancho de banda o simplemente ancho de banda.

siendo $S = D^{\,t}\,D$ una matriz cuadrada de orden k. Queremos, por tanto, obtener los valores de β que minimicen $||Y - X\beta||_2^2 + \lambda\beta^{\,t}\,S\beta$, *i.e.*, el estimador por mínimos cuadrados penalizados es

$$\hat{\beta} = \arg_{\beta\in\mathbb{R}^p} \min ||Y - X\beta||_2^2 + \lambda\beta^{\,t}\,S\beta. \tag{4.9}$$

PROPOSICIÓN 4.1. *El estimador mínimo cuadrático (penalizado) de β es $\hat{\beta} = (X^{\,t}\,X + \lambda S)^{-1}X^{\,t}\,Y$.*

PROPOSICIÓN 4.2. *Se verifica*

$$\left|\left|\begin{bmatrix} Y \\ 0_{k-2} \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda}D \end{bmatrix}\beta\right|\right|_2^2 = ||Y - X\beta||_2^2 + \lambda\beta^{\,t}\,S\beta, \tag{4.10}$$

que es una expresión computacionalmente más eficiente de calcular el estimador $\hat{\beta}$.

Demostración.

$$\begin{aligned}\left|\left|\begin{bmatrix} Y \\ 0_{k-2} \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda}D \end{bmatrix}\beta\right|\right|_2^2 &= (Y^{\,t} - \beta^{\,t}\,X^{\,t}, 0_{k-2}{}^{\,t} - \beta^{\,t}\,\sqrt{\lambda}D^{\,t}) \begin{bmatrix} Y - X\beta \\ 0_{k-2} - \sqrt{\lambda}D\beta \end{bmatrix} = \\ &= (Y^{\,t} - \beta^{\,t}\,X^{\,t})(Y - X\beta) + \sqrt{\lambda}\sqrt{\lambda}\beta^{\,t}\,D^{\,t}\,D\beta = \\ &= (Y^{\,t} - X\beta)^{\,t}(Y - X\beta) + \lambda\beta^{\,t}\,S\beta = \\ &= ||Y - X\beta||_2^2 + \lambda\beta^{\,t}\,S\beta\end{aligned} \tag{4.11}$$

□

Elección de λ por validación cruzada. Ya indicamos que valores altos de λ nos conducirán a estimaciones más lisas de f, mientras que aquellos próximos a 0 se traducen en una estimación sobreajustada (más afectada por la rugosidad). Sea $\hat{f}$ la estimación de f mediante $\hat{f} \equiv \sum_j b_j\hat{\beta}_j$. Sea M un valor que mide la *distancia* entre la función f y su estimación $\hat{f}$,

$$M = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2 \tag{4.12}$$

con $f_i = f(x_i)$ y $\hat{f}_i = \hat{f}(x_i)$. Querremos que M sea lo menor posible.

DEFINICIÓN 4.2 (*Score* de validación cruzada). Sea $\hat{f}^{[-i]}$ el valor de $\hat{f}$ obtenido cuando ajustamos todas las observaciones salvo y_i, $1 \leq i \leq n$. Llamamos *score* de validación cruzada, y lo denotamos ν_0, al valor real positivo

$$\nu_0 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2 \tag{4.13}$$

Como $y_i = f_i + \varepsilon_i$, desarrollando la expresión de ν_0 se llega a

$$\begin{aligned}\nu_0 &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i - \varepsilon_i)^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}\left((\hat{f}_i^{[-i]} - f_i)^2 + \varepsilon_i^2 - 2(\hat{f}_i^{[-i]} - f_i)\varepsilon_i\right)\end{aligned} \tag{4.14}$$

PROPOSICIÓN 4.3. $\mathrm{E}[\nu_0] \to \mathrm{E}[M] + \sigma^2$ *si* $n \to \infty$.

Demostración. Por la linealidad de la esperanza:

$$\begin{aligned}\mathrm{E}[\nu_0] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[(\hat{f}_i^{[-i]} - f_i)^2 + \varepsilon_i^2 - 2(\hat{f}_i^{[-i]} - f_i)\varepsilon_i]\right] = \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{E}\left[(\hat{f}_i^{[-i]} - f_i)^2\right] + \mathrm{E}\left[\varepsilon_i^2\right] - 2\,\mathrm{E}\left[(\hat{f}_i^{[-i]} - f_i)\varepsilon_i\right]\right] = \\ &= \frac{\sum_{i=1}^{n}\mathrm{E}\left[(\hat{f}_i^{[-i]} - f_i)^2\right]}{n} + \sigma^2\end{aligned} \tag{4.15}$$

Si $n \to \infty$, entonces $\hat{f}^{[-i]} \approx \hat{f}$ porque la influencia prescindir de cada y_i en el ajuste es ínfima, y $\mathrm{E}[\nu_0] \approx \mathrm{E}[M] + \sigma^2$. □

Este resultado nos permite calcular el valor de λ que minimiza M.

Puede obtenerse mediante métodos computacionales una matriz cuadrada A de orden n, que aseguraría

$$\nu_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{f}_i)^2}{(1 - a_{i,i})^2} \tag{4.16}$$

donde $a_{i,i}$ es el i-ésimo elemento diagonal de A. Debido a los altos costes computacionales del método excluyendo una observación en cada ajuste, puede formularse el *score* de validación cruzada generalizada como sigue:

DEFINICIÓN 4.3 (*Score* de validación cruzada generalizada). Llamamos *score* de validación cruzada, y lo denotamos ν_g, al valor real positivo

$$\nu_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{(n - \mathrm{tr}(A))} \tag{4.17}$$

siendo A la misma matriz de la ecuación (4.16).

Dos variables

Sean X y V dos variables explicativas de la variable aleatoria Y observadas n veces y expresadas mediante la forma funcional

$$y_i = f_1(x_i) + f_2(v_i) + \varepsilon_i \tag{4.18}$$

donde α es un término independiente, f_1 y f_2 son funciones de suavizado y ε_i son las componentes de un vector de errores aleatorios independientes de dimensión n tales que $\mathrm{E}[\varepsilon] = 0 \in \mathbb{R}^n$ y tiene por matriz de covarianzas $\mathrm{Cov}[\varepsilon] = \Sigma_\varepsilon = \sigma^2 I_n$ que es una matriz cuadrada de orden n.

OBSERVACIÓN. Tengamos en cuenta que la aditividad $f_1(x) + f_2(v)$ es un caso particular de plantear $y = f(x, v)$.

Representación de la regresión por tramos penalizados. Sean b_j, $1 \leq j \leq k_1$, y B_j, $1 \leq j \leq k_2$, dos bases de funciones para f_1 y f_2 respectivamente. Podremos escribir $f_1(x) = \sum_{j=1}^{k_1} b_j(x)\delta_j$ y $f_2(v) = \sum_{j=1}^{k_2} B_j(v)\gamma_j$ donde δ_j, $1 \leq j \leq k_1$ y γ_j, $1 \leq j \leq k_2$ son parámetros reales.

Desarrollamos vectorial y matricialmente estas ideas. Sean $\vec{f_1} = (f_1(x_1), \ldots, f_1(x_n))^{\mathrm{t}}$, $X_1 = (b_j(x_i))_{i,j}$ una matriz de dimensión $n \times k_1$ y el vector $\delta = (\delta_1, \ldots, \delta_{k_1})^{\mathrm{t}} \in \mathbb{R}^{k_1}$. Se verifica

$$\vec{f_1} = X_1\delta = \begin{bmatrix} b_1(x_1) & \cdots & b_{k_1}(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \cdots & b_{k_1}(x_n) \end{bmatrix} \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_{k_1} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{k_1} b_j(x_1)\delta_j \\ \vdots \\ \sum_{j=1}^{k_1} b_j(x_n)\delta_j \end{bmatrix} = \begin{bmatrix} f_1(x_1) \\ \vdots \\ f_1(x_n) \end{bmatrix} \tag{4.19}$$

Análogamente, para $f_2(v)$ escribiremos

$$\vec{f_2} = X_2\gamma = \begin{bmatrix} B_1(v_1) & \cdots & B_{k_2}(v_1) \\ \vdots & \ddots & \vdots \\ B_1(v_n) & \cdots & B_{k_2}(v_n) \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{k_2} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{k_2} B_j(v_1)\gamma_j \\ \vdots \\ \sum_{j=1}^{k_2} B_j(v_n)\gamma_j \end{bmatrix} = \begin{bmatrix} f_2(v_1) \\ \vdots \\ f_2(v_n) \end{bmatrix} \tag{4.20}$$

siendo $X_2 = (B_j(v_i))_{i,j}$ una matriz de dimensión $n \times k_2$ y el vector $\gamma = (\gamma_1, \ldots, \gamma_{k_2})^{\mathrm{t}} \in \mathbb{R}^{k_2}$.

Sea $\delta^{\mathrm{t}} D_1{}^{\mathrm{t}} D_1 \delta = \delta^{\mathrm{t}} \overline{S}_1 \delta$ para $\vec{f_1}$ y $\gamma^{\mathrm{t}} D_2{}^{\mathrm{t}} D_2 \gamma = \gamma^{\mathrm{t}} \overline{S}_2 \gamma$ para $\vec{f_2}$, donde D_1 y D_2 se definen respectivamente para X y V como en el caso univariante. Vamos a exigir además al modelo

$$1_n{}^{\mathrm{t}} \vec{f_1} = \sum_{i=1}^{n} f_1(x_i) = 0 \in \mathbb{R} \tag{4.21}$$

i.e., $1_n{}^{\mathrm{t}} X_1 \delta = 0_{k_1}$ siendo δ los coeficientes para la base $b_j(\cdot)$ elegida, *i.e.*, $1_n{}^{\mathrm{t}} X_1 = 0_{k_1}$ (que es la matriz columna nula de longitud k_1).

$$\begin{bmatrix} 1, \cdots, 1 \end{bmatrix} \begin{bmatrix} b_1(x_1) & \cdots & b_{k_1}(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \cdots & b_{k_1}(x_n) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.22}$$

Veamos que si $\tilde{X}_1 = X_1 - \frac{1_n 1_n{}^{\mathrm{t}} X_1}{n}$, podemos definir $\tilde{\vec{f}}_1 = \tilde{X}_1\delta - 1_n c = \vec{f}_1 - c$, definicido c como veremos a continuación.

En primer lugar introducimos la notación $1_{m\times n}$ para la matriz de m filas y n columnas cuyos elementos son todos la unidad.

$$\begin{aligned}
\tilde{X}_1 &= X_1 - \frac{1_n 1_n{}^{\mathrm{t}} X_1}{n} = X_1 - \frac{1_{n\times n} X_1}{n} = \\
&= X_1 - n^{-1}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} b_1(x_1) & \cdots & b_{k_1}(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \cdots & b_{k_1}(x_n) \end{bmatrix} = \\
&= X_1 - n^{-1}\begin{bmatrix} \sum_{i=1}^n b_1(x_i) & \cdots & \sum_{i=1}^n b_{k_1}(x_i) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^n b_1(x_i) & \cdots & \sum_{i=1}^n b_{k_1}(x_i) \end{bmatrix} = \\
&= \begin{bmatrix} b_1(x_1) - n^{-1}\sum_{i=1}^n b_1(x_i) & \cdots & b_{k_1}(x_1) - n^{-1}\sum_{i=1}^n b_{k_1}(x_i) \\ \vdots & \ddots & \vdots \\ b_1(x_n) - n^{-1}\sum_{i=1}^n b_1(x_i) & \cdots & b_{k_1}(x_n) - n^{-1}\sum_{i=1}^n b_{k_1}(x_i) \end{bmatrix}
\end{aligned} \tag{4.23}$$

Cada elemento (i,j) de $\tilde{X}_1$ es el correspondiente $b_j(x_i)$ de X_1 menos la media de los valores de la columna, j, a la que pertenece; $(\tilde{X}_1)_{i,j} = (b_j(x_i) - n^{-1}\sum_{l=1}^n b_j(x_l))_{i,j}$.

Luego $\tilde{\vec{f}}_1 = \tilde{X}_1\delta = X_1\delta - 1_n 1_n{}^{\mathrm{t}} X_1\delta n^{-1} = X_1\delta - 1_n c$, y c resulta

$$c = \frac{1}{n} 1_n{}^{\mathrm{t}} X_1\delta \tag{4.24}$$

La matriz X_1 es de rango $\mathrm{rg}\,(X_1) = k_1$. Por la definición de $\tilde{X}_1$ a partir de combinaciones lineales, esta pierde una columna independiente y resulta que $\mathrm{rg}\left(\tilde{X}_1\right) = k - 1$.

Todo lo anterior puede ser expresado análogamente para $\vec{f}_2$, la base B_j, $1 \le j \le k_2$, los parámetros γ_j, $1 \le j \le k_2$, y las n observaciones de V.

Ajuste mediante mínimos cuadrados penalizados. De manera análoga al caso de una variable en la ecuación (4.9), el estimador $\hat{\beta}$ de β se alcanza minimizando la siguiente función

$$\hat{\beta} = \arg_\beta \min \left(||Y - X\beta||_2 + \lambda_1 \beta^{\mathrm{t}} S_1 \beta + \lambda_2 \beta^{\mathrm{t}} S_2 \beta \right) \tag{4.25}$$

donde los parámetros λ_j expresan el parámetro de suavizado de f_j, $j = 1, 2$.

PROPOSICIÓN 4.4. *Dado un modelo aditivo en dos variables, el mejor estimador $\hat{\beta}$ de β en el sentido de minimizar $||Y - X\beta||_2^2 + \lambda_1 \beta^{\mathrm{t}} S_1 \beta + \lambda_2 \beta^{\mathrm{t}} S_2 \beta$ es*

$$\hat{\beta} = (X^{\mathrm{t}} X + \lambda_1 S_1 + \lambda_2 S_2)^{-1} X^{\mathrm{t}} Y \tag{4.26}$$

Computacionalmente genera demasiadas dificultades y el cálculo se plantea en términos de minimizar

$$\left\| \begin{bmatrix} Y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ B \end{bmatrix} \right\|_2^2 \tag{4.27}$$

siendo B tal que $B^{\,\mathrm{t}}\, B = \lambda_1 S_1 + \lambda_2 S_2$, análogamente a como se hizo con una variable.

4.1.2. Definición

A continuación, ofrecemos una definición de modelo aditivo dada por Faraway (2016) [4]:

Definición 4.4 (Modelo aditivo). Sean Y un vector aleatorio de dimensión n cuyas componentes son de distribución libre, X una matriz de modelo de dimensión $n \times p$ ($p \leq n$), $\beta_0 \in \mathbb{R}$ y f_j funciones reales de una variable real desconocidas, $1 \leq j \leq p$. Diremos que Y verifica las condiciones de un modelo aditivo si se expresa en la forma

$$\mathrm{E}\,[Y_i] = \beta_0 + \sum_{j=1}^{p} f_j(x_{i,j}),\ 1 \leq i \leq n. \tag{4.28}$$

La ecuación (4.28) puede ser reformulada como $Y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{i,j}) + \varepsilon_i$, $1 \leq i \leq n$, siendo ε_i las coordenadas de un vector aleatorio de dimensión n y distribución libre sobre el que únicamente impondremos $\mathrm{E}\,[\varepsilon] = 0_n$ y $\mathrm{Cov}\,[\varepsilon] = \sigma^2 I_n$.

Ejemplo 4.2. El modelo de regresión lineal es un modelo aditivo de regresión no paramétrica en el que $f_j(x_{i,j}) = \beta_j x_{i,j}$, $1 \leq j \leq p$ y $1 \leq i \leq n$.

4.1.3. Estimación

Como señala Faraway (2016) [4], el modelo aditivo es mucho más flexible que el modelo lineal, y la aditividad de las funciones f_j hace interpretable la relación entre las variables X_j y la respuesta Y. La desventaja del modelo estará en la inconsistencia de sus estimaciones cuando existe una fuerte interacción entre variables del tipo $f_{j,k}(x_j x_k)$ ó $f_{j,k}(x_j, x_k)$.

Algoritmo *backfitting*

Un método de ajuste de un modelo aditivo viene dado en Hastie y Tibshirani (1990) basado en la teoría de suavizado que hemos presentado en el Capítulo 4.1.1. Es llamado algoritmo *backfitting*. Sea $\mathrm{E}\,[Y] = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$ un modelo aditivo y sea

$y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ el vector de observaciones de Y. Sea $\mathtt{s} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ una función cualquiera de suavizado, ya sea no paramétrica tipo *splines* o LOESS (toda la teoría más elemental sobre *splines* y LOESS puede consultarse en el Apéndice A.5), o paramétrica, tipo polinómico.

1. Sea $\hat{\beta}_0 = n^{-1} \sum_{i=1}^{n} y_i$ y sea $\hat{\beta}^{(0)}$ el estimador de β mediante mínimos cuadrados ordinarios, de modo que las f_j se inicializan[2] como $\hat{f}_j^{(0)}(x) = \hat{\beta}_j^{(0)} x$.

2. Fijada una condición de convergencia, el algoritmo itera en m, para $1 \leq j \leq p$, sobre los pasos:

   a) $e_{i,j} = y_i - (\beta_0 + \sum_{k<j} \hat{f}^{(m+1)}(x_{i,k}) + \sum_{k>j} \hat{f}^{(m)}(x_{i,k}))$,

   b) $e_j = (e_{1,j}, \ldots, e_{n,j})^{\mathrm{t}}$,

   c) $\hat{f}_j^{(m+1)} = \mathtt{s}(x_j, e_j)$, $m = 0, 1, 2, \ldots$, donde x_j denota la j-ésima columna de X.

   para $1 \leq j, k \leq p$.

Desde un punto de vista práctico, el ajuste de un modelo aditivo puede hacerse desde el paquete `gam`, basado en Hastie y Tibshirani (1990). El paquete `mgcv` proporcionado por Wood implementa este método considerando para el suavizado una base de *splines*. De hecho, este método maximiza la función de verosimilitud penalizada

$$\ell(\beta; y) - \sum_{j=1}^{p} \lambda_j \beta_j{}^{\mathrm{t}} S_j \beta_j \tag{4.29}$$

siendo los λ_j, obtenidos mediante validación cruzada generalizada (ver Definición 4.3) y las matrices S_j son las correspondientes a cada variable X_j de manera análoga a como definimos en el caso univariante de la ecuación (4.8).

4.1.4. Ejemplo

Para ilustrar el modelo aditivo, tomaremos el conjunto de datos `ozone` recogido en el paquete `faraway` de `R`.

[2] Es una propuesta de Faraway (2016) [4], Wood (2007) [16] propone inicializar las $\hat{f}_j$ como funciones nulas.

```
data(ozone, package="faraway")
```

Se trata de datos sobre calidad del aire en el área metropolitana de Los Ángeles (California) en 1976 y fueron recogidos por Breiman y Friedman (1985), pero la versión de Faraway omite las filas con valores perdidos, de manera que las observaciones sean completas. Se toma como variable respuesta la concentración de ozono (`O3`) en partes por millón (ppm). Las variables explicativas serán:

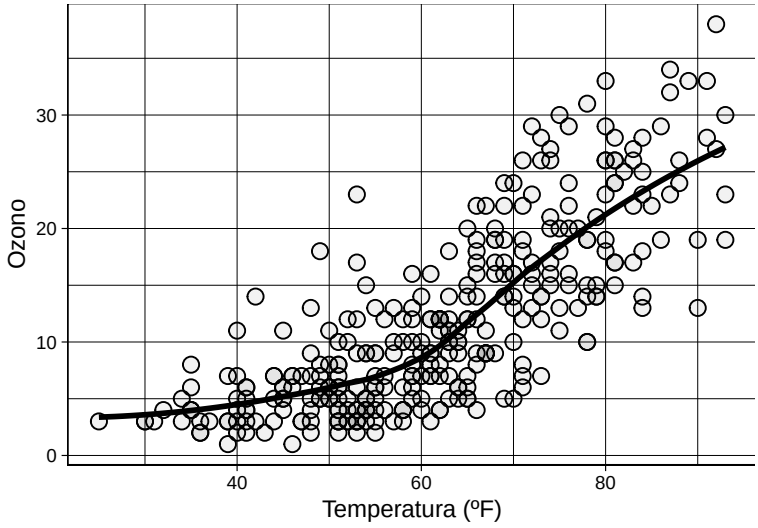
- Temperatura, `temp`, en ºF medida en El Monte.
- *Inversion base height*, `ibh`, en pies medida en el Aeropuerto de Los Ángeles (LAX).
- *Inversion base temperature*[3], `ibt`, en ºF medida en LAX.

El modelo aditivo resultante deberá tener la forma

$$\mathrm{E}\left[Y\right] = \beta_0 + f_1(\texttt{temp}) + f_2(\texttt{ibh}) + f_3(\texttt{ibt}) \tag{4.30}$$

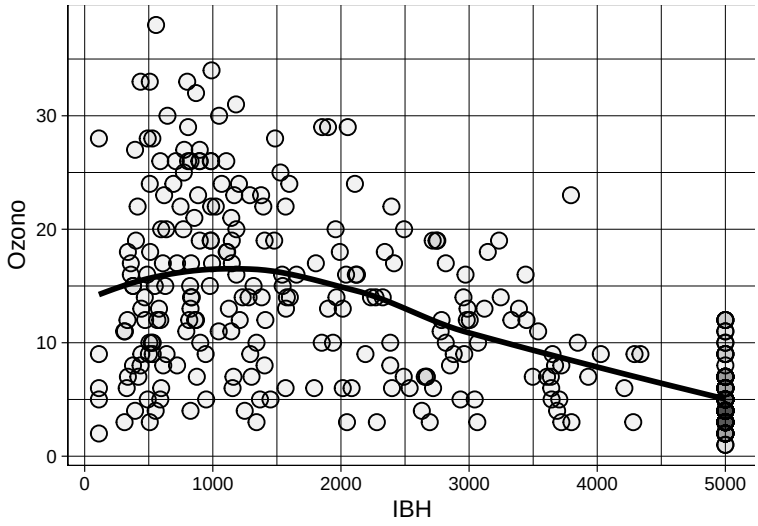
Visualización de las variables. Se muestra la relación que tiene, por separado, cada variable X_j con la respuesta `O3`:

- Y frente a X_1 (temperatura):

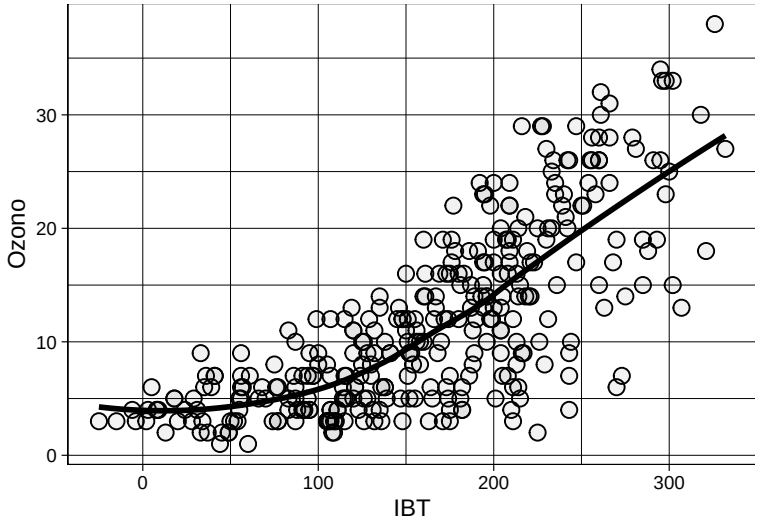


- Y frente a X_3 (IBH):

[3]En la edición impresa aparece «*inversion top temperature*». El 6 de mayo de 2022 notifiqué al profesor J. Faraway esta errata.



- Y frente a X_3 (IBT):



La primera modelización planteada es siguiendo un modelo lineal (que, como ya hemos señalado en el Ejemplo 4.2) sería un caso particular de modelo aditivo.

```
mlineal = lm(O3~temp+ibh+ibt,data=ozone)
summary(mlineal)
##
## Call:
## lm(formula = O3 ~ temp + ibh + ibt, data = ozone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3224  -3.1913  -0.2591   2.9635  13.2860
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -7.7279822  1.6216623  -4.765 2.84e-06 ***
## temp         0.3804408  0.0401582   9.474  < 2e-16 ***
## ibh         -0.0011862  0.0002567  -4.621 5.52e-06 ***
## ibt         -0.0058215  0.0101793  -0.572    0.568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.748 on 326 degrees of freedom
## Multiple R-squared:  0.652,Adjusted R-squared:  0.6488
## F-statistic: 203.6 on 3 and 326 DF,  p-value: < 2.2e-16
```

Vemos que sí se da una relación lineal entre `temp` e `ibh` y `O3`, sin embargo no para `ibt`. Queremos no descartar la variable como habríamos hecho en el modelo de regresión lineal normal, sino aplicar lo visto para modelos aditivos, veamos cómo. Como hemos dicho, el paquete `mgcv` de `R` es una revisión del algoritmo *backfitting* tomando como *smooth* los *splines* y que asume un modelo normal el ajuste (`Family: gaussian` en el resultado de ejecución).

```
library(mgcv)
maditivo = gam(O3~s(temp)+s(ibh)+s(ibt), data=ozone)
summary(maditivo)
##
## Family: gaussian
## Link function: identity
##
## Formula:
## O3 ~ s(temp) + s(ibh) + s(ibt)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.7758     0.2382   49.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F  p-value
## s(temp) 3.386  4.259 20.681  < 2e-16 ***
## s(ibh)  4.174  5.076  7.338 1.74e-06 ***
```

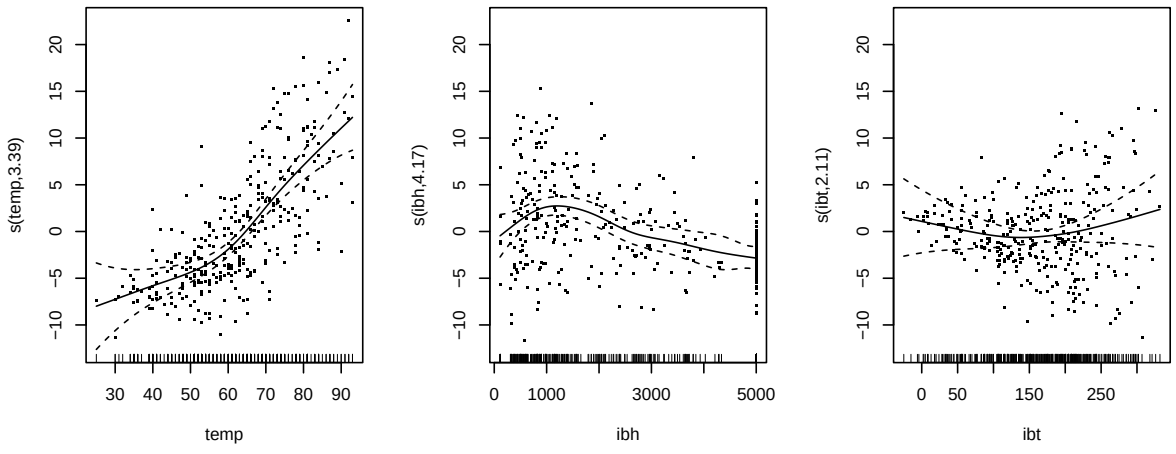
```
## s(ibt)  2.112  2.731  1.400    0.214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.708   Deviance explained = 71.7%
## GCV = 19.346  Scale est. = 18.72     n = 330
```

El único término paramétrico del modelo es β_0. La orden en `R` devuelve el estimador $\hat{\beta}_0 = 11.776$ y una significación $p < 2 \times 10^{16} < 0.05 = \alpha$.

Como hemos visto en el Capítulo 4.1.3, la estimación se efectúa sobre mínimos cuadrados, podemos calcular el test F de manera equivalente al modelo lineal cuando la función de suavizado `s` es lineal por tramos. Sin embargo, el resultado de salida de la instrucción es una modificación de este[4] obteniendo p-valores aproximados para las funciones de suavizado de las distintas variables X_j, $1 \leq j \leq 3$. La única variable que rechazamos del modelo es `ibt`.

El modelo excluye el valor R^2 de los modelos lineales y lo sustituye por el de devianza explicada. En este ejemplo, R^2 como devianza explicada es $R^2 = 0.717$, mayor que el $R^2 = 0.6488$ que obtuvimos en el modelo lineal `mlineal`.

Para observar qué tipo de transformaciones de suavizado se han realizado en cada variable, representamos gráficamente:



De estos gráficos obtenemos las siguientes ideas:

Temperatura: La temperatura parece ser una función lineal a tramos, con un cambio de tendencia próximo a los 60 ^{o}F, de un crecimiento más lento a una pendiente más

[4]Puede verse en la referencia de `R` del paquete.

acusada.

IBH: Igualmente parece haber un cambio de tendencia, esta vez incluso de los signos de estas. Primero, una fase creciente y luego decreciente.

IBT: Las bandas de confianza del ajuste (líneas discontinuas) y la línea de suavizado ajusta lo bastante próximo al 0 como para considerar la nulidad de esta variable.

Estudio del cambio de cambio de tendencia. Veamos si en efecto los cambios de tendencia arriba señalados son significativos, una vez rechazada la variable `ibt` del modelo.

Temperatura. Sean los modelos

$$\begin{cases} \mathrm{M}_A: & \mathrm{E}\left[Y_i\right] = \beta_0 + f_1(x_{i,1}) + f_2(x_{i,2}) \\ \mathrm{M}_B: & \mathrm{E}\left[Y_i\right] = \beta_0 + \beta_1 x_{i,1} + f_2(x_{i,2}) \end{cases} \tag{4.31}$$

```
modA <- gam(O3 ~ s(temp)+s(ibh),data=ozone)
modB <- gam(O3 ~ temp+s(ibh),data=ozone)
anova(modA,modB,test="F")
## Analysis of Deviance Table
##
## Model 1: O3 ~ s(temp) + s(ibh)
## Model 2: O3 ~ temp + s(ibh)
##   Resid. Df Resid. Dev      Df Deviance      F    Pr(>F)
## 1    319.11       6054
## 2    322.74       6950 -3.6237  -895.98 13.109 3.146e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se tiene un p valor (aproximado) muy inferior a la significación usual, $p = 3.14 \times 10^{-9} < 0.05 = \alpha$, luego asumimos el cambio de tendencia como en el modelo M_A.

IBH. Análogamente, evaluamos el cambio de tendencia en `ibh`: Sean los modelos

$$\begin{cases} \mathrm{M}_C: & \mathrm{E}\left[Y_i\right] = \beta_0 + f_1(x_{i,1}) + f_2(x_{i,2}) \\ \mathrm{M}_D: & \mathrm{E}\left[Y_i\right] = \beta_0 + f_1(x_{i,1}) + \beta_2 x_{i,2} \end{cases} \tag{4.32}$$

```
modC <- gam(O3 ~ s(temp)+s(ibh),data=ozone)
modD <- gam(O3 ~ s(temp)+ ibh,data=ozone)
anova(modC,modD,test="F")
## Analysis of Deviance Table
##
## Model 1: O3 ~ s(temp) + s(ibh)
## Model 2: O3 ~ s(temp) + ibh
##   Resid. Df Resid. Dev      Df Deviance     F  Pr(>F)
## 1    319.11     6054.0
## 2    323.18     6316.2 -4.0607  -262.25 3.424 0.00892 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Y siguiendo el mismo criterio el p-valor (aproximado) resultante es lo bastante inferior a $\alpha = 0.05$ y damos por bueno el modelo de cambio de tendencia en `ibh`.

Faraway (2016) [4] continua desarrollando este ejemplo con teoría sobre funciones de suavizado más elaboradas, basada en producto tensorial entre variables, `te(temp,ibh)`, cuya exposición queda fuera de los límites de este trabajo.

4.2. Modelo aditivo generalizado

4.2.1. Definición

La siguiente definición no es la original en Hastie y Tibshirani (1986) [6] sino que vine dada por Wood (2007) [16].

Definición 4.5 (Modelo aditivo generalizado). Sean Y un vector aleatorio de dimensión n, x_j, $1 \leq j \leq p$ columnas de una matriz X de orden $n \times p$ de constantes conocidas, f_j, $1 \leq j \leq p \leq n$, una colección de funciones $f_j : \mathbb{R} \to \mathbb{R}$ desconocidas $\gamma \in \mathbb{R}^p$ un parámetro vectorial desconocido, y A una matriz de modelo de dimensión $n \times p$. Además, sea $g : \mathbb{R} \to \mathbb{R}$ una función monótona y derivable. Diremos que Y se ajusta a un modelo aditivo generalizado si verifica:

1. Las coordenadas de Y son variables aleatorias independientes con funciones de densidad $p(y_i; \theta_i, \phi_i)$, $1 \leq i \leq n$, pertenecientes todas ellas a la misma familia exponencial, y con media $\mathrm{E}[Y_i] = \mu_i = b'(\theta_i)$.

2. $g(\mu_i) = \eta_i = a_i{}^{\mathrm{t}}\,\gamma + \sum_{j=1}^{p} f_j(x_{i,j})$, siendo $a_i{}^{\mathrm{t}}$ la i-ésima fila de la matriz A, $1 \leq i \leq n$.

Wood (2007) [16] propone definiciones alternativas a la segunda condición, una de ellas es $g(\mu_i) = \eta_i = a_i{}^{\mathrm{t}}\,\gamma + \sum_{j=1}^{p} L_{i,j} f_j(x_j)$ siendo $L_{i,j}$ una función lineal que acota a f_j de modo que coincide con la definición original si $L_{i,j} f_j(x_j) = f_j(x_{i,j})$. Existen definiciones alternativas para $L_{i,j}$ orientadas al modelo aditivo generalizado mixto, que no trataremos aquí.

4.2.2. Estimación

El proceso de estimación para el modelo aditivo generalizado sigue la idea del algoritmo *backfitting* descrito en el Capítulo 4.1.3. Wood (2007) [16] apunta que, si bien en el modelo aditivo la estimación se basaba en mínimos cuadrados penalizados, el modelo aditivo generalizado buscará maximizar una función de verosimilitud penalizada,

$$\ell(\beta; y) - \sum_{j=1}^{p} \lambda_j \beta_j{}^{\mathrm{t}}\, S_j \beta_j, \tag{4.33}$$

que en la práctica se traduce en un proceso iterativo de mínimos cuadrados penalizados[5].

Algoritmo PIRLS. Veamos el proceso iterativo descrito por Wood (2007) [16][6]:

1. Sean $\hat{\mu}_i^{(0)} = y_i + \delta_i$ y $\hat{\eta}_i^{(0)} = g(\hat{\mu}_i^{(0)})$, $1 \leq i \leq n$, y donde δ_i suele ser cero o una cantidad que garantice que $\hat{\eta}_i^{(0)}$ esté bien definido.

2. Iteramos en m hasta alcanzar convergencia:

   a) Iteramos en i recorriendo $1 \leq i \leq n$:

      1) Sean $w_i^{(m)} = (V(\hat{\mu}_i^{(m)}) g'(\mu_i^{(m)})^2)^{-1}$ y $z_i^{(m)} = g'(\hat{\mu}_i^{(m)})(y_i - \hat{\mu}_i^{(m)}) + \hat{\eta}_i^{(m)}$.

      2) Sea la matriz diagonal $W^{(m)}$ con $W_{i,i}^{(m)} = w_i^{(m)}$ y $W^{\frac{1}{2}\,(m)}$ tal que $W^{\frac{1}{2}\,(m)} W^{\frac{1}{2}\,(m)} = W^{(m)}$.

   b) Hallar

   $$\hat{\beta}^{(m+1)} = \arg_\beta \min \left(\left|\left| W^{\frac{1}{2}\,(m)} z^{(m)} - W^{\frac{1}{2}\,(m)} X\beta \right|\right|_2^2 \right) + \sum_j \lambda_j \beta^{\mathrm{t}}\, S_j \beta \tag{4.34}$$

   donde todos los argumentos dados lo son para la iteración m, $m = 1, 2, \ldots$.

   c) Actualizamos $\eta^{(m+1)} = X\hat{\beta}^{(m+1)}$ y $\hat{\mu}_i = g^{-1}(\eta^{(m+1)})$.

[5] *Penalized iterative least squares*, PIRLS

[6] Se señala que es un algoritmo análogo al de los modelos mixtos lineales generalizados presentados en el mismo libro, pero que no ha sido tratado en este trabajo.

4.2.3. Ejemplo

El siguiente ejemplo viene dado por Wood (2007) [16] e ilustra la estiamción mediante la biblioteca `mcgv` de `R` para el conjunto de datos `trees` nativo en `R`.

Se trata de 31 observaciones en árboles de la especie cerezo criollo (*Prunus serotina*) y recogidas por primera vez por A.C. Atkinson en su obra *Plots, Transformations and Regression*, 1985 [1].

```
data(trees)
```

El conjunto consta de tres variables:

1. Diámetro (`girth`) del tronco del árbol, medido en pulgadas.

2. Altura (`height`) total del árbol, medido en pies, X_1.

3. Volumen (`volume`) de madera del árbol, medido en pies cúbicos, X_2.

Antes de continuar, mostramos un resumen numérico del conjunto de datos

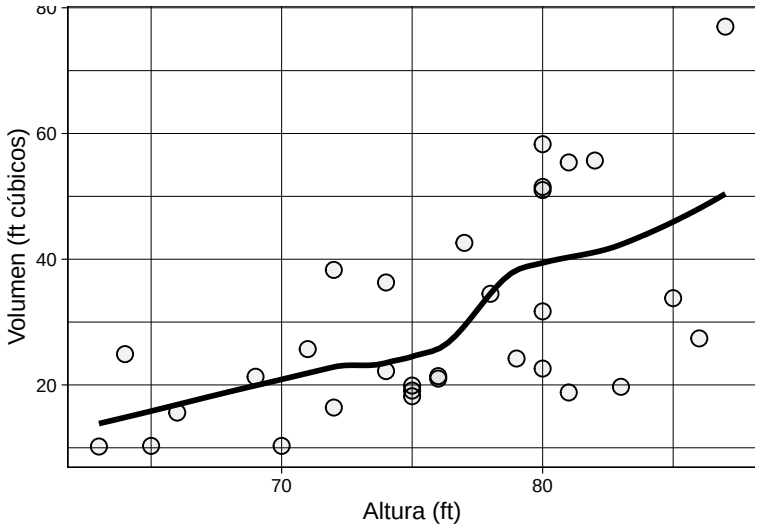
```
summary(trees)
##      Girth           Height       Volume
##  Min.   : 8.30   Min.   :63   Min.   :10.20
##  1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
##  Median :12.90   Median :76   Median :24.20
##  Mean   :13.25   Mean   :76   Mean   :30.17
##  3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
##  Max.   :20.60   Max.   :87   Max.   :77.00
```

El problema a plantear es que una función de enlace g que relacione el diámetro con alguna expresión funcional (f_1 y f_2) de la altura y el volumen. Como se señala en el planteamiento del problema, resulta razonable que la varianza y media del volumen aumenten conjuntamente, por lo que suponemos $Y_i \sim \text{Gamma}\,(\alpha_i, \beta_i)$. En tal caso $\mu_i = \frac{\alpha_i}{\beta_i}$, luego una formulación del modelo aditivo generalizado sería

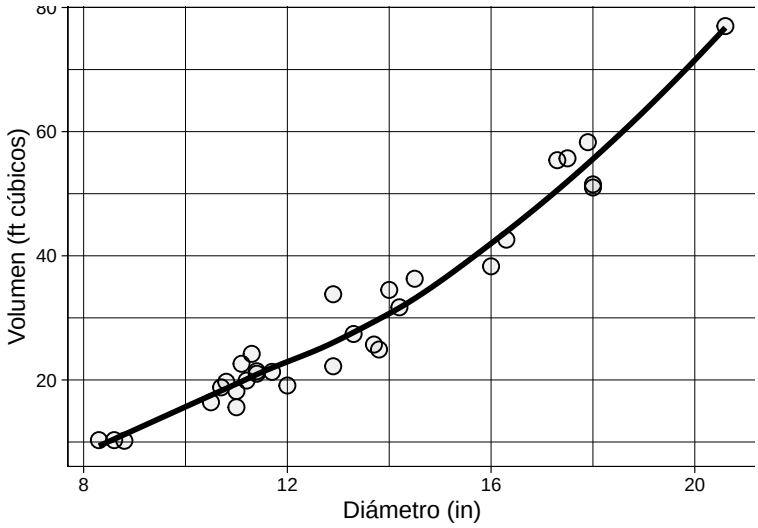
$$\log(\mu_i) = f_1(X_1) + f_2(X_2) \tag{4.35}$$

Visualización de las variables. Veamos la relación entre Y y las distintas variables regresoras del modelo:

- Y frente a X_1 (altura):



- Y frente a X_2 (diámetro):



Modelo.

OBSERVACIÓN. Es importante señalar que, en lo que sigue no buscaremos un modelo que se ajuste por completo a la Definición 4.2.1, sino una forma más simple y parecida al ejemplo del modelo aditivo calculado antes, esto es

$$g(\mathrm{E}\,[Y_i]) = \beta_0 + \sum_{j=1}^{p} f_j(x_{i,j}). \tag{4.36}$$

Wood (2007) [16] desarrolla este ejemplo ofreciendo estimaciones de los términos paramétricos ${a_i}^{\mathrm{t}}\,\gamma$, pero, de nuevo, recurre a bases de funciones y técnicas que escapan a lo expuesto en este trabajo.

Según la definición de función de varianza (ver Definición 3.6),

$$V(\mu_i) = \mu_i^2,\ 1 \leq i \leq n, \tag{4.37}$$

y

$$g'(\mu_i) = \log'(\mu_i) = \frac{1}{\mu_i},\ 1 \leq i \leq n, \tag{4.38}$$

luego los coeficientes diagonales de W serían

$$w_i = \frac{1}{V(\mu_i)g'(\mu_i)^2} = \frac{1}{\mu_i^2\mu_i^{-2}} = 1,\ 1 \leq i \leq n, \tag{4.39}$$

i.e., $W = I_n = W^{\frac{1}{2}}$ en este caso; y el vector $z \in \mathbb{R}^n$ se construye como

$$z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i = \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} + \hat{\eta}_i,\ 1 \leq i \leq n. \tag{4.40}$$

Computacionalmente, este será la iteración $m = 0$ sobre la que trabajará la función `gam()` del paquete `mgcv`.

```
library(mgcv)
mag <- gam(Volume~s(Height)+s(Girth), family=Gamma(link=log), data=trees)
```

Por defecto, la función toma los λ_j, $j = 1, 2$ mediante validación cruzada generalizada y hemos indicado que la base se funciones de suavizado sean *splines*, `s`.

```
summary(mag)
##
## Family: Gamma
## Link function: log
##
## Formula:
## Volume ~ s(Height) + s(Girth)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.27570    0.01492   219.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df      F  p-value
## s(Height) 1.000  1.000  31.32 7.07e-06 ***
```

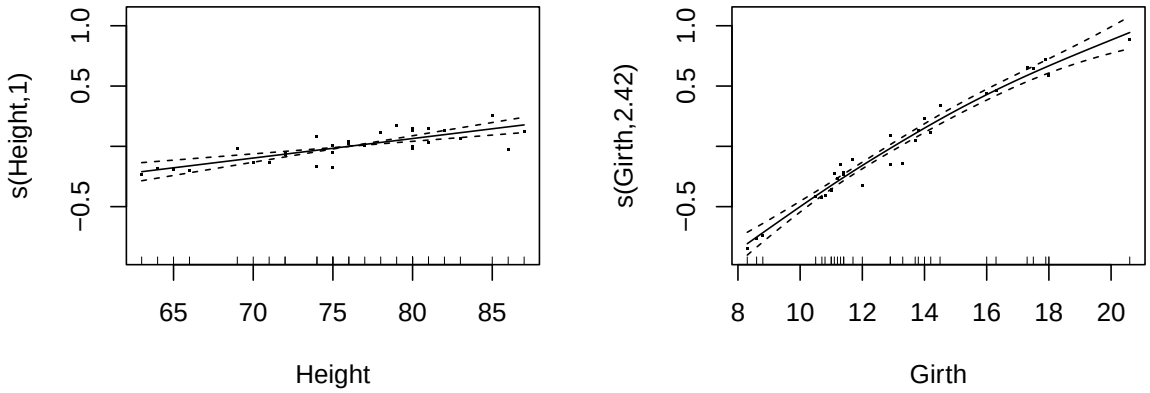
```
## s(Girth)  2.422  3.044 219.28  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.973   Deviance explained = 97.8%
## GCV = 0.0080824  Scale est. = 0.006899  n = 31
```

El único término paramétrico es $\beta_0 = 3.27570$ que resulta signficativo con un p-valor $p < 2 \times 10^{-16} < 0.05 = \alpha$.

En cuanto a las variables regresoras, el ajuste del modelo mediante las funciones f_1 y f_2 con bases de *spline*, con también significativas con sendos p-valores $7.07 \times 10{-}6$ y 2×10^{-16}.

La devianza explicada del modelo es $R^2 = 0.978$, muy próximo a 1, que indica un ajuste casi perfecto, *i.e.*, el modelo explica el 97.8 % de la varianza total.

La representación gráfica de los residuos es



Vemos en la orden `summary(mag)` y en el gráfico de los residuos que los grados de libertad de la altura, `Height`, es igual a 1, es decir, f_1 es casi una función lineal.

Apéndice A

Otros resultados

A continuación se presentan algunas definiciones y resultados que son de utilidad a lo largo de este trabajo. Cabe tener en cuenta en este apéndice son presentados de la manera más general, puede cambiar la notación en las definiciones, proposiciones, teoremas y sus demostraciones según son formuladas en el campo de los modelos lineales, lineales generalizados o aditivos generalizados.

A.1. Resultados de Análisis Matemático

Definición A.1 (Gradiente). Sea la función $f : \mathbb{R}^n \to \mathbb{R}$. Se define el gradiente de f en $x \in \mathbb{R}^n$ como $\nabla f(x) = (\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n})^{\mathrm{t}}$.

Definición A.2 (Jacobiano). Sea la función $f : \mathbb{R}^n \to \mathbb{R}^m$. Se define el jacobiano de f en $x \in \mathbb{R}^n$ como

$$\mathrm{J}_{f(x)} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_j} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial f_i(x)}{\partial x_1} & \cdots & \frac{\partial f_i(x)}{\partial x_j} & \cdots & \frac{\partial f_i(x)}{\partial x_n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_j} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix}. \tag{A.1}$$

A.1.1. Método de Newton-Raphson

Escalar

Sea $f : \mathbb{R} \to \mathbb{R}$ continua en el intervalo $[a, b] \subseteq \mathbb{R}$, de la que queremos encontrar una raíz r. Supongamos que f' está bien definida cualquiera que sea $x \in [a, b]$. Si $f'(r) \neq 0$,

entonces existe $\delta > 0$ tal que la sucesión $(x^{(k)})$, $k = 0, 1, \ldots$, definida por el proceso iterativo

$$x^{(k)} = x^{(k-1)} - \frac{f(x^{(k-1)})}{f'(x^{(k-1)})},\ k = 1, 2, \ldots \tag{A.2}$$

converge a r sea cualquiera que sea el valor inicial $x^{(k)} \in [x - \delta, x + \delta]$.

Vectorial

Sea $f : \mathbb{R}^n \to \mathbb{R}^n$ continua en el n-intervalo $[a, b] \subseteq \mathbb{R}^n$, de la que queremos encontrar una raíz $r = (r_1, \ldots, r_n)^{\mathrm{t}}$. Sea f_i diferenciable con respecto a cada una de las x_j. La expresión vectorial del método de Newton-Raphson para funciones $\mathbb{R}^n$-valoradas resulta

$$x^{(k)} = x^{(k-1)} - \mathrm{J}^{-1}_{f(x^{(k-1)})} f(x^{(k-1)}),\ k = 1, 2, \ldots \tag{A.3}$$

Para un estudio más exhaustivo de los métodos numéricos, se remite al lector a Mathews y Fink (2000) [8].

A.1.2. Diferenciación bajo el signo integral

TEOREMA A.1 (Convergencia dominada). *Sea (f_n) una sucesión de funciones reales, integrables respecto de una cierta medida μ, que converge puntualmente a una función medible f. Si existe una función g integrable tal que $|f_n| \leq g$ para todo n, entonces $\int f \,\mathrm{d}\,\mu$ existe y coincide con el límite de las integrales de (f_n).*

Se trata de un resultado de la teoría de la medida muy general pero que sirve de punto de partida para enunciar el siguiente teorema, de gran utilidad en la teoría de los modelos lineales generalizados.

TEOREMA A.2. *Sea $f : \mathbb{R} \times \Theta \to \mathbb{R}$ donde Θ es un abierto de $\mathbb{R}$, verificando:*

1. *$f(\cdot, \theta) : \Theta \to \mathbb{R}$ es una función Lebesgue-integrable para cualquier $\theta \in \Theta$.*
2. *$\frac{\partial f(x,\theta)}{\partial \theta}$ existe y es continua para cualquier x fijado salvo en un conjunto de medida nula de $\mathbb{R}$.*
3. *Existe $g : \mathbb{R} \to \mathbb{R}$ Lebesgue-integrable que acota a $\left|\frac{\partial f(x,\theta)}{\partial \theta}\right|$ en todo $\theta \in \Theta$ y $x \in \mathbb{R}$ salvo un conjunto de medida nula.*

Entonces,

$$\frac{\partial \int_{\mathbb{R}} f(x, \theta) \,\mathrm{d}\, x}{\partial \theta} = \int_{\mathbb{R}} \frac{\partial f(x, \theta)}{\partial \theta} \,\mathrm{d}\, x. \tag{A.4}$$

A.2. Conceptos de Álgebra Lineal

Definición A.3 (Matriz simétrica). Una matriz cuadrada A de dimensión $n \times n$ es simétrica si $a_{i,j} = a_{j,i}$, para cualesquiera $1 \leq i, j \leq n$.

Definición A.4 (Matriz semidefinida positiva). Una matriz cuadrada A de dimensión $n \times n$ es semidefinida positiva si cualquier vector $v \in \mathbb{R}^n$ no nulo verifica $v^{\,\mathrm{t}}\, Av \geq 0$.

Definición A.5 (Matriz definida positiva). Una matriz cuadrada A de dimensión $n \times n$ es definida positiva si cualquier vector $v \in \mathbb{R}^n$ no nulo verifica $v^{\,\mathrm{t}}\, Av > 0$.

A.3. Distribuciones de probabilidad

A.3.1. Distribuciones absolutamente continuas

Distribución normal

Definición A.6 (Distribución normal). Una variable aleatoria X sigue una distribución normal de parámetros $\mu \in \mathbb{R}$ y $\sigma^2 \in \mathbb{R}^+$, y la denotamos $X \sim \mathrm{N}(\mu, \sigma^2)$, si tiene por función de densidad de probabilidad

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}},\ x \in \mathbb{R}. \tag{A.5}$$

Proposición A.1. *Si* $X \sim \mathrm{N}(\mu, \sigma^2)$, *entonces* $\mathrm{E}[X] = \mu$ *y* $\mathrm{Var}[X] = \sigma^2$.

Definición A.7 (Distribución normal multivariante). Un vector aleatorio X de dimensión n sigue una distribución normal multivariante de parámetros $\mu \in \mathbb{R}^n$ y Σ (siendo esta una matriz cuadrada de dimensión $n \times n$ simétrica y semidefinida positiva), y lo denotamos $X \sim \mathrm{N}_n(\mu, \Sigma)$, si tiene por función de densidad de probabilidad

$$p(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^{\mathrm{t}}\, \Sigma^{-1}(x-\mu)},\ x \in \mathbb{R}^n. \tag{A.6}$$

Proposición A.2. *Si* $X \sim \mathrm{N}_n(\mu, \Sigma)$, *entonces* $\mathrm{E}[X] = \mu$ *y* $\mathrm{Cov}[X] = \Sigma$.

Proposición A.3. *Si* $X \sim \mathrm{N}_n(\mu, \Sigma)$, *entonces cada coordenada* $X_i \sim \mathrm{N}(\mu_i, \Sigma_{i,i} = \sigma_i^2)$, $1 \leq i \leq n$.

Distribución ji-cuadrado

DEFINICIÓN A.8 (Distribución ji-cuadrado). Una variable aleatoria absolutamente continua X sigue una distribución ji-cuadrado de n grados de libertad, y la denotamos $X \sim \chi^2(n)$, si tiene por función de densidad de probabilidad[1]

$$p(x;n) = \frac{x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} I_{]0,\infty[}(x). \tag{A.7}$$

PROPOSICIÓN A.4. *Si $Z \sim \mathrm{N}_n(0, I_n)$, entonces $X = \sum_i Z_i^2$ sigue una distribución ji-cuadrado de n grados de libertad, $X \sim \chi^2(n)$.*

DEFINICIÓN A.9 (Distribución ji-cuadrado no central). Una variable aleatoria absolutamente continua X sigue una distribución ji-cuadrado no central de n grados de libertad y parámetro de localización ν, y la denotamos $X \sim \chi^2(n,\nu)$, si tiene por función de densidad de probabilidad

$$p(x;n,\nu) = \sum_{k=0}^{\infty} \mathrm{e}^{-\nu}\frac{\nu^k}{k!}\frac{x^{\frac{n+2k}{2}-1}\mathrm{e}^{-\frac{x}{2}}}{\Gamma(\frac{n+2k}{2})2^{\frac{n+2k}{2}}} I_{]0,\infty[}(x). \tag{A.8}$$

PROPOSICIÓN A.5. *Si $Z \sim \mathrm{N}_n(\mu, I_n)$, entonces $X = \sum_i Z_i^2$ sigue una distribución ji-cuadrado no central de n grados de libertad y parámetro de localización $\nu = \frac{||\mu||_2^2}{2}$, $X \sim \chi^2(n,\nu)$.*

Se trata de una generalización de la distribución ji-cuadrado de n grados de libertad y coincide con esta cuando el parámetro de centralidad se anula, *i.e.*, $\chi^2(n) \equiv \chi^2(n,0)$.

Distribución F

DEFINICIÓN A.10 (Distribución F). Una variable aleatoria absolutamente continua S sigue una distribución F-Snedecor de parámetros enteros positivos m y n, y la denotamos $S \sim F(m,n)$, si tiene por función de densidad de probabilidad

$$p(s;m,n) = \frac{m^{\frac{m}{2}}n^{\frac{n}{2}}\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\frac{s^{\frac{m}{2}-1}}{(ms+n)^{\frac{m+n}{2}}} I_{]0,\infty[}(s). \tag{A.9}$$

PROPOSICIÓN A.6. *Si $X \sim \chi^2(m)$ e $Y \sim \chi^2(n)$ son independientes, entonces $S = \frac{X}{Y}\frac{n}{m}$ sigue una distribución F-Snedecor de parámetros m y n, $S \sim F(m,n)$.*

[1] Si $x > 0$, se define la función gamma como $\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}\,t$.

DEFINICIÓN A.11 (Distribución F no central). Una variable aleatoria absolutamente continua S sigue una distribución F-Snedecor no central de parámetros enteros positivos m y n, y parámetro de centralidad ν, y la denotamos $S \sim F(m, n, \nu)$, si tiene por función de densidad de probabilidad

$$p(s; m, n, \nu) = \sum_{k=0}^{\infty} \mathrm{e}^{-\nu} \frac{\nu^k}{k!} \frac{\left(\frac{m}{n}\right)^{\frac{m+2k}{2}} \Gamma\left(\frac{m+n+2k}{2}\right)}{\Gamma\left(\frac{m+2k}{2}\right)\Gamma\left(\frac{n}{2}\right)} s^{\frac{n_1+2k}{2}-1} \left(2 + \frac{m}{n}s\right)^{\frac{m+n+2k}{2}} I_{]0,\infty[}(s). \quad \text{(A.10)}$$

PROPOSICIÓN A.7. *Si $X \sim \chi^2(m, \nu)$ e $Y \sim \chi^2(n, 0)$ son independientes, entonces $S = \frac{X}{Y}\frac{n}{m} \sim F(m, n, \nu)$.*

Se trata de una generalización de la distribución F-Snedecor central y coincide con esta cuando el parámetro de centralidad se anula, *i.e.*, $F(m, n) \equiv F(m, n, 0)$.

Distribución t-Student

DEFINICIÓN A.12 (Distribución t-Student). Una variable aleatoria absolutamente continua T sigue una distribución t-Student con n grados de libertad, y la denotamos $T \sim t(n)$, si tiene por función de densidad de probabilidad

$$p(t; n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}}, \; t \in \mathbb{R}. \quad \text{(A.11)}$$

PROPOSICIÓN A.8. *Si $X \sim \mathrm{N}(0, 1)$ e $Y \sim \chi^2(n)$ son independientes, entonces $T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n)$.*

Distribución gamma

DEFINICIÓN A.13 (Distribución gamma). Una variable aleatoria absolutamente continua X sigue una distribución gamma de parámetros[2] $\alpha > 0$ y $\beta > 0$, y la denotamos $X \sim \mathrm{Gamma}(\alpha, \beta)$, si tiene por función de densidad de probabilidad

$$p(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-x\beta} I_{]0,\infty[}(x). \quad \text{(A.12)}$$

PROPOSICIÓN A.9. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes con distribución gamma y parámetros (α_i, β), $X_i \sim \mathrm{Gamma}(\alpha_i, \beta)$, $1 \leq i \leq n$. La variable aleatoria $S = \sum_{i=1}^n X_i$ sigue una distribución gamma, $S \sim \mathrm{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.*

[2]Existe una definición alternativa $\mathrm{Gamma}(\kappa, \theta)$, donde $\kappa = \alpha$ y $\theta = \beta^{-1}$.

DEFINICIÓN A.14 (Distribución Erlang). Una variable aleatoria X sigue una distribución Erlang de parámetros $\kappa \in \mathbb{Z}^+$ y $\lambda > 0$, y la denotamos $X \sim \text{Erlang}(\kappa, \lambda)$ si tiene por función de densidad de probabilidad

$$p(x; \kappa, \lambda) = \frac{\lambda^\kappa}{(\kappa - 1)!} x^{\kappa-1} \mathrm{e}^{-x\lambda} I_{]0,\infty[}(x). \tag{A.13}$$

PROPOSICIÓN A.10. *Se verifica que* $\text{Gamma}(\alpha, \beta) \equiv \text{Erlang}(\alpha, \beta)$ *si* $\alpha \in \mathbb{Z}^+$.

DEFINICIÓN A.15 (Distribución exponencial). Una variable aleatoria X sigue una distribución exponencial de parámetro $\lambda > 0$, y la denotamos $X \sim \text{Exp}(\lambda)$, si tiene por función de densidad de probabilidad

$$p(x; \lambda) = \lambda \mathrm{e}^{-\lambda x} I_{]0,\infty[}(x). \tag{A.14}$$

PROPOSICIÓN A.11. *Se verifica que* $\text{Gamma}(1, \beta) \equiv \text{Erlang}(1, \beta) \equiv \text{Exp}(\beta)$.

PROPOSICIÓN A.12. *Si* $X \sim \text{Gamma}(\alpha, \beta)$, *entonces* $\text{E}[X] = \frac{\alpha}{\beta}$ *y* $\text{Var}[X] = \frac{\alpha}{\beta^2}$*; y por las Proposiciones A.11 y A.12 son inmediatas la media y la varianza de las distribuciones Erlang y exponencial.*

A.3.2. Distribuciones discretas

Distribución de Bernoulli

DEFINICIÓN A.16 (Distribución de Bernoulli). Una variable aleatoria discreta X sigue una distribución de Bernoulli con parámetro $\pi \in]0, 1[$, y la denotamos $X \sim \text{Be}(\pi)$, si tiene por función de masa de probabilidad

$$p(x; \pi) = \pi^x (1 - \pi)^{1-x} I_{\{0,1\}}(x). \tag{A.15}$$

Distribución geométrica

DEFINICIÓN A.17 (Distribución geométrica). Una variable aleatoria discreta X sigue una distribución geométrica con parámetro $\pi \in]0, 1[$, y la denotamos $X \sim \text{Geom}(\pi)$, si tiene por función de masa de probabilidad

$$p(x; \pi) = \pi(1 - \pi)^x I_{\mathbb{Z}_0^+}(x). \tag{A.16}$$

PROPOSICIÓN A.13. *Si* $X \sim \text{Geom}(\pi)$, *entonces* $\text{E}[X] = \frac{1-\pi}{\pi}$ *y* $\text{Var}[X] = \frac{1-\pi}{\pi^2}$.

Distribución binomial

DEFINICIÓN A.18 (Distribución binomial). Una variable aleatoria discreta X sigue una distribución binomial, fijado n entero positivo, de parámetro $\pi \in]0,1[$, y la denotamos $X \sim \mathrm{b}_n(\pi)$, si tiene por función de masa de probabilidad

$$p(x;\pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x}I_{\{0,\ldots,n\}}(x). \tag{A.17}$$

PROPOSICIÓN A.14. *Si* $X \sim \mathrm{b}_n(\pi)$, *entonces* $\mathrm{E}[X] = n\pi$ *y* $\mathrm{Var}[X] = n\pi(1-\pi)$.

PROPOSICIÓN A.15. *Sean* $Z_1, \ldots, Z_n$ *variables aleatorias independientes e idénticamente distribuidas, tales que* $Z_i \sim \mathrm{Be}(\pi)$, $1 \leq i \leq n$. *La variable aleatoria* $X = \sum_{i=1}^n Z_i \sim \mathrm{b}_n(\pi)$.

Distribución de Poisson

DEFINICIÓN A.19 (Distribución de Poisson). Una variable aleatoria discreta X sigue una distribución de Poisson con parámetro de intensidad $\lambda \in \mathbb{R}^+$, y la denotamos $X \sim \mathrm{Poisson}(\lambda)$, si tiene por función de masa de probabilidad

$$p(x;\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}I_{\mathbb{Z}_0^+}(x). \tag{A.18}$$

PROPOSICIÓN A.16. *Si* $X \sim \mathrm{Poisson}(\lambda)$, *entonces* $\mathrm{E}[X] = \lambda$ *y* $\mathrm{Var}[X] = \lambda$.

PROPOSICIÓN A.17. *Sean* $X_1, \ldots, X_n$ *variables aleatorias independientes con distribución de Poisson,* $X_i \sim \mathrm{Poisson}(\lambda_i)$, $1 \leq i \leq n$. *La variable aleatoria* $S = \sum_{i=1}^n X_i$ *sigue una distribución de Poisson con parámetro igual a la suma de los parámetros de cada variable aleatoria de Poisson,* $\sum_{i=1}^n X_i = S \sim \mathrm{Poisson}(\sum_{i=1}^n \lambda_i)$.

Distribución multinomial

DEFINICIÓN A.20 (Distribución multinomial). Diremos que el vector aleatorio X de dimensión k sigue una distribución multinomial de parámetros $(n, \pi_1, \ldots, \pi_k)$ con $n \in \mathbb{Z}^+$, cada $\pi_i \in]0,1[$ tales que $\sum_{i=1}^k \pi_i = 1$ y $\sum_{i=1}^k x_i = n$, y lo denotamos[3] $X \sim \mathrm{M}(n, \pi_1, \ldots, \pi_k)$, si tiene por función de masa de probabilidad

$$p(x; n, \pi_1, \ldots, \pi_k) = n!\frac{\prod_{i=1}^k \pi_i^{x_i}}{\prod_{i=1}^k x_i!},\ x \in \mathbb{Z}^{+k}. \tag{A.19}$$

[3]En ocasiones se omite el último parámetro, ya que se deduce inmediatamente como $\pi_k = 1 - \pi_1 - \ldots - \pi_{k-1}$.

PROPOSICIÓN A.18. *El caso $k = 2$ se corresponde con una distribución binomial* $\mathrm{b}_n(\pi) \equiv \mathrm{M}(n, \pi, 1-\pi)$.

PROPOSICIÓN A.19. *Si* $X \sim \mathrm{M}(n, \pi_1, \ldots, \pi_k)$, *entonces* $\mathrm{E}[X_i] = n\pi_i$, $\mathrm{Var}[X_i] = n\pi_i(1-\pi_i)$ *y* $\mathrm{Cov}[X] = (-n\pi_i\pi_l)_{1\leq i,l\leq k}$.

A.4. Resultados de estadística

A.4.1. Estimación por máxima verosimilitud

Sea X una variable aleatoria con función de densidad de probabilidad $p(x;\theta)$ perteneciente a la familia paramétrica $\{p(\,\cdot\,;\theta) : \theta \in \Theta\}$.

DEFINICIÓN A.21 (Función de verosimilitud). Sea $\underline{x} = (x_1, \ldots, x_n)$ una muestra aleatoria simple, de tamaño n, de la variable aleatoria X. Llamamos función de verosimilitud del parámetro θ a partir de la muestra $\underline{x}$ a $L(\theta;\underline{x}) = \prod_{i=1}^{n} p(x_i;\theta) > 0$.

El método de estimación por máxima verosimilitud consistirá en hallar aquellos valores $\hat{\theta} \in \Theta$ que maximizan la función de verosimilitud de θ dada la muestra $\underline{x}$, $L(\theta;\underline{x})$. En la práctica, se tratará de encontrar los $\hat{\theta}$ en los que $L'(\theta;\underline{x}) = 0$ y $L''(\theta;\underline{x}) < 0$. En ocasiones $L(\theta;\underline{x})$ tiene una expresión exponencial con la que no es fácil operar. Por la monotonía del logaritmo, los extremos relativos de $L(\theta;\underline{x})$ y $L'(\theta;\underline{x})$ se mantienen en $\log(L(\theta;\underline{x}))$ y $\log(L'(\theta;\underline{x}))$. Se tiene, por tanto,

$$\arg_{\theta\in\Theta} \max L(\theta;\underline{x}) = \arg_{\theta\in\Theta} \max \log(L(\theta;\underline{x})). \tag{A.20}$$

DEFINICIÓN A.22 (Función de log-verosimilitud). Sea $\underline{x}$ una muestra aleatoria simple, de tamaño n, de X. Llamamos función de log-verosimilitud del parámetro θ a partir de la muestra $\underline{x}$ a $\ell(\theta;\underline{x}) = \log(L(\theta;\underline{x})$.

DEFINICIÓN A.23 (Vector de *score*). Sea X una variable aleatoria cuya función de densidad de probabilidad pertenece a una familia p-paramétrica $\{p(\,\cdot\,;\theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$. Llamamos vector de *score* de θ al gradiente del logaritmo de $p(x;\theta)$,

$$U_\theta(x) = \nabla \log(p(x;\theta)) = \left(\frac{\partial \log(p(\theta;x))}{\partial\theta_1}, \ldots, \frac{\partial \log(p(\theta;x))}{\partial\theta_p}\right)^{\mathrm{t}} \tag{A.21}$$

DEFINICIÓN A.24 (Información de Fisher). Sea X una variable aleatoria cuya función de densidad de probabilidad pertenece a una familia p-paramétrica. Llamamos matriz de

información de Fisher a la matriz de covarianzas de $U_\theta(X)$, $\mathcal{I}(\theta) = \operatorname{Cov}[U_\theta(X)]$, que es cuadrada de orden p.

PROPOSICIÓN A.20. *Dado un vector aleatorio X de dimensión n y el espacio p-paramétrico $\Theta \subseteq \mathbb{R}^p$, $\mathrm{E}[U_\theta(X)] = 0 \in \mathbb{R}^p$.*

Demostración. Cada coordenada $U_{\theta,j}(\theta; x)$, $1 \leq j \leq p$, de $U_\theta(x)$ es de media nula:

$$\begin{aligned} \mathrm{E}[U_{\theta,j}(X)] &= \int_{\mathbb{R}^n} U_{\theta,j}(X) p(x;\theta)\,\mathrm{d}\,x = \int_{\mathbb{R}^n} \frac{\partial \log(p(\theta;x))}{\partial \theta_j} p(x;\theta)\,\mathrm{d}\,x = \\ &= \int_{\mathbb{R}^n} \frac{1}{p(x;\theta)} \frac{\partial p(x;\theta)}{\partial \theta_j} p(x;\theta)\,\mathrm{d}\,x = \frac{\partial \int_{\mathbb{R}^n} p(x;\theta)\,\mathrm{d}\,x}{\partial \theta_j} = \frac{\partial 1}{\partial \theta_j} = 0, \end{aligned} \tag{A.22}$$

y la antepenúltima igualdad viene garantizada por el Teorema A.2. □

Luego, para $p = 1$, $\mathcal{I}(\theta) = \mathrm{E}[U_\theta(X)^2]$.

PROPOSICIÓN A.21. *Si X una variable aleatoria con función de densidad de probabilidad $p(x;\theta)$ dependiente del parámetro θ de dimensión p, entonces*

$$(\mathcal{I}(\theta))_{i,j} = -\mathrm{E}\left[\frac{\partial^2 \log(p(x;\theta))}{\partial \theta_j \partial \theta_k}\right],\ 1 \leq j,k \leq p. \tag{A.23}$$

Demostración. Su definición completa es larga y escapa al propósito de este apéndica. Algunas ideas que desarrollan la demostración son:

1. La nulidad de la esperanza de $U_\theta(X)$, hace que $\mathcal{I}(\theta)_{j,k} = \operatorname{cov}[U_{\theta,j}(X), U_{\theta,k}(X)]$, $1 \leq j,k \leq p$.

2. El hessiano de $\log(p(x;\theta))$ es

$$\mathrm{H}_{\log(p(x;\theta))} = \mathrm{J}_{\frac{U_\theta(x)}{p(x;\theta)}} = \frac{\mathrm{H}_{p(x;\theta)}}{p(x;\theta)} - \left(\frac{U_\theta(x)}{p(x;\theta)}\right)\left(\frac{U_\theta(x)}{p(x;\theta)}\right)^{\mathrm{t}} \tag{A.24}$$

3. Finalmente

$$\mathrm{E}\left[\mathrm{H}_{\log(p(x;\theta))}\right] = \left(\mathrm{E}\left[\frac{\partial^2 \log(p(x;\theta))}{\partial \theta_j \partial \theta_k}\right]\right)_{j,k} = -\mathcal{I}(\theta). \tag{A.25}$$

□

Si $p = 1$, la igualdad queda $\mathcal{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2 \log(p(x;\theta))}{\partial \theta^2}\right] = -\mathrm{E}\left[\frac{\partial^2 \ell(\theta;x)}{\partial \theta^2}\right]$.

PROPOSICIÓN A.22. *Sea $\underline{x} = (x_1, \ldots, x_n)$ es una muestra aleatoria de tamaño n de la variable aleatoria X. Si denotamos $\mathcal{I}(\theta)$ a la información de θ para una observación e $\mathcal{I}_n(\theta)$ a la información de θ para dada toda la muestra, entonces $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$.*

Demostración. Por la Proposición A.21 y la linealidad de la esperanza, es inmediato $\mathcal{I}_n(\theta) = -\mathrm{E}\left[\frac{\partial^2 p(\underline{x};\theta)}{\partial \theta^2}\right] = -\mathrm{E}[\ell''(\theta;\underline{x})] = -\sum_{i=1}^n \mathrm{E}[\ell''(\theta;x_i)] = n\mathcal{I}(\theta)$. □

A.5. Suavizado

A.5.1. *Splines* cúbicos naturales

DEFINICIÓN A.25 (*Spline* cúbico natural). Sean $\{(x_i, y_i)\}_{i=1}^n$ n puntos sobre $\mathbb{R}^2$, verificando $x_i < x_{i+1}$, $1 \leq i \leq n-1$. Una función $s : \mathbb{R} \to \mathbb{R}$ dada por trozos de polinomios de tercer grado sobre cada intervalo de puntos consecutivos, $]x_i, x_{i+1}[$, es un *spline* cúbico natural si verifica:

1. $s(x_i) = y_i$, en cada $1 \leq i \leq n$,
2. s, s' y s'' existen y son continuas, y
3. $s''(x_1) = s''(x_n) = 0$.

TEOREMA A.3. *De todas las funciones f de variable real que verifican:*

1. *f es continua sobre el intervalo $[x_1, x_n]$,*
2. *Existe f' y es continua sobre el intervalo $[x_1, x_n]$ y*
3. *$f(x_i) = y_i$, en cada $1 \leq i \leq n$, .*

el spline cúbico natural s es la función que mejor suaviza f en el sentido de minimizar $J(f) = \int_{x_1}^{x_n} f''(x)^2 \,\mathrm{d}\, x$.

Demostración. Sea $h(x) = f(x) - s(x)$, equivalentemente podemos escribir $f(x) = h(x) + s(x)$. Entonces $J(f) \equiv J(s) = \int_{x_1}^{x_n} (s(x) + h(x))^2 \,\mathrm{d}\, x$. Desarrollamos

$$J(s) = \int_{x_1}^{x_n} (s''(x) + h''(x))^2 \,\mathrm{d}\, x = \int_{x_1}^{x_n} s''(x)^2 + h''(x)^2 \,\mathrm{d}\, x + 2s''(x)h''(x) \,\mathrm{d}\, x \qquad \text{(A.26)}$$

Vamos a ver que la integral en el tercer sumando se anula, $\int_{x_1}^{x_n} s''(x)h''(x) \,\mathrm{d}\, x = 0$. $\int_{x_1}^{x_n} s''(x)h''(x) = s''(x_n)h'(x_n) - s''(x_1)h'(x_1) - \int_{x_1}^{x_n} s'''(x)h'(x) \,\mathrm{d}\, x = -\int_{x_1}^{x_n} s'''(x)h'(x) \,\mathrm{d}\, x$. Como $s'''(x) = \alpha_i$ dado cualquier $x \in]x_i, x_{i+1}[$, $1 \leq i \leq n-1$, se tiene $-\int_{x_1}^{x_n} s'''(x)h'(x) \,\mathrm{d}\, x = \sum_{i=1}^{n-1} \alpha_i \int_{x_i}^{x_{i+1}} h(x) \,\mathrm{d}\, x = 0$. Luego, la expresión A.26 concluye como

$$J(s) = \int_{x_1}^{x_n} s''(x)^2 \,\mathrm{d}\, x + \int_{x_1}^{x_n} h''(x)^2 \,\mathrm{d}\, x \geq \int_{x_1}^{x_n} s''(x)^2 \,\mathrm{d}\, x \qquad \text{(A.27)}$$

y la igualdad solo se da en el caso $h''(x) = 0$. Como además, como $h(x_1) = h(x_n)$, es necesario que la igualdad se dé si, y solo si, $h \equiv 0$, luego s es la función que mejor suaviza en el sentido de minimizar $J(f) = \int_{x_1}^{x_n} f''(x)^2 \,\mathrm{d}\, x$. □

Los y_i suelen ir acompañados de un ruido y es conveniente suavizar los (x_i, y_i) en lugar de interpolarlos. En lugar de fijar $y_i = s(x_i)$, consideremos $s(x_i)$ como n parámetros libres de s. El objetivo será minimizar $\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int_{x_1}^{x_n} s''(x)\,\mathrm{d}\,x$.

PROPOSICIÓN A.23. *Si f y f' son continuas sobre $[x_1, x_n]$, entonces s minimiza*

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_{x_i}^{x_n} f''(x)\,\mathrm{d}\,x \tag{A.28}$$

A.5.2. Regresión local: LOESS

Un método, usual en series de tiempo, para el ajuste de un conjunto de n pares de datos $\{(x_i, y_i)\}_{i=1}^{n}$ es el basado en regresiones (ponderadas) locales. Es descrito de manera práctica por Venables y Ripley (2002) [14].

DEFINICIÓN A.26 (*Span*). Dados $\underline{x} = (x_1, \ldots, x_n)$ e $\underline{y} = (y_1, \ldots, y_n)$, llamamos *span*, y lo denotamos $\alpha \in]0, 1]$, a la proporción de valores de $\underline{x}$ escogidos para el ajuste LOESS en el punto x_0.

Procedemos fijando las siguientes ideas:

1. Sean las funciones:

   a) $\mathrm{d} : (x, y) \in \mathbb{R} \times \mathbb{R} \to \mathrm{d}(x, y) \in [0, +\infty[$ una distancia sobre $\mathbb{R}$ y

   b) $W : u \in [0, 1[\to W(u) = (1 - u^3)^3 \in]0, 1]$.

2. Procedemos:

   a) Fijamos $x_0 \in \mathbb{R}$,

   b) $N_{x_0} \subset \underline{x}$ tal que $\mathrm{d}(x_i, x_0) < \mathrm{d}(x_l, x_0)$ para todo $x_i \in N_{x_0}$ y $x_l \in \underline{x} \backslash N(x_0)$, y $\mathtt{n}(N_{x_0}) = q = \lfloor \alpha n \rfloor$, *i.e.*, dado α, q es el $100\alpha\,\%$ de puntos de $\underline{x}$ más próximos a x_0.

   c) $\Delta_{x_0} = \mathrm{máx}_{x_i \in N_{x_0}}\, \mathrm{d}(x_0, x_i)$.

3. $\omega_i = W(\Delta_{x_0}^{-1}\mathrm{d}(x_0, x_i))$.

4. Calculamos el ajuste de mínimos cuadrados ponderados de y sobre N_{x_0}. Se toma el valor ajustado $\hat{y}_0 = S(x_0)$ a partir de los pesos ω_i.

La función `loess()` viene implementada de manera nativa mediante el paquete `stats` en `R`. Venables y Ripley (2002) [14] no dan ningún resultado teórico que justifique una u otra elección de α, pero es evidente por la definición de la función W que valores próximos a 1 supondrán ajustes más lisos (mayor suavizado), y próximos a 0, más rugososo (sobreajuste). Puede regularse mediante la instrucción `span` en `loess()`, y toma como valor por defecto `span=0.75`.

A.6. *Software* `R`

Para los cálculos computacionales, se ha recurrido a la versión `4.2.0` de `R`. Los paquetes involucrados en los cálculos, la representación gráfica y inlusión de código en LaTeX son:

1. `dobson v.0.4`
2. `faraway v.1.0.7`
3. `gam v.1.20.1`
4. `gamair v.1.0`
5. `ggplot2 v.3.3.6`
6. `knitr v.1.39`
7. `lmtest v.0.9`
8. `MASS v.7.3`
9. `mgcv v.1.8`
10. `splines v.4.2.0`
11. `stats v.4.2.0`

Bibliografía

[1] A.C. Atkinson. *Plots, Transformations and Regression*. Oxford University Press, 1985.

[2] D.R. Cox and E.J. Snell. *Applied Statistics: Principles and Examples*. Chapman and Hall, 1981.

[3] A.J. Dobson. *An Introduction to Generalized Linear Models*. Chapman and Hall, 2002.

[4] J.J. Faraway. *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regressions Models*. Chapman and Hall, 2016.

[5] F.A. Graybill. *Theory and Applications of the Linear Model*. Duxbury Press, 1976.

[6] T. Hastie and R. Tibshirani. *Generalized Additive Models*. *Statistical Science*, 1(3):297–310, 1986.

[7] E.L. Lehmann. *Theory of Point Estimation*. Wiley Publication in Mathematical Statistics, 1983.

[8] J.H. Mathews and K.D. Fink. *Métodos Numéricos con MATLAB*. Prentice Hall, 2000.

[9] J. Montanero. *Modelos lineales*. Servicio de Publicaciones Universidad de Extremadura, 2008.

[10] A.J. Nelder and R.W.M. Wedderburn. *Generalized Linear Models*. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[11] A.G. Nogales. *Estadística matemática*. Servicio de Publicaciones Universidad de Extremadura, 1998.

[12] V.K. Rohatgi. *An Introduction to Probability Theory and Mathematical Statistics*. John Wiley and Sons, Inc., 1976.

[13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[14] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag New York, 2002.

[15] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[16] S.N. Wood. *Generalized Additive Models. An Introduction with R*. Chapman and Hall, 2007.

[17] Y. Xie. *knitr: A Comprehensive Tool for Reproducible Research in R*. Chapman and Hall, 2014.