Modelos lineales generalizados: Modelos con coeficiente de variación constante y otros.

Laura Morte Sarmiento

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Directora del trabajo: Ana C. Cebrián Guajardo
Diciembre de 2018

Prólogo

Los modelos lineales generalizados se definen como una extensión de los modelos lineales clásicos. Estos, se dieron a conocer debido al estudio de Legendre, quien propuso un método aplicado a datos astronómicos basándose en un criterio intuitivo, en el que la variabilidad en las observaciones fue, en gran parte, debido a un error de medida.

Años más tarde, en 1809, Gauss introdujo la distribución Normal de errores como un método para describir la variabilidad. De esta manera, mostró que muchas de las propiedades de las estimaciones de mínimos cuadrados no dependen de la normalidad, sino de supuestos de varianza e independencia constantes. Una propiedad estrechamente relacionada se aplica a los modelos lineales generalizados.

Los modelos lineales generalizados, propuestos por Nelder y Wedderburn en 1972, fueron formulados como una manera de unificar distintos modelos estadísticos bajo un solo marco teórico. La introducción de estos modelos tuvo un gran impacto en la estadística aplicada y, actualmente, tienen aplicaciones en diferentes campos como la medicina, la geografía, la psicología o la climatología, entre otros.

Estos modelos son una solución especialmente adecuada para modelos de dependencia con datos no métricos, los cuales no se ajustan al modelo lineal clásico e incumplen los supuestos de linealidad y normalidad.

Resumen

Throughout this work, generalized linear models, which are a generalization of the linear models, are studied. Mainly, they are used when it is no possible to verify one or more features of linear models. For this reason, the first chapter starts remembering the properties of these ones in order to define the generalized linear models. Once the generalized linear models are defined, their random component, the systematic component, and the link function are defined explaining in the first one the exponential family. Moreover, in the link function section the canonical link is studied.

After that, it is shown how to estimate a generalized linear model with the maximum likelihood method and it is also explained how to calculate the deviance and the scaled deviance in these models giving a concrete formula for both of them.

Once all these topics have been seen, the three kinds of residuals of the generalized linear models are studied. They are known as Pearson residual, Anscombe residual and deviance residual.

Finally, the first chapter finishes with the explanation of the dummy covariates and the deviance test used in the selection of the covariates of the model.

In the second chapter, some particular cases of the generalized linear models are considered. It starts showing that generalized linear models with Poisson error belong to the exponential family, and giving an expression of its canonical link and its deviance. Their main characteristic is that its mean and its variance are the same. However, sometimes this assumption is not verified because the variance is bigger than the mean, thus overdispersion occurs. In order to solve this problem, an error with Binomial Negative distribution is introduced. In addition, as in models with Poisson error, it is shown the belonging to this distribution to the exponential family, and an expression of the canonical link is given.

After generalized linear models with Poisson and Binomial Negative error, generalized linear models with Gamma error are studied and, again, it is shown that this model belongs to the exponential family. In addition, a particular expression of its canonical link and deviance are given. In this case, apart from studying these characteristics, the estimation of the dispersion parameter is also taken into consideration.

Finally, in the third and last chapter, the theory seen during the work is used to create two models applied to heat waves using R Commander. It starts defining heat waves and doing a descriptive analysis of the covariates which are used to create both models.

The first model is applied to the study of the length of heat waves in Zaragoza during May, June, July, August and September in an interval of time of 65 years, from 1951 to 2016. For that model, both Poisson distribution or Binomial Negative distribution can be used. The most frequent count data distribution is Poisson distribution, for this reason it is started creating a model with this distribution.

When the model is created, the existence of overdispersion is tested. As it is shown, overdispersion occurs, so a new model with Binomial Negative distribution must be studied.

Once this model is generated, it is compulsory to check it by the application of different techniques. The first one is the validity of the model with a goodness-of-fit test, Kolmogorov-Smirnov test, with the DHARMa package.

Finally, in order to see the obtained result with this model, it is created a plot between the covariates and the fitted values.

On the other hand, the second model is applied in the study of the maximum intensity of heat waves under the same features than the first one. In this case, Gamma distribution is used and, as in the first model, once the model is finished it is necessary to chek it to ensure that it is correct using the same techniques than in the previous model.

Also, to finish the study of this model, a graphic between the covariates and the fitted values is created.

Índice general

Capítulo 1

Introducción a los modelos lineales generalizados

Los modelos lineales clásicos presentan varias restricciones, las cuales en ocasiones pueden no verificarse. Para solucionar este problema se estudian los modelos lineales generalizados. A lo largo de este capítulo se va a dar una definición específica de estos modelos así como de sus componentes. También se va a estudiar el proceso de estimación y se dará una definición de la desviación de los modelos lineales generalizados. A continuación se verán tres tipos de residuos: Pearson, Anscombe y desviación y, por último, se hará un breve estudio de las variables dummy y de un proceso para la selección de covariables en el modelo.

1.1. Definición

Un modelo lineal generalizado (GLM) es una extensión del modelo lineal clásico; por ello, antes de definir un modelo lineal generalizado conviene recordar las tres condiciones que debe verificar un modelo lineal:

1. Los errores se distribuyen normalmente.
2. La varianza es constante.
3. Las variables independientes están relacionadas con la variable dependiente de manera lineal.

De manera analítica se tiene que, dada una muestra $(Y_i, X_{i1}, \ldots, X_{ip})$ con i=1,...,n, la relación entre las observaciones Y_i y las variables independientes se expresa como:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i \qquad \text{con i=1,...,n.}$$

o equivalentemente

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

suponiendo que el error verifica $\varepsilon_i \sim N(0, \sigma^2)$, es decir, $Y_i \sim N(\mu, \sigma^2)$.

El modelo lineal generalizado se define como una generalización del modelo lineal anterior que puede ser utilizado cuando una o varias de sus condiciones no se satisfacen. En particular, esta generalización permite varianzas no constantes y errores con distribuciones no Normales, como Binomial, Poisson o Gamma entre otras. Además, tampoco requiere una relación lineal entre la respuesta y las variables independientes.

1.2. Componentes de un modelo lineal generalizado

La definición de un modelo lineal generalizado requiere de tres componentes: la componente aleatoria, la componente sistemática y la función de enlace, las cuales se explican a continuación.

1.2.1. Componente aleatoria

La componente aleatoria es el vector aleatorio $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ cuyos elementos son independientes y están idénticamente distribuidos con función de distribución perteneciente a la familia exponencial. La densidad de la familia exponencial es:

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

con *a, b y c* funciones conocidas, θ el parámetro canónico, que se verá más adelante, y ϕ un parámetro de dispersión.

Algunas de estas distribuciones son la distribución Normal, Gamma, Poisson o Binomial, entre otras. Cada una de ellas tiene una función *a, b y c* diferente. Por ejemplo, sabiendo que la distribución Normal viene dada por la función de densidad

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

se tiene, de acuerdo a (1.1), que $\theta = \mu$ y $\phi = \sigma^2$ y, por lo tanto, sus funciones *a, b y c* son:

$$a(\phi) = \phi \quad , \quad b(\theta) = \theta^2/2 \quad , \quad c(y;\phi) = -\frac{1}{2}\left(y^2/\sigma^2 + \log(2\pi\sigma^2)\right).$$

Media y varianza de la familia exponencial

Considerando $l(\theta,\phi;y) = \log f_Y(y;\theta,\phi)$ la función de log-verosimilitud para un y dado, la media y la varianza de $\mathbf{Y}$ se pueden obtener haciendo uso de relaciones conocidas como

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \tag{1.2}$$

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\frac{\partial l}{\partial \theta}\right)^2 = 0 \tag{1.3}$$

De (1.1) se tiene

$$l(\theta,\phi;y) = \{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)$$

y haciendo la primera y segunda derivada respecto a θ queda

$$\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} \tag{1.4}$$

$$\frac{\partial^2 l}{\partial \theta^2} = \frac{-b''(\theta)}{a(\phi)} \tag{1.5}$$

Tomando esperanzas en (1.4) y aplicando (1.2) se tiene que $E\left(\frac{y - b'(\theta)}{a(\phi)}\right) = 0$, de donde se deduce la media:

$$E(Y) = \mu = b'(\theta) \tag{1.6}$$

De la misma manera, tomando esperanzas en (1.4) y (1.5) y aplicando (1.3) se llega a:

$$E\left(\frac{-b''(\theta)}{a(\phi)}\right)+E\left(\frac{y-b'(\theta)}{a(\phi)}\right)^2=0.$$

Y, con el resultado de (1.6), queda:

$$E\left(\frac{-b''(\theta)}{a(\phi)}\right)+E\left(\frac{y-\mu}{a(\phi)}\right)^2=0$$

de donde se obtiene la varianza:

$$Var(Y)=b''(\theta)a(\phi). \tag{1.7}$$

De esta manera se tiene que la varianza de Y es el producto de las funciones $b''(\theta)$ y $a(\phi)$. La primera es conocida como *función varianza* y depende únicamente del parámetro canónico (y, por tanto, también de la media), mientras que $a(\phi)$ depende solo de ϕ. A lo largo del trabajo, se va a denotar $V(\mu)=b''(\theta)$ a la función varianza considerada como una función de μ.

La función $a(\phi)$ suele definirse como

$$a(\phi)=\phi/\omega$$

donde la constante ϕ es el parámetro de dispersión y ω es un peso conocido.

1.2.2. Componente sistemática

La componente sistemática está formada por variables explicativas, cuya combinación lineal se denomina predictor lineal. Dicho predictor se puede expresar de la siguiente manera

$$\eta_i=\sum_{j=1}^{p}\beta_j X_{ij}, \qquad \text{con i=1,...,n}$$

y, equivalentemente,

$$\eta=(\eta_1,...,\eta_n)=\mathbf{X}\beta$$

siendo $\mathbf{X}$ la matriz del modelo $(n\times p)$ y $\beta=(\beta_1,...,\beta_p)$ el vector de parámetros $(p\times 1)$.

1.2.3. Función de enlace

Para permitir relaciones no lineales entre la variable respuesta y las variables explicativas, se introduce la *función de enlace*, $g(\cdot)$, la cual relaciona el predictor lineal y la media de la variable respuesta como sigue:

$$g(\mu)=\mathbf{X}\beta=\eta.$$

El *enlace canónico* de una distribución es aquel que verifica

$$\theta=g(\mu)$$

siendo θ el parámetro canónico.

Cada distribución posee una única función de este tipo, aunque esta puede coincidir para distintas distribuciones. Algunas de ellas son:

Distribución Normal $\theta=\mu$

Distribución Gamma $\theta=\mu^{-1}$

Distribución Poisson $\theta = \ln(\mu)$

Distribución Binomial $\theta = \ln\left(\dfrac{\mu}{1-\mu}\right)$

1.3. Estimación del modelo

En la estimación de un GLM se busca estimar los parámetros β_i que definen el predictor lineal $\eta = \mathbf{X}\beta$ así como el parámetro ϕ. El método de estimación más habitual es el método de máxima verosimilitud. Vamos a ver cómo los estimadores máximos verosímiles de los parámetros β se pueden obtener por un procedimiento iterativo de mínimos cuadrados con pesos, véase [1].

Para ello se considera un modelo de regresión donde la variable dependiente no es y sino z, una forma linealizada de la función de enlace aplicada a y, y los pesos son funciones de los valores ajustados $\widehat{\mu}$.

Notar que z es la forma linealizada de la función de enlace de primer orden, es decir:

$$g(y) \simeq g(\mu) + (y-\mu)g'(\mu) = \eta + (y-\mu)\frac{d\eta}{d\mu} = z.$$

El proceso es iterativo ya que tanto la variable dependiente ajustada z como el peso W dependen de los valores ajustados, los cuales solo están definidos para estimaciones actuales. Dicho procedimiento es como sigue:

Sea $\widehat{\eta}_0$ el estimador actual del predictor lineal, y $\widehat{\mu}_0$ su correspondiente valor, $\widehat{\mu}_0 = g^{-1}(\widehat{\eta}_0)$. Entonces, se forma la variable dependiente ajustada:

$$z_0 = \widehat{\eta}_0 + (y - \widehat{\mu}_0)\left(\frac{d\eta}{d\mu}\right)\Bigg]_{\widehat{\mu}_0}$$

tal que $z_0 = \mathbf{X}\beta$.

Por su lado, el peso cuadrático se define como:

$$W_0^{-1} = \left(\frac{d\eta}{d\mu}\right)\Bigg]^2_{\widehat{\mu}_0} V_0$$

donde V_0 es la función varianza evaluada en $\widehat{\mu}_0$.

A continuación, se estima el modelo $z_0 = \mathbf{X}\beta$, con peso W_0 para obtener nuevos estimadores $\widehat{\beta}_1$ de los parámetros, a partir de los cuales se forma un nuevo estimador $\widehat{\eta}_1$ del predictor lineal, y se va repitiendo el proceso hasta que las modificaciones sean suficientemente pequeñas.

1.4. Desviación de un modelo lineal generalizado

La desviación de un GLM es una medida de bondad de ajuste formada a partir del logaritmo de razón de verosimilitudes. Dadas n observaciones, podemos ajustar modelos que contengan hasta n parámetros. El modelo más simple es el *modelo nulo* y, en el otro extremo, está el *modelo completo*, el cual tiene n parámetros, uno por observación. Este último modelo da una base de medida de discrepancia para un modelo intermedio con p parámetros.

Es conveniente expresar la log-verosimilitud en términos de μ en lugar de θ. Sea $l(\widehat{\mu},\phi;y)$ la log-verosimilitud maximizada sobre β para un valor fijo del parámetro de dispersión ϕ. La verosimilitud del modelo completo con n parámetros es $l(y,\phi;y)$. La discrepancia de un ajuste es proporcional a dos veces la diferencia entre $l(y,\phi;y)$ y $l(\widehat{\mu},\phi;y)$. Entonces, denotando $\widehat{\theta}=\theta(\widehat{\mu})$ el estimador del parámetro canónico del modelo bajo estudio y $\widetilde{\theta}=\theta(y)$ el del modelo completo, sus correspondientes funciones de log-verosimilitud son:

$$l(y,\phi;y)=\frac{y\widetilde{\theta}-b(\widetilde{\theta})}{a(\phi)}+c(y,\phi) \qquad ; \qquad l(\widehat{\mu},\phi;y)=\frac{y\widehat{\theta}-b(\widehat{\theta})}{a(\phi)}+c(y,\phi)$$

Considerando además $a_i(\phi)=\phi/\omega_i$, la diferencia entre las funciones de log-verosimilitud anteriores se puede escribir como (véase [1]):

$$D(y;\widehat{\mu})/\phi=2\sum_{i=1}^{n}\left[\omega_i\frac{y_i\widetilde{\theta}_i-b(\widetilde{\theta}_i)}{\phi}+c(y,\phi)-\omega_i\frac{y_i\widehat{\theta}-b(\widehat{\theta}_i)}{\phi}-c(y,\phi)\right]=$$
$$2\sum_{i=1}^{n}\omega_i\left[y_i(\widetilde{\theta}_i-\widehat{\theta}_i)-b(\widetilde{\theta}_i)+b(\widehat{\theta}_i)\right]/\phi$$

donde $D(y;\widehat{\mu})$ se conoce como *desviación*.

En ocasiones, el valor del parámetro ϕ puede ser desconocido, para que esto no afecte se puede escalar la desviación obteniendo así la *desviación escalada* definida como $D(y;\mu)/\phi$.

1.5. Residuos

Para comprobar la adecuación del ajuste de un modelo, se utilizan los residuos. A continuación se van a definir tres tipos de residuos: Pearson, Anscombe y residuos de la desviación.

Residuo de Pearson Se define como:

$r_{p,i}=\dfrac{y_i-\widehat{\mu}_i}{\sqrt{\widehat{V}(\widehat{\mu}_i)}}$, siendo $\sqrt{\widehat{V}(\widehat{\mu}_i)}$ la raíz cuadrada de la varianza del valor ajustado.

Este residuo tiene la desventaja de que su distribución para datos no Normales es asimétrica, lo que impide que tengan propiedades similares a las que tienen bajo normalidad.

Residuo Anscombe Anscombe define un residuo utilizando una función $A(y)$ en lugar de y, donde $A(\cdot)$ se elige para conseguir una distribución tan Normal como sea posible.

Wedderburn (véase [6]) mostró que, para las funciones de verosimilitud de los modelos lineales generalizados, la función $A(\cdot)$ viene dada por

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

Por lo tanto, los residuos Anscombe varían dependiendo de la distribución con la que se trabaje. Por ejemplo, en el caso de la distribución Poisson se definen como

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

Y en el caso de la Gamma como

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

Los valores que toma este residuo para distribuciones no Normales suelen ser muy parecidos a los que toma el residuo en la desviación, definido a continuación.

Residuo de la desviación Si se usa la desviación como una medida de discrepancia de un modelo lineal generalizado, cada observación contribuye una cantidad d_i a la desviación D tal que $\sum d_i = D$. Así, el residuo en la desviación se define como

$$r_{D,i} = sign(y - \widehat{\mu})\sqrt{d_i}$$

con i=1,...,n.

De nuevo este residuo varía dependiendo de la distribución con la que se trabaje. Por ejemplo, el residuo de la desviación para una distribución Poisson es:

$$r_D = sign(y - \widehat{\mu})\,\{2(y\log(y/\widehat{\mu}) - y + \widehat{\mu})\}^{1/2}.$$

Y para una Gamma:

$$r_D = sign(y - \widehat{\mu})\left\{2\left(\log(\widehat{\mu}_i/y_i) + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i}\right)\right\}^{1/2}.$$

1.6. Predictor lineal

1.6.1. Variables cualitativas

Los modelos lineales generalizados, al igual que los modelos lineales clásicos, permiten la introducción de variables tanto cuantitativas como cualitativas. La introducción de variables cualitativas requiere el uso de las variables dummy o ficticias, definidas a continuación.

La introducción de una variable cualitativa con k niveles en el modelo, debe permitir que el valor medio de la respuesta en cada nivel i de la variable pueda ser distinto, α_i. Para ello se debe introducir en el predictor lineal

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + ... + \alpha_k\mathbf{u}_k$$

donde $\mathbf{u}_j$, $j = 1,...,k$, son las variables binarias, las cuales toman el valor 1 si la observación pertenece al nivel j y cero en otro caso.

Se verifica que $\mathbf{u}_1 + \mathbf{u}_2 + ... + \mathbf{u}_k = \mathbf{1}$, donde $\mathbf{1}$ es el vector identidad. Dado que el término independiente de un modelo está asociado al vector $\mathbf{1}$, si en el modelo existe término independiente, se introducen $k-1$ variables dummy para evitar así una combinación lineal entre las variables, mientras que si no existe dicho término se introducen las k variables.

1.6.2. Selección de las covariables

Para la creación de un modelo es necesario saber qué covariables deben estar en este. Estas covariables se seleccionan contrastando si sus coeficientes pueden ser cero o no mediante inferencia sobre el vector de parámetros β.

La inferencia sobre el vector de parámetros β puede realizarse mediante distintos métodos, como el método de Wald o el test Score (véase [2]). En esta sección nos vamos a centrar en el *test de la desviación*.

Este test consiste en la comparación del ajuste de dos modelos anidados, es decir, de dos modelos con la misma distribución de probabilidad y con la misma función de enlace, pero siendo la componente lineal del modelo más simple, M_0, un caso particular de la componente lineal del modelo general M_1.

Se considera la hipótesis nula correspondiente a M_0

$$H_0 : \beta = (\beta_0, \beta_1, ..., \beta_q)$$

y la hipótesis correspondiente a M_1

$$H_1 : \beta = (\beta_0, \beta_1, ..., \beta_q, \beta_{q+1}, ..., \beta_p)$$

con $q < p < n$. Notar que el modelo M_0 se obtiene al imponer en M_1: $\beta_{q+1} = ... = \beta_p = 0$.

El objetivo del test es comparar H_0 y H_1 mediante la diferencia entre sus desviaciones. Denotando como D_0 a la desviación del modelo M_0 y como D_1 a la de M_1 se tiene

$$\triangle D = D_0 - D_1 = [2l(y;y) - 2l(\mu_0;y)] - [2l(y;y) - 2l(\mu_1;y)] = 2[l(\mu_1;y) - l(\mu_0;y)].$$

Si ambos modelos son adecuados, entonces $D_0 \sim \chi^2(n-q)$ y $D_1 \sim \chi^2(n-p)$, por lo tanto, $\triangle D = D_0 - D_1 = \chi^2(p-q)$, bajo ciertas condiciones generales de independencia, véase [2]. Si el contraste nos lleva a no rechazar H_0, se elige el modelo M_0 ya que es más sencillo.

Si el parámetro de dispersión es conocido, el modelo es adecuado. Sin embargo, en ocasiones este parámetro no es del todo conocido, esto ocurre en distribuciones como la Normal o la Gamma. Veamos cómo se puede solucionar el problema para esta última distribución, que será la que utilicemos a lo largo del trabajo:

Para un modelo con distribución Gamma, su desviación es (los cálculos se verán en 2.2.2):

$$D(\widehat{\mu};y) = \frac{2}{\sigma_V^2} \sum_{i=1}^{n} \left[-\log(y_i/\widehat{\mu}_i)) + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} \right]$$

siendo σ_V^2 el coeficiente de variación.

Sean $\widehat{\mu}_i(0)$ y $\widehat{\mu}_i(1)$ los valores ajustados para M_0 y M_1 respectivamente. Entonces:

$$D_0 = \frac{2}{\sigma_V^2} \sum_{i=1}^{n} \left[-\log(y_i/\widehat{\mu}_i(0)) + \frac{y_i - \widehat{\mu}_i(0)}{\widehat{\mu}_i(0)} \right] \quad ; \quad D_1 = \frac{2}{\sigma_V^2} \sum_{i=1}^{n} \left[-\log(y_i/\widehat{\mu}_i(1)) + \frac{y_i - \widehat{\mu}_i(1)}{\widehat{\mu}_i(1)} \right].$$

Se asume que M_1 es un modelo adecuado y, por tanto, $D_1 \sim \chi^2(n-p)$. Si M_0 también lo es, $D_0 \sim \chi^2(n-q)$ y, de esta manera se tiene $\triangle D = D_0 - D_1 = \chi^2(p-q)$.

Se puede definir el siguiente estadístico que no depende del parámetro σ_V^2:

$$F = \frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} = \frac{\sum_{i=1}^{n} \left[-\log(y_i/\widehat{\mu}_i(0)) + \frac{y_i - \widehat{\mu}_i(0)}{\widehat{\mu}_i(0)} \right] /(p-q)}{\sum_{i=1}^{n} \left[-\log(y_i/\widehat{\mu}_i(1)) + \frac{y_i - \widehat{\mu}_i(1)}{\widehat{\mu}_i(1)} \right] /(n-p)}.$$

Si H_0 es correcta, F tendrá una distribución central $F(p-q, n-p)$, mientras que si no lo es, el valor de F será mayor de lo esperado en la distribución $F(p-q, n-p)$.

Capítulo 2

Algunos casos particulares de GLM

Los modelos lineales clásicos asumen una varianza constante para todos los valores. Esta propiedad es necesaria para garantizar una estimación de los parámetros correcta. Sin embargo, a menudo se pueden dar casos en los que la varianza no es constante sino que aumenta con la media.

En este capítulo se van a tratar los modelos lineales generalizados con error de Poisson y Binomial Negativo así como sus propiedades principales y los modelos lineales generalizados con coeficiente de variación constante llegando así a definir la distribución Gamma y sus características.

2.1. Modelos lineales generalizados para variables de conteo

La distribución Poisson se usa generalmente para representar datos de conteo, es decir, la frecuencia de un determinado suceso. Por lo tanto toma valores enteros positivos, incluido el cero. Si la variable aleatoria Y sigue una distribución Poisson, su distribución de probabilidad es:

$$P(Y=y)=\frac{e^{-\mu}\mu^{y}}{y!} \qquad y=0,1,2,... \tag{2.1}$$

con $\mu>0$. Su media y varianza coinciden:

$$E(Y)=Var(Y)=\mu.$$

Pertenencia a la familia exponencial

Se puede demostrar que esta distribución pertenece a la familia exponencial vista en la Sección *1.2.1*. En efecto, escribiendo (2.1) como

$$\exp\{y\log\mu-\mu-\log(y!)\}$$

y denotando $\theta=\log\mu$, es decir, $\mu=e^{\theta}$, se tiene:

$$\exp\{y\theta-e^{\theta}-\log(y!)\}.$$

Considerando

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

queda demostrado que la distribución de Poisson pertenece a la familia exponencial.

Además, la media y varianza de esta distribución verifican (1.6) y (1.7):

$$E(Y)=b'(\theta)=e^{\theta}=\mu \qquad ; \qquad Var(Y)=b''(\theta)a(\phi)=e^{\theta}=\mu.$$

Enlace canónico

Para calcular el *enlace canónico* de la distribución de Poisson, consideramos $E(Y) = \mu = e^{\theta}$. Así se tiene que $\log \mu = \theta$.

Por tanto, el enlace canónico de esta distribución es:

$$\eta = \theta = \log \mu.$$

Desviación

Para n observaciones independientes, la función de log-verosimilitud es

$$l(\mu; y) = \sum_{i=1}^{n} (y_i \log \mu_i - \mu_i - \log(y_i)!).$$

Con ella, se puede calcular la función de desviación de la siguiente manera:

$$D(y; \widehat{\mu}) = 2\sum_{i=1}^{n} \{-y_i + y_i \log y_i + \widehat{\mu}_i - y_i \log \widehat{\mu}_i\} = 2\sum_{i=1}^{n} \{y_i \log(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)\}.$$

2.1.1. Sobredispersión

La propiedad más destacada de la distribución de Poisson es que la media y la varianza coinciden. Sin embargo, esta hipótesis puede ser muy restrictiva y, al trabajar con datos reales, con frecuencia la varianza es mayor que la media. A este hecho se le llama *sobredispersión*.

Cuando aparece un problema de sobredispersión, existen distintas soluciones posibles. Zeilis, Kleiber y Jackman proponen distintas alternativas (véase [4]). Aquí nos vamos a centrar en considerar un error con distribución Binomial Negativa.

2.1.2. Modelos lineales generalizados con error Binomial Negativo

La distribución Binomial Negativa se puede obtener considerando el número de ensayos Bernoulli independientes e idénticamente distribuidos hasta conseguir un número determinado de éxitos.

La función de probabilidad de una Binomial Negativa de parámetros r y p es

$$P(Y = y, r, p) = \binom{y + r - 1}{y} (1-p)^y p^r$$

con $y \in \mathbb{N} \cup \{0\}$, $p \in [0,1]$, $r \in \mathbb{R}^+$.

Su media y varianza son

$$E(Y) = \frac{r(1-p)}{p} \qquad , \qquad Var(Y) = \frac{r(1-p)}{p^2}.$$

Binomial Negativa en mixturas

La distribución Binomial Negativa también aparece cuando se considera la mixtura entre las distribuciones Poisson y Gamma. Consideramos una variable aleatoria Y con distribución Poisson cuya media no es constante sino que es una variable aleatoria Z con distribución Gamma tal que $E(Z) = \mu$ y $Var(Z) = \mu/\phi$. Se puede probar que, véase [3, pág. 6]:

$$P(Y = y; \mu, \phi) = \frac{\Gamma(y + \phi\mu)\phi^{\phi\mu}}{y!\Gamma(\phi\mu)(1+\phi)^{y+\phi\mu}} = \binom{y + \phi\mu - 1}{y} \left(1 - \frac{\phi}{1+\phi}\right)^y \left(\frac{\phi}{1+\phi}\right)^{\mu\phi} \qquad , \qquad \text{y=0,1,2,...}$$

Esta distribución es una Binomial Negativa, con $p = \dfrac{\phi}{1+\phi}$ y $r = \mu\phi$:

$$P(Y = y; \mu, \phi) = \binom{y+r-1}{y}(1-p)^y p^r.$$

cuya media y varianza es

$$E(Y) = \mu \qquad ; \qquad Var(Y) = \frac{\mu(1+\phi)}{\phi} = \mu + \frac{\mu}{\phi} = \mu\left(1 + \frac{1}{\phi}\right).$$

Observamos que en la varianza, μ está multiplicada por un factor mayor que 1, $1 + 1/\phi$, el cual es mayor cuanto menor sea ϕ, por lo tanto la distribución Binomial Negativa se puede utilizar en los casos en los que existe sobredispersión, es decir, en los casos en los que la varianza es mayor que la media.

De esta manera, dicha distribución Binomial Negativa se puede considerar como una generalización de la distribución de Poisson ya que tiene su misma media y varianza, añadiéndole a esta última un parámetro adicional para la sobredispersión.

Pertenencia a la familia exponencial

Si consideramos r como un valor fijo, la distribución Binomial Negativa pertenece a la familia exponencial. En efecto:

$$P(Y = y; r, p) = \binom{y+r-1}{y}(1-p)^y p^r = \exp\left\{y\log(1-p) + r\log p + \log\binom{y+r-1}{y}\right\}.$$

Considerando $\theta = \log(1-p)$, es decir, $\log p = \log(1-e^\theta)$, se tiene:

$$P(Y = y; r, \theta) = \exp\left\{y\theta - (-r\log(1-e^\theta)) + \log\binom{y+r-1}{y}\right\}.$$

Y, con

$$b(\theta) = -r\log(1-e^\theta) \qquad , \qquad a(\phi) = \phi = 1 \qquad , \qquad c(y,\phi) = \log\binom{y+r-1}{y}$$

queda demostrado que la distribución Binomial Negativa pertenece a la familia exponencial.

Además,

$$E(Y) = b'(\theta) = \frac{re^\theta}{1-e^\theta} = \frac{r(1-p)}{p}$$

y

$$Var(Y) = b''(\theta)a(\phi) = \frac{re^\theta(1-e^\theta) + re^\theta e^\theta}{(1-e^\theta)^2} = \frac{r(1-p)}{p^2}.$$

Enlace canónico

Para calcular el *enlace canónico* de esta distribución, consideramos

$$\mu = E(Y) = \frac{re^\theta}{1-e^\theta} = \frac{r}{e^{-\theta}-1}.$$

Entonces,

$$\mu^{-1} = \frac{e^{-\theta}-1}{r} \Rightarrow r\mu^{-1}+1 = e^{-\theta} \Rightarrow \log(r\mu^{-1}+1) = -\theta \Rightarrow \log\left(\frac{1}{r\mu^{-1}+1}\right) = \theta\ .$$

Y, como $\mu = E(Y)$, el enlace canónico de esta distribución es:

$$\eta = \theta = \log\left(\frac{\mu}{r+\mu}\right).$$

2.2. Modelos lineales generalizados con coeficiente de variación constante

A lo largo de esta sección se va a asumir que Y es una variable aleatoria con coeficiente de variación σ_V^2 constante. En consecuencia:

$$Var(Y) = \sigma_V^2 (E(Y))^2 = \sigma_V^2 \mu^2.$$

Para σ_V pequeño, una posibilidad para estabilizar la varianza y obtener una distribución Normal es transformar la respuesta mediante un logaritmo, esto es suponer que la distribución de Y es log-Normal. De esta manera se tienen los siguientes momentos:

$$E(\log(Y)) = \log(\mu) - \sigma_V^2/2 \qquad ; \qquad Var(\log(Y)) \simeq \sigma_V^2$$

Y así hemos llegado a obtener un modelo con varianza constante.

La demostración de los resultados anteriores se sigue aplicando el desarrollo en serie de Taylor:

$$\log Y \approx \log\mu + \frac{Y-\mu}{\mu} - \frac{(Y-\mu)^2}{2\mu^2}. \tag{2.2}$$

Para demostrar la primera expresión tomamos esperanzas en (2.2):

$$E(\log(Y)) = E(\log\mu) + E\left(\frac{Y-\mu}{\mu}\right) - E\left(\frac{(Y-\mu)^2}{2\mu^2}\right) = \log\mu - \frac{E(Y-\mu)^2}{2\mu^2} = \log\mu - \frac{Var(Y)}{2\mu^2} = \log\mu - \frac{\sigma_V^2\mu^2}{2\mu^2} = \log\mu - \frac{\sigma_V^2}{2}.$$

Por su lado, la demostración de la varianza se obtiene tomando varianzas en (2.2):

$$Var(\log(Y)) = Var(\log\mu) + Var\left(\frac{Y-\mu}{\mu}\right) - Var\left(\frac{(Y-\mu)^2}{2\mu^2}\right) \simeq \frac{Var(Y)}{\mu^2} = \frac{\sigma_V^2\mu^2}{\mu^2} = \sigma_V^2.$$

En otros casos es más conveniente no transformar la variable respuesta Y y utilizar una función de enlace. Si la parte sistemática del modelo es multiplicativa, se tiene, (véase [1]):

$$\log(E(Y)) = x^T\beta.$$

Con esta función de enlace, se tiene linealidad sin transformar la escala y una función varianza que es función cuadrática de la media. Con estas características, se pueden usar iterativamente mínimos cuadrados no lineales con pesos para obtener estimaciones para β, véase [1]. Este método de estimación es equivalente a asumir que Y sigue una distribución Gamma con $\nu = \dfrac{1}{\sigma_V^2}$ constante e independiente de la media, como se muestra a continuación.

2.2.1. La distribución gamma

Sea Y una variable aleatoria con distribución Gamma, $G(\mu,\nu)$. Su función de densidad es:

$$f(y)=\frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu}\exp\left(-\frac{\nu y}{\mu}\right)\frac{1}{y}, \qquad y\geq 0,\nu>0,\mu>0. \tag{2.3}$$

La esperanza y varianza son:

$$E(Y)=\mu \qquad ; \qquad Var(Y)=\mu^2/\nu.$$

El valor de ν determina la forma de la distribución. En este capítulo nos vamos a centrar en el caso en el que $\nu=\sigma_V^{-2}$ es constante para todas las observaciones.

2.2.2. Definición de modelos lineales generalizados con error Gamma

Pertenencia a la familia exponencial

La distribución Gamma pertenece a la familia exponencial. En efecto, cogiendo logaritmos a ambos lados de (2.3):

$$\log f(y)=-\log\Gamma(\nu)+\nu\log\nu+\nu\log y-\nu\log\mu-\nu y/\mu-\log y$$

y despejando $f(y)$:

$$f(y)=\exp\{-\log\Gamma(\nu)+\nu\log\nu+\nu\log y-\nu\log\mu-\nu y/\mu-\log y\}.$$

Denotando $\theta=-1/\mu$ y $\phi=1/\nu$ queda:

$$f(y)=\exp\{\nu y\theta+\nu\log(-\theta)-\log\Gamma(1/\phi)+1/\phi\log(1/\phi)+1/\phi\log y-\log y\}.$$

Por último, considerando

$$b(\theta)=-\log(-\theta) \quad , \quad a(\phi)=\phi \quad , \quad c(y,\phi)=-\log\Gamma(1/\phi)+1/\phi\log(1/\phi)+1/\phi\log y-\log y$$

queda demostrado que la distribución Gamma pertenece a la familia exponencial.

Además la media y la varianza son

$$E(Y)=b'(\theta)=-\frac{1}{\theta}=\mu$$

$$Var(Y)=b''(\theta)a(\phi)=\frac{1}{\theta^2\nu}=\frac{\mu^2}{\nu}.$$

Enlace canónico

Para calcular el *enlace canónico* de esta distribución consideramos $E(Y)=\mu=-1/\theta$. Despejando θ queda

$$\theta=-1/\mu.$$

Por lo tanto, el enlace canónico de la distribución Gamma es la función inversa:

$$\eta=\theta=\mu^{-1}.$$

La transformación recíproca no garantiza una estimación positiva de la media, la cual debe ser siempre positiva en una distribución Gamma, por lo que en ocasiones puede ser inadecuada esta función de enlace. Para evitar este problema, se utiliza el ***enlace logarítmico***, es decir

$$\eta=\log\mu.$$

Desviación

Para n observaciones independientes, la función de log-verosimilitud es:

$$l(\mu;y) = \sum_{i=1}^{n} \left[-\log\Gamma(\nu) + \nu\log\nu + \nu\log y_i - \nu\log\mu_i - \frac{\nu y_i}{\mu_i} - \log y_i \right].$$

A partir de ella, y sabiendo que

$$l(y;y) = \sum_{i=1}^{n} \left(-\log\Gamma(\nu) + \nu\log\nu + \nu\log y_i - \nu\log y_i - \nu - \log y_i \right)$$

$$l(\widehat{\mu};y) = \sum_{i=1}^{n} \left(-\log\Gamma(\nu) + \nu\log\nu + \nu\log y_i - \nu\log\widehat{\mu}_i - \frac{\nu y_i}{\widehat{\mu}_i} - \log y_i \right)$$

podemos calcular la desviación escalada mediante la diferencia entre las expresiones anteriores:

$$D(y;\widehat{\mu})/\phi = 2\sum_{i=1}^{n} \left[\nu\log(\widehat{\mu}_i/y_i) + \frac{\nu y_i}{\widehat{\mu}_i} - \nu \right] = 2\sum_{i=1}^{n} \nu \left[\log(\widehat{\mu}_i/y_i) + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} \right].$$

Multiplicando la desviación escalada por el parámetro de escala, $\phi = 1/\nu$, obtenemos la expresión de la desviación:

$$D(y;\widehat{\mu}) = 2\sum_{i=1}^{n} \left[\log(\widehat{\mu}_i/y_i) + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} \right].$$

2.2.3. Estimación del parámetro de dispersión

En un GLM, la matriz de covarianzas del vector de parámetros β es aproximadamente

$$cov(\widehat{\beta}) \simeq \sigma_V^2 (X^T W X)^{-1}$$

donde $W = diag\{(d\mu_i/d\eta_i)^2/V(\mu_i)\}$ es la matriz $n \times n$ diagonal de pesos, véase [1, pág. 40], ya que los estimadores de máxima verosimilitud de los parámetros β en el predictor lineal η pueden ser obtenidos a partir de un método iterativo de mínimos cuadrados con pesos. Si σ_V^2 es conocido, la matriz de covarianzas de $\widehat{\beta}$ puede calcularse directamente. Sin embargo, en la práctica suele ser desconocido pero puede ser estimado a partir de los residuos.

Para un modelo Gamma, la estimación de máxima verosimilitud de $\nu = \sigma_V^{-2}$ se obtiene como la solución de la siguiente ecuación:

$$D(y;\widehat{\mu}) = 2n\{\log\nu - \Gamma'(\nu)/\Gamma(\nu)\}.$$

Este resultado se puede comprobar igualando a cero la derivada de la función de log-verosimilitud $l(\mu,y)$ respecto a ν:

$$\frac{\partial l(\mu,y)}{\partial\nu} = \sum_{i=1}^{n} \left(-\frac{y_i}{\mu_i} - \log\mu_i + \log y_i + 1 + \log\nu - \frac{\Gamma'(\nu)}{\Gamma(\nu)} \right) = 0 \Rightarrow$$

$$n\left(\log\nu - \frac{\Gamma'(\nu)}{\Gamma(\nu)} \right) = \sum_{i=1}^{n} \left(\log(\mu_i/y_i) - \frac{y_i - \mu_i}{\mu_i} \right).$$

Al multiplicar por dos a ambos lados de la igualdad y aplicar la definición de desviación se tiene el resultado.

El principal problema del estimador de máxima verosimilitud es que es extremadamente sensible a errores de redondeo en observaciones pequeñas. Además, si la suposición de Gamma es falsa, ν^{-1} no estima consistentemente el coeficiente de variación. Por ello es preferible el siguiente estimador, véase [1]:

$$\widetilde{\sigma}_V^2 = \sum \{(y-\widehat{\mu})/\widehat{\mu}\}^2/(n-p)$$

el cual es consistente para σ_V^2 siempre y cuando β haya sido estimado consistentemente. Este estimador para σ_V^2 se puede usar en la fórmula $\sigma_V^2(X^TWX)^{-1}$ para obtener un estimador de $cov(\widehat{\beta})$.

Capítulo 3

Aplicación a las olas de calor

Una ola de calor se define como un periodo de varios días consecutivos en el cual la temperatura es excesivamente alta, es decir, supera un determinado umbral. Este umbral es diferente para cada localidad y se calcula como el percentil 95 de la temperatura máxima diaria de la localidad a estudiar en los meses de julio y agosto en un periodo de referencia de 1971 a 2000.

En este trabajo nos vamos a centrar en las olas de calor de Zaragoza, donde el umbral es de 37ºC, en los meses de mayo, junio, julio, agosto y septiembre en un intervalo de tiempo de 65 años, desde 1951 hasta 2016. Durante este tiempo, en Zaragoza se dieron 198 olas de calor.

El objetivo de este trabajo es realizar dos estudios relacionados con las olas de calor, aplicando la teoría vista en los capítulos anteriores. El primero de ellos va a consistir en el estudio de la duración de las olas de calor, es decir, en el número de días con observaciones por encima de un umbral, mientras que en el segundo se va a tratar la intensidad de las mismas, es decir, los grados por encima del umbral.

3.1. Descripción del problema

Los datos que se van a utilizar están proporcionados por AEMET (Agencia Estatal de Meteorología), y las variables que van a intervenir en la creación de los modelos anteriores son:

LZ: Duración de las olas de calor. Se define como el número de días por encima del umbral máximo definido a partir del cual se producen las olas de calor.

IxZ: Intensidad máxima de las olas de calor. Se define como el número de grados por encima de dicho umbral.

CTxm31Z: Tendencia a corto plazo de la temperatura.

CTTxZ: Tendencia a largo plazo de la temperatura.

cospi y sinpi: Términos estacionales. La estacionalidad, véase [7], se tiene en cuenta considerando como covariables la restricción a los meses de verano de las funciones armónicas que describen el ciclo anual:

$$cos(2\pi i) \text{ y } sin(2\pi i)$$

donde $i = 152/365, ..., 243/365$ señala la posición del día en el año.

3.2. Duración de la ola de calor

3.2.1. Selección del modelo

Se va a comenzar creando un primer modelo lineal generalizado para estudiar la duración de una ola de calor de las características nombradas al principio.

En cuanto a la distribución a elegir para la creación de este modelo, de acuerdo a las características de nuestros datos (son discretos y solo toman valores positivos), se debe seleccionar una distribución discreta que solo tome valores positivos. La distribución de conteo más frecuente es la de Poisson, por lo que vamos a empezar a trabajar con ella.

Notar que la distribución Poisson toma valores desde cero, mientras que la variable respuesta *LZ* comienza a tomar valores a partir de 1. Para solucionar este problema es necesario trabajar con la duración desplazada una unidad, es decir con *LZ-1 = LZmodif*.

Al trabajar con una distribución Poisson, para la elección de las covariables que van a formar parte del modelo, vamos a estudiar distintos contrastes de hipótesis utilizando el Test de Razón de Verosimilitudes mediante la orden **lrtest**. Con esta orden realizamos un Test de Razón de Verosimilitudes con una distribución Chi-cuadrado obteniendo un p-valor resultante, si este p-valor es menor que un nivel de significación 0.05 se rechaza la hipótesis nula.

Empezamos analizando si existe un comportamiento estacional, para ello consideramos la introducción de un armónico en el modelo, dado por *cospi* y *sinpi*. Entonces, con el modelo:

$$\log(LZmodif) = \beta_0 + \beta_1 sinpi + \beta_2 cospi$$

hacemos el siguiente contraste de hipótesis:

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \beta_1 \neq 0 \text{ y/o } \beta_2 \neq 0$$

Con el Test de Razón de Verosimilitudes obtenemos el siguiente resultado:

```
> lrtest(GLM.1,GLM.2)
## Likelihood ratio test
##
## Model 1: LZmodif ~ cospi + sinpi
## Model 2: LZmodif ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -280.10
## 2   1 -280.55 -2 0.8961     0.6389
```

Como el p-valor es 0.6389 > 0.05, no se rechaza la hipótesis nula, es decir, se rechaza la entrada del armónico en el modelo y, por tanto, no hay un comportamiento estacional en el modelo.

Veamos ahora si entra en el modelo la variable *CTxm31Z*. Para ello consideramos el modelo:

$$\log(LZmodif) = \beta_0 + \beta_1 CTxm31Z.$$

Y realizamos un contraste de hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Con el Test de Razón de Verosimilitudes obtenemos el siguiente resultado:

```
## Likelihood ratio test
##
## Model 1: LZmodif ~ CTxm31Z
## Model 2: LZmodif ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -264.16
## 2   1 -280.55 -1 32.772 0.00000001036 ***
```

Como el p-valor es $0.00000001036 < 0.05$ se rechaza la hipótesis nula, es decir, la variable *CTxm31Z* tiene una influencia significativa a un nivel $\alpha = 0{,}05$ en la respuesta.

Por otro lado, vamos a analizar si entra en el modelo la variable que muestra la evolución a largo plazo, es decir, veamos si entra *CTTxZ*. Para ello consideramos el modelo:

$$\log(LZmodif) = \beta_0 + \beta_1 CTTxZ.$$

Y realizamos un contraste de hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Con el Test de Razón de Verosimilitudes obtenemos el siguiente resultado:

```
## Likelihood ratio test
##
## Model 1: LZmodif ~ CTTxZ
## Model 2: LZmodif ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -277.77
## 2   1 -280.55 -1 5.5582    0.01839
```

Como el p-valor es $0.01839 < 0.05$ se rechaza la hipótesis nula, es decir, la variable *CTTxZ* tiene una influencia significativa a un nivel $\alpha = 0{,}05$ en la respuesta.

Vemos así que tanto la variable *CTTxZ* como la variable *CTxm31Z* entran por separado en el modelo, sin embargo el p-valor obtenido con la inclusión de esta última es mucho mayor al p-valor obtenido incluyendo *CTTxZ*.

Por último, veamos si, una vez la variable *CTxm31Z* está en el modelo, *CTTxZ* entra en este considerando el modelo:

$$\log(LZmodif) = \beta_0 + \beta_1 CTxm31Z + \beta_2 CTTxZ.$$

Y realizando un contraste de hipótesis:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Con el Test de Razón de Verosimilitudes obtenemos el siguiente resultado:

```
## Likelihood ratio test
##
## Model 1: LZmodif ~ CTxm31Z + CTTxZ
## Model 2: LZmodif ~ CTxm31Z
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -263.88
## 2   2 -264.16 -1 0.5739     0.4487
```

Como el p-valor es 0.4487 > 0.05, no se rechaza la hipótesis nula, es decir, se rechaza la entrada de *CTTxZ* en el modelo.

Por tanto, el modelo creado con la distribución Poisson para el estudio de la duración de una ola de calor es:

$$\log(LZmodif) = -9{,}22 + 0{,}27CTxm31Z.$$

Como hemos visto en el capítulo 2, en un modelo Poisson la media y varianza coinciden: $E(Y) = \mu = Var(Y)$; pero, si la varianza es mayor que la media: $Var(Y) = \mu + cf(\mu)$, siendo $f(\cdot)$ una función monónotona, se da el caso de sobredispersión. Para ver si en el modelo anterior existe sobredispersión vamos a realizar el siguiente contraste de hipótesis:

$$H_0 : c = 0$$
$$H_1 : c \neq 0$$

mediante la orden **testDispersion**. Esta orden aparece en el paquete *DHARMa* y realiza un test de dispersión basado en la simulación de los residuos que compara la dispersión de los residuos simulados con los residuos observados obteniendo un p-valor resultante. Si dicho p-valor es menor que 0.05, se rechaza la hipótesis nula aceptando así la existencia de sobredispersión en el modelo, mientras que si el p-valor es mayor que 0.05, no se rechaza la hipótesis nula y, por lo tanto, no se acepta la existencia de sobredispersión:

```
##  DHARMa nonparametric dispersion test
##
## data:  simulationOutput
## dispersion = 1.7108, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

Como el p-valor es 2.2e-16 < 0.05 se rechaza la hipótesis nula, es decir, hay sobredispersión en el modelo. Para solucionar el problema de la sobredispersión se utiliza, como hemos visto anteriormente, la distribución Binomial Negativa. Vamos entonces a crear un nuevo modelo para el estudio de la duración de las olas de calor en Zaragoza basado en la Binomial Negativa aplicando un procedimiento análogo al utilizado para la distribución Poisson.

Empezamos analizando si existe un comportamiento estacional, considerando la introducción de un armónico en el modelo, dado por *cospi* y *sinpi*:

$$\log(LZmodif) = \beta_0 + \beta_1 sinpi + \beta_2 cospi.$$

Realizamos un contraste de hipótesis:

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \beta_1 \neq 0 \text{ y/o } \beta_2 \neq 0$$

Con el Test de Razón de Verosimilitudes obtenemos el p-valor, que es 0.798. Como dicho p-valor es mayor que 0.05 no se rechaza la hipótesis nula, es decir, el armónico no entra en el modelo.

Veamos ahora si entra en el modelo la variable *CTxm31Z* considerando el modelo:

$$\log(LZmodif) = \beta_0 + \beta_1 CTxm31Z.$$

Y realizando un contraste de hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Con el Test de Razón de Verosimilitudes obtenemos el p-valor, que es 0.00003439. Como dicho p-valor es menor que 0.05 se rechaza la hipótesis nula, es decir, la variable *CTxm31Z* tiene influencia significativa a un nivel $\alpha = 0{,}05$ en la respuesta.

A continuación, vamos a analizar si la variable *CTTxZ* entra en el modelo, para ello consideramos:

$$\log(LZmodif) = \beta_0 + \beta_1 CTTxZ.$$

Y realizamos un contraste de hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Con el Test de Razón de Verosimilitudes obtenemos el p-valor, que es 0.08295. Como dicho p-valor es mayor que 0.05 no se rechaza la hipótesis nula, es decir, la covariable *CTTxZ* no entra en el modelo.

Por último, veamos si, una vez que *CTxm31Z* está en el modelo, la covariable *CTTxZ* entra en este considerando:

$$\log(LZmodif) = \beta_0 + \beta_1 CTxm31Z + \beta_2 CTTxZ.$$

Y realizando un contraste de hipótesis:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Con el Test de Razón de Verosimilitudes obtenemos el p-valor, que es 0.6054. Como dicho p-valor es mayor que 0.05 no se rechaza la hipótesis nula, es decir, la variable *CTTxZ* no entra en el modelo.

Por tanto, el modelo creado para el estudio de la duración de una ola de calor viene dado únicamente por la variable *CTxm31Z*. Veamos ahora la posible existencia de términos cuadráticos considerando el modelo:

$$\log(LZmodif) = \beta_0 + \beta_1 CTTxm31Z + \beta_2 (CTTxm31Z)^2.$$

Y realizando un contraste de hipótesis:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Con el Test de Razón de Verosimilitudes obtenemos el p-valor, que es 0.9168. Como dicho p-valor es mayor que $\alpha = 0{,}05$ no se rechaza la hipótesis nula, es decir, se descarta la inclusión de términos cuadráticos en el modelo, obteniendo así este modelo para el estudio de la duración de una ola de calor:

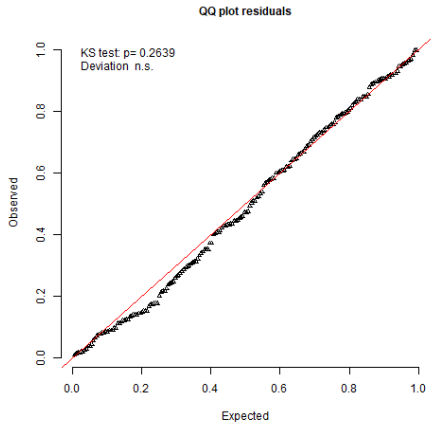
$$\log(LZmodif) = -9{,}02 + 0{,}27 CTxm31Z.$$

3.2.2. Validación del modelo

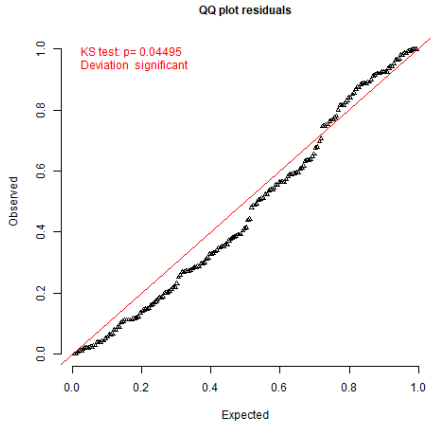
Una vez se ha hecho la selección de covariables, se debe comprobar la validez del modelo, para ello hay varias formas y métodos. Los errores de especificación de un GLM no se pueden evaluar de forma fiable con las gráficas estándar de residuos. La razón es que la distribución esperada de los residuos es una función de los valores ajustados.

Por ello, vamos a utilizar el paquete *DHARMa*. Con este paquete se puede realizar la prueba de *Kolmogorov-Smirnov*, que es un procedimiento de bondad de ajuste el cual permite medir el grado de concordancia que hay entre la distribución de un conjunto de datos y la distribución teórica específica.

El contraste de hipótesis consiste en considerar como hipótesis nula un modelo con error Binomial Negativo, obteniendo lo siguiente:

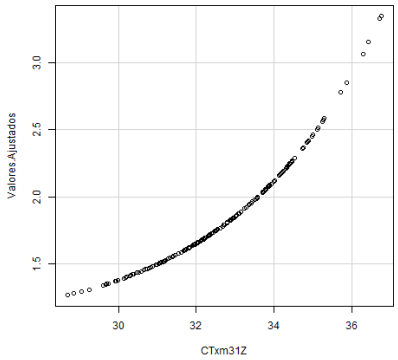


Como el p-valor es 0.2639 > 0.05 no se rechaza la hipótesis nula, es decir, la distribución Binomial Negativa es adecuada para este modelo. Además, gracias a este test nos podemos asegurar de que efectivamente la distribución Poisson no era adecuada para este modelo ya que el p-valor resultante es 0.04495 < 0.05 y la gráfica obtenida está mucho más desajustada:



De esta manera se observa que el modelo creado con una distribución Binomial Negativa para el estudio de la duración de las olas de calor en Zaragoza está bien construido.

Por último, vamos a hacer una gráfica enfrentando las covariables con los valores ajustados para ver el resultado obtenido con este modelo:



Como se ve en la gráfica, si la covariable es de 30º, la duración media de la ola de calor será de más de un día; mientras que si la covariable es de 36º, estará en torno a los tres días de duración.

3.3. Intensidad de la ola de calor

3.3.1. Selección del modelo

El segundo modelo lineal generalizado que se va a crear nos va a permitir estudiar la intensidad de las olas de calor en Zaragoza en los meses de mayo a septiembre desde 1951 hasta 2016. La distribución elegida para tratar con este tipo de datos es la distribución Gamma, con una función de enlace logarítmica para, como hemos visto en teoría, evitar una media negativa.

Como en la duración de la ola de calor, vamos a empezar analizando si existe un comportamiento estacional, para ello consideramos la introducción de un armónico en el modelo, dado por *cospi* y *sinpi*, en el modelo.

Como se vio en la Sección *1.6.2*, en el caso de la distribución Gamma el parámetro de dispersión no es conocido, por lo que para el proceso de selección de covariables en el modelo se debe utilizar la distribución F. Para ello vamos a estudiar distintos contrastes de hipótesis utilizando el Test de Desviación mediante la función **anova**.

Entonces, consideramos el modelo:

$$\log(IxZ) = \beta_0 + \beta_1 sinpi + \beta_2 cospi.$$

Hacemos el siguiente contraste de hipótesis:

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \beta_1 \neq 0 \text{ y/o } \beta_2 \neq 0$$

Con el Test de Desviación obtenemos el siguiente resultado:

```
> anova(GLM.1,GLM.2,test="F")
## Analysis of Deviance Table
##
## Model 1: IxZ ~ cospi + sinpi
## Model 2: IxZ ~ 1
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1       195     124.85
## 2       197     125.42 -2 -0.56881 0.4075 0.6659
```

Como el p-valor es 0.6659 > 0.05 no se rechaza la hipótesis nula, es decir, el armónico no entra en el modelo.

Veamos ahora si entra el término *CTxm31Z*, para ello consideramos el modelo:

$$\log(IxZ) = \beta_0 + \beta_1 CTxm31Z.$$

y realizamos un contraste de hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Con el Test de Desviación obtenemos el p-valor, que es 0.005928 < 0.05, por lo que se rechaza la hipótesis nula, es decir, la variable *CTxm31Z* tiene una influencia significativa a un nivel $\alpha = 0{,}05$ en la respuesta.

A continuación, veamos si entra en el modelo la variable *CTTxZ* considerando el modelo:

$$\log(IxZ) = \beta_0 + \beta_1 CTTxZ$$

y realizamos un contraste de hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Con el Test de Desviación obtenemos el p-valor, que es 0.05228 > 0.05, por lo que no se rechaza la hipótesis nula, es decir, la covariable *CTTxZ* no entra en el modelo.

Por último, veamos si, una vez dentro del modelo la covariable *CTxm31Z*, entra en el mismo *CTTxZ* con un razonamiento análogo al anterior, es decir, consideramos el modelo:

$$\log(IxZ) = \beta_0 + \beta_1 CTxm31Z + \beta_2 CTTxZ$$

y realizamos el contraste de hipótesis:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Con el Test de Desviación obtenemos el p-valor, que es 0.6077 > 0.05, por lo que no se rechaza la hipótesis nula, es decir, la covariable *CTTxZ* no entra en el modelo.

Por tanto, el modelo creado para el estudio de la intensidad de una ola de calor viene dado únicamente por la variable *CTxm31Z*. Veamos ahora la posible existencia de términos cuadráticos considerando el modelo:

$$\log(IxZ) = \beta_0 + \beta_1 CTxm31Z + \beta_2 (CTxm31Z)^2$$

Y realizando el contraste de hipótesis:

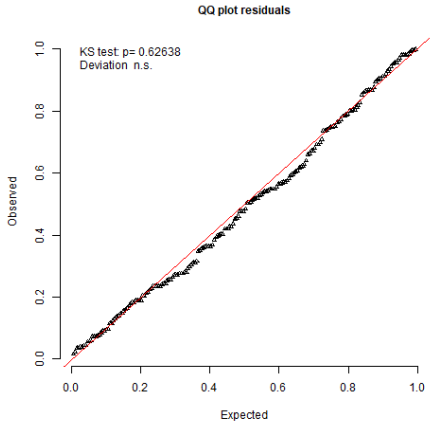
$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Con el Test de Desviación obtenemos el p-valor, que es 0.7455 > 0.05, por lo que no se rechaza la hipótesis nula, es decir, se descarta la existencia de un término cuadrático en el modelo, obteniendo así este modelo para el estudio de la intensidad de una ola de calor:

$$\log(IxZ) = -2{,}71584 + 0{,}09550 CTxm31Z.$$

3.3.2. Validación del modelo

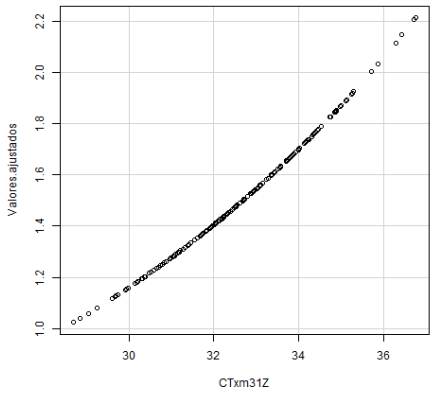
De la misma manera que en el modelo creado para la duración de una ola de calor, vamos a comprobar la validez del modelo creado para la intensidad con el paquete *DHARMa*, para ello realizamos la prueba de *Kolmogorov-Smirnov* considerando como hipótesis nula un modelo con error Gamma, obteniendo lo siguiente:



Como el p-valor es 0.62638 > 0.05 no se rechaza la hipótesis nula, es decir, la distribución Gamma es adecuada para este modelo.

De esta manera se observa que el modelo creado con una distribución Gamma para el estudio de la intensidad de las olas de calor en Zaragoza está bien construido.

Para terminar con el estudio del modelo, vamos a hacer una gráfica enfrentando las covariables con los valores ajustados:



Como se aprecia en la gráfica, si la covariable es de 30 grados, la media de la intensidad máxima es de 1.2; mientras que cuando la temperatura es de 36 grados, la media de dicha intensidad supera los 2 grados.

Bibliografía

[1] P. MCCULLAGH Y J.A.NELDER, *Generalized Linear Models*, 2.ª ed., Chapman and Hall, 1983.

[2] ANNETTE J. DOBSON, *Introduction to generalized linear models*, 2.ª ed., Chapman and Hall.

[3] ROBIN L. PLACKETT, *The Analysis of Categorical Data*, 1981.

[4] A. ZEILIS, C. KLEIBER Y S. JACKMAN, *Regression Models for Count Data in R*, 2008.

[5] STEPHEN M. STIGLER, *The history of statistics*, Harvard University Press.

[6] BARNDORFF-NIELSEN, 1978.

[7] J.ABAURREA, J.ASÍN, A.C.CEBRIÁN Y A. CENTELLES, *Modeling and forecasting extreme hot events in the central Ebro valley, a continental-Mediterranean area*, Global and Planetary Change, 2007.