

# REGRESIÓN GAMMA

Oihane Álvarez, Gabriel Carbonell, Daniel Hernández, Celia Sifre

# INTRODUCCIÓN

- ▶ El Modelo Lineal Generalizado (MLG) gamma no se encuentra entre los modelos más comúnmente utilizados, sin embargo, es de gran utilidad cuando nos enfrentamos a ciertos tipos de datos. Lo que tienen en común los datos que se modelizan con una Gamma es que son de tipo continuo y asimétricos por la derecha.

# INTRODUCCIÓN

- Y una variable aleatoria sigue una distribución gamma de parámetros  $\nu > 0$  y  $\lambda > 0$ ,  $Y \sim Ga(\nu, \lambda)$ , si su función de densidad de probabilidad es:

$$f(y) = \frac{1}{\Gamma(\nu)} \lambda^\nu y^{\nu-1} e^{-\lambda y} \quad y > 0$$

Donde  $\Gamma(\cdot)$  es la función gamma es:

$$\Gamma(\nu) = (\nu - 1)! \quad \text{si } \nu > 0$$

Esperanza:

$$E(y) = \nu/\lambda$$

Varianza:

$$V(y) = \nu/\lambda^2$$

# INTRODUCCIÓN

Sin embargo, para el propósito de un Modelo Lineal Generalizado, es conveniente reparametrizar la ecuación sustituyendo

$$\lambda = \nu/\mu$$

tal que la densidad quedaría:

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} e^{-\left(\frac{\nu y}{\mu}\right)} \quad y > 0$$

# INTRODUCCIÓN

Con la reparametrización,

$$E(y) = \nu/(\nu/\mu) = \mu$$

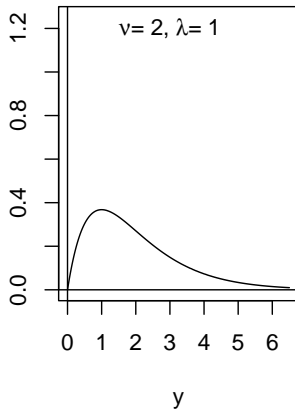
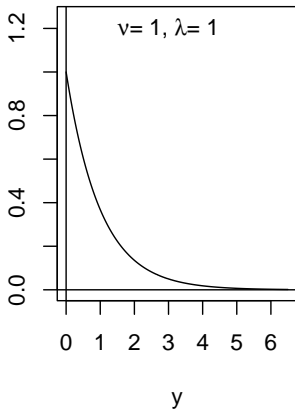
y la varianza es

$$Var(y) = \nu/(\nu^2/\mu^2) = \mu^2/\nu = E(y)^2/\nu$$

# INTRODUCCIÓN

Además  $\nu$  describe la forma y  $\lambda$  describe la escala de la distribución. Observamos que significa esto:

**Función de densidad distribuciones gamma**



# PROPIEDADES DISTRIBUCIÓN GAMMA

- ▶ La distribución exponencial es un caso particular de la distribución gamma,  $\text{Gamma}(1, \lambda) \sim \text{Expo}(\lambda)$
- ▶ La distribución  $\chi^2$  es un caso particular de la distribución gamma, donde  $\lambda = 1/2$  y  $\nu = \text{df}/2$  ( $\text{df} \equiv$  grados de libertad).

# EJEMPLOS SITUACIONES

Variables respuesta que toman valores continuos, positivos y asimétricos a la derecha. Se emplea comúnmente en estudios de fiabilidad, análisis de supervivencia, gestión de riesgos. . .

- ▶ Tiempo de vida útil de una maquina o electrodoméstico
- ▶ Altura a la que se inician las precipitaciones
- ▶ Consumo diario de energía (en millones de kW\*h) en una ciudad



# FAMILIA EXPONENCIAL (I)

Se dice que una variable aleatoria  $Y$  con distribución gamma pertenece a la familia exponencial si su función de densidad se puede expresar de la forma siguiente:

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

- ▶  $a()$ ,  $b()$ ,  $c()$  son funciones específicas de la distribución
- ▶  $\theta$  es un parámetro denominado natural o canónico
- ▶  $\phi$  es un parámetro muy relacionado con la dispersión.

## FAMILIA EXPONENCIAL (II)

Así nuestro objetivo es a partir de la función de densidad de una distribución gamma obtener la expresión anterior. Con este objetivo aplicaremos la exponencial del logaritmo a la función de densidad.

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} e^{-\left(\frac{\nu}{\mu} y\right)} \quad y > 0$$

Primeramente usamos el logaritmo.

$$\log(f(y)) = -\log(\Gamma(\nu)) + \nu \log(\nu) - \nu \log(\mu) + \nu \log(y) - \log(y) - \frac{\nu y}{\mu}$$

Así, aplicando la exponencial para despejar  $f(y)$  obtenemos:

$$f(y) = \exp \left\{ -\log(\Gamma(\nu)) + \nu \log(\nu) - \nu \log(\mu) - \frac{\nu y}{\mu} + (\nu - 1) \log(y) \right\}$$

## FAMILIA EXPONENCIAL (III)

Así denotando  $\theta = -1/\mu$ , donde  $\theta$  es el parámetro denominado natural o canónico y  $\phi = 1/\nu$ , donde  $\phi$  es el parámetro relacionado con la dispersión. Sustituyendo queda:

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

Así, finalmente tendremos que

$$b(\theta) = -\log(-\theta),$$

$$a(\phi) = \phi,$$

$$c(y, \phi) = -\log(\Gamma(1/\phi)) + 1/\phi \log(1/\phi) + (1/\phi - 1) \log(y)$$

# MODELO DE REGRESIÓN GAMMA

La variable respuesta  $Y_i$  es continua y positiva:

$$Y_i \sim \text{Gamma}(\mu_i, \nu_i)$$

Donde  $i \in \{i = 1, \dots, n\}$ .

Predictor lineal formado por una combinación lineal de las componentes  $X^{(1)}, X^{(2)}, \dots, X^{(p)}$  que indican tanto covariables como variables indicadoras e interacciones de ellas.

Hay varias funciones de enlace que unen el predictor lineal con la respuesta media, las cuales son: el logaritmo  $g(\mu) = \log(\mu)$ , la identidad  $g(\mu) = \mu$  y la inversa  $g(\mu) = 1/\mu$ .

Es posible llevar a cabo una transformación sobre la varianza y unirla con el predictor lineal.

El ajuste con regresión Gamma evita la realización de transformaciones de la variables respuesta.

# FUNCIONES *LINK*

Enlace logaritmo:

$$\log(\mu_i) = \beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)}, i = 1, \dots, n$$

$$\mu_i = \exp \left\{ \beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)} \right\}$$

$$\text{Var}(Y_i) = \frac{\mu^2}{\nu} = \frac{1}{\nu} \mu^2 = \frac{1}{\nu} \left( \exp \left\{ \beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)} \right\} \right)^2$$

Interpretación: como cambios porcentuales en  $E(Y)$  por cada incremento de una unidad en  $X_i$  ( $\% = 100 \cdot (e^{\beta_j})$ ).

## FUNCIONES *LINK* (II)

Enlace identidad:

$$\mu_i = \beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)}, i = 1, \dots, n$$
$$\text{Var}(Y_i) = \frac{\mu^2}{\nu} = \frac{1}{\nu} \mu^2 = \frac{1}{\nu} \left( \beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)} \right)^2$$

Interpretación: el incremento de una unidad en  $X_i$  hace crecer  $E(Y)$  en  $\beta_j$ .

Enlace inverso (canónico)

$$\frac{1}{\mu_i} = \beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)}, i = 1, \dots, n$$
$$\text{Var}(Y_i) = \frac{\mu^2}{\nu} = \frac{1}{\nu} \mu^2 = \frac{1}{\nu} \left( \frac{1}{\beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)}} \right)^2$$

# EJEMPLOS DE APLICACIÓN

El uso de gamma GLM es adecuado para datos continuos, positivos y con sesgo a la derecha. Estos son algunos ejemplos de aplicación para los diferentes enlaces.

Enlace logaritmo:

- ▶ *Tiempo de supervivencia*. Tiempo de Supervivencia de los Pacientes con Nefropatía Diabética (Grover et al., 2013).

Enlace identidad:

- ▶ *BMI*. Índice de masa corporal (IMC) (Kaggle, 2017).

Enlace canónico:

- ▶ *Número de reclamaciones*. Número de reclamaciones por daños en automóviles (IBM, 2021).

# AJUSTE DEL MODELO (I)

Ajuste del modelo por Ajuste por Mínimos Cuadrados Ponderados Iterados. Especificando una estimación inicial de  $\hat{\beta}$  se obtienen los parámetros del predictor lineal del modelo  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ .

Utilizando el estadístico de Wald, se puede contrastar si los valores de los parámetros  $\beta_i$  valen 0 y obtener un intervalo de confianza al  $(1 - \alpha)100\%$  de cada parámetro. Contraste:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Si  $\beta_i = 0$  la covariable  $X_i$  no tendrá influencia en la variable respuesta (*siempre en presencia del resto*).



## AJUSTE DEL MODELO (II)

Si no se conoce el parámetro de forma  $\nu$ , podemos conocer dicho parámetro y el de la dispersión  $\phi$  usando el método de máxima verosimilitud que es aproximadamente:

$$\hat{\phi} = \frac{1}{\nu} = \frac{D(y, \hat{\mu})}{n - p}$$

Donde  $n$  es el número total de datos y  $p$  el número parámetros.  $D$  es la DEVIANCE para una distribución Gamma, que se define de la siguiente manera,  $D(y, \hat{\mu}) = -2 \sum (\ln(y_i/\hat{\mu}_i) - ((y_i - \hat{\mu}_i)/\hat{\mu}_i))$

## AJUSTE DEL MODELO (III)

La estimación de la dispersión mediante esta expresión es sensible a valores pequeños de  $y_i$  y además no están definidas para cuando  $y_i = 0$ . Por ello, es preferible la estimación de la dispersión mediante el método de Pearson:

$$\hat{\phi} = \frac{1}{\nu} = \frac{X^2}{n - p}$$

Donde el estadístico de chi-cuadrado  $X^2$  de Pearson se basa en la diferencia al cuadrado entre los valores observados y los esperados, entre los esperados. Es decir,

$$X^2 = \sum_i (\text{observados}_i - \text{esperados}_i)^2 / \text{esperados}_i = \sum_i (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i$$

# DIAGNÓSTICO DEL MODELO (I)

Para valorar el ajuste del modelo, se emplea el estadístico DEVIANCE, y alternativamente a éste, se puede emplear el estadístico chi-cuadrado  $X^2$  de Pearson.

Para conocer si nuestro modelo es adecuado:

- ▶ Se comprueba que los residuos DEVIANCE o Pearson estén entre -2 y 2, así como si tienen un comportamiento normal.

# DIAGNÓSTICO DEL MODELO (II)

Para determinar si el modelo tiene un buen ajuste:

- ▶ El estadístico DEVIANCE será pequeño
- ▶ El estadístico DEVIANCE deberá distribuirse como una  $X^2$  con  $n - (p + 1)$  grados de libertad. Contraste:

$H_0$  : El modelo propuesto tiene un buen ajuste

$H_1$  : El modelo propuesto no tiene un buen ajuste

- ▶ La diferencia de DEVIANCES entre el modelo nulo y el modelo propuesto nos da una idea de la calidad del ajuste de dicho modelo. Además, esta aproximación es más precisa que la de la DEVIANCE misma.

# ELECCIÓN DEL MEJOR MODELO (I)

A nivel de mejor ajuste, tres estrategias:

- ▶ Diferencia de DEVIANCES. Un modelo más complejo siempre tendrá una DEVIANCE menor, pero se puede determinar si esa disminución de DEVIANCE es significativa mediante la diferencia de DEVIANCES.
- ▶ Se pueden emplear criterios AIC, AICc, BIC, etc. Estos criterios penalizan la complejidad del modelo y cuanto menor sea su valor, mejor será el ajuste del modelo.
- ▶ Cuando hay un número de modelos grande, se puede hacer una selección forward o backward, mientras que si el número es pequeño, se puede hacer un análisis detallado de los modelos.

# ELECCIÓN DEL MEJOR MODELO (II)

A nivel de mejor capacidad predictiva:

- ▶ Validación cruzada (CV). También sirve para seleccionar el conjunto de variables que mejor predicen la variable respuesta.

# BIBLIOGRAFÍA (I)

- ▶ Faraway, J.J. (2006). Extending the Linear Model with R. CRC/Chapman and Hall.
- ▶ Camargo Lozano, B. (2018). Regresión Gamma generalizada: Extensiones y aplicaciones al análisis de datos espaciales.
- ▶ Grover, Gurprit & Sabharwal, Alka & Mittal, Juhi. (2013). An Application of Gamma Generalized Linear Model for Estimation of Survival Function of Diabetic Nephropathy Patients.
- ▶ Johnson, P.E. (2014). GLM with a Gamma-distributed Dependent Variable.

## BIBLIOGRAFÍA (II)

- ▶ ibm.com (2021). Fitting a Gamma Regression to Car Insurance Claims (Generalized Linear Models). [online] Available at: <https://www.ibm.com/docs/en/spss-modeler/18.1.1?topic=smt-fitting-gamma-regression-car-insurance-claims-generalized-linear-models> [Accessed 5 May 2022].
- ▶ Tatman, R. (2018). Regression Challenge: Day 4 (Gamma Distribution). [online] Kaggle.com. Available at: <https://www.kaggle.com/code/rtatman/regression-challenge-day-4-gamma-distribution/notebook> [Accessed 6 May 2022].
- ▶ DHSC Analysts (2021). Chapter 11 Testing regression assumptions | Intermediate R - R for Survey Analysis. Bookdown.org.