Sesión 7: Modelos Lineales Generalizados

Modelos de la Familia Exponencial y Tópicos

Cristian E García

cegarcia@uao.edu.co

Maestría en Inteligencia Artificial y Ciencias de Datos
Facultad de Ingeniería
2025

Universidad Autónoma de Occidente

Outline

1. Familia Exponencial
   1.1. Introducción
   1.2. Definición
   1.3. Modelos

2. Regresión Beta y GAM

3. Ejemplos

Introducción

Los métodos estadísticos para múltiples variables suelen analizar cómo el resultado de una **variable de respuesta** está asociado o puede ser predicho por los valores de las **variables explicativas**.

Por ejemplo, un estudio podría analizar cómo la cantidad anual donada a la caridad está relacionada con variables explicativas como el ingreso anual de una persona, el número de años de educación, la religiosidad, la edad y el género. Para la **inferencia estadística**, estos métodos asumen una distribución de probabilidad para la variable de respuesta en cada combinación de valores de las variables explicativas, sin necesidad de asumir nada sobre la distribución de las variables explicativas.

Familia Exponencial

La densidad de una familia exponencial univariada para la variable respuesta y es definida como:

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

El logaritmo de la densidad está dado por:

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

El parámetro θ es llamado el parámetro natural o canónico. $b(\theta)$ es una función con primera y segunda derivada. ϕ es un parámetro de dispersión. w es un valor conocido, habitualmente representa el peso de cada observación.

uao

Familia Exponencial

Ejemplo 1: distribución normal

$$
\begin{aligned}
f(y \mid \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \\
&= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right]
\end{aligned}
$$

Aquí, $\theta = \mu$, $b(\theta) = \frac{\mu^2}{2}$, $\phi = \sigma^2$, $w = 1$ y $c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$. Además,

$$
b'(\theta) = \mu = \mathbb{E}(y),\ b''(\theta) = 1 \text{ y}
$$

$$
\mathbb{V}ar(y) = \phi b''(\theta) = \sigma^2
$$

Familia Exponencial

Ejemplo 2: distribución poisson

$$f(y \mid \lambda) = \mathbb{P}(Y = y) = \frac{\lambda^y \exp(-\lambda)}{y!}, \qquad y = 0, 1, \ldots \tag{3}$$

el logaritmo de la densidad es:

$$\ln f(y \mid \lambda) = y \ln(\lambda) - \lambda - \ln(y!) \tag{4}$$

Así,

$$f(y \mid \lambda) = \exp[y \ln(\lambda) - \lambda - ln(y!)] \tag{5}$$

El parámetro natural es $\theta = \ln(\lambda)$, $b(\theta) = \exp(\theta)$, $c(y, \phi) = \ln(y!)$, $\phi = 1$ y $w = 1$. $b'(\theta) = b''(\theta) = \exp(\theta) = \lambda$. Así $\mathbb{E}(y) = \lambda$ y $\mathbb{V}ar(y) = \lambda$.

Familia Exponencial

Ejemplo 3: distribución Bernoulli

$$f(y \mid \pi) = \pi^{y}(1-\pi)^{(1-y)}, \quad y \in \{0,1\}, \tag{6}$$

el logaritmo de la densidad es:

$$\ln f(y \mid \pi) = y \ln(\pi) - y \ln(1-\pi) + \ln(1-\pi) \tag{7}$$

El parámetro canónico es $\theta = \ln(\pi) - \ln(1-\pi) = \ln\left(\frac{\pi}{(1-\pi)}\right)$. Por otro lado, $\ln(1-\pi) = -\ln(1+\exp(\theta))$. La densidad en la forma de la familia exponencial nos queda:

$$f(y \mid \theta) = \exp(y\theta - \ln(1+\exp(\theta)) \tag{8}$$

uao

Familia Exponencial

en este ejemplo, $b(\theta) = \ln(1 + \exp(\theta))$, $\phi = 1$, $c = 0$ y $w = 1$.
$b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}$ y $b''(\theta) = \frac{\exp(\theta)}{(1+\exp(\theta))^2}$. Como $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$,
despejando π, nos queda

$$\pi = \frac{\exp(\theta)}{(1 + \exp(\theta))}$$

Así, $\mathbb{E}(y) = b'(\theta) = \pi$ y $\mathbb{V}ar(y) = b''(\theta) = \pi(1 - \pi)$

uao

Familia Exponencial

Distribution		$\theta(\mu)$	$b(\theta)$	ϕ
Normal	$N(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2
Bernoulli	$B(1, \pi)$	$\log(\pi/(1-\pi))$	$\log(1+\exp(\theta))$	1
Poisson	$Po(\lambda)$	$\log(\lambda)$	$\exp(\theta)$	1
Gamma	$G(\mu, \nu)$	$-1/\mu$	$-\log(-\theta)$	ν^{-1}
Inverse Gaussian	$IG(\mu, \sigma^2)$	$-1/(2\mu^2)$	$-(-2\theta)^{1/2}$	σ^2

(c) Expectation and variance

Distribution	$E(y) = b'(\theta)$	$b''(\theta)$	$\text{Var}(y) = b''(\theta)\phi/w$
Normal	$\mu = \theta$	1	σ^2/w
Bernoulli	$\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$	$\pi(1-\pi)$	$\pi(1-\pi)/w$
Poisson	$\lambda = \exp(\theta)$	λ	λ/w
Gamma	$\mu = -1/\theta$	μ^2	$\mu^2\nu^{-1}/w$
Inverse Gaussian	$\mu = (-2\theta)^{-1/2}$	μ^3	$\mu^3\sigma^2/w$

uao

Familia Exponencial

Dado un conjunto de covariables $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})^t$, las variables respuesta y_i se asumen condicionalmente independientes y pertenecientes a la familia exponencial:

$$f(y_i \mid \theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i)\right).$$

θ_i es el parámetro natural, ϕ es un parámetro de dispersión común independiente de i. Además, se cumple que:

$$\mathbb{E}(y_i) = \mu_i = b'(\theta_i), \quad \text{Var}(y_i) = \sigma_i^2 = \phi b''(\theta_i)/w_i\,.$$

Para datos no agrupados, $w_i = 1$.

Supuestos estructurales

La media condicional μ_i está conectada a los predictores lineales $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ a través de:

$$\mu_i = h(\eta_i) = h(\mathbf{x}_i'\boldsymbol{\beta})$$

o

$$\eta_i = g(\mu_i),$$

donde:

- h es una función uno a uno, dos veces diferenciable,
- g es la función de enlace, es decir, $g = h^{-1}$.

Modelos Lineales Generalizados

El modelo lineal general es completamente determinado por:

- La familia exponencial seleccionada (Gaussiana, Binomial, Poisson, Gamma, Gaussiana Inversa)
- La elección de la función de enlace
- La definición y selección de covariables

Cada familia exponencial tiene una función de enlace canónica única, dada por

$$\theta_i = \eta_i = \mathbf{x}_i' \boldsymbol{\beta} \, .$$

Para las funciones de enlace canónicas, el logaritmo de la log-verosimilitud siempre es cóncava, así el estimador de máxima verosimilitud (MV) siempre es único si es que existe. Además, se puede probar que las matrices de información esperadas y observadas coinciden:

$$F(\boldsymbol{\beta}) = H(\boldsymbol{\beta}) \, .$$

GLM

Modelos de regresión para datos binarios

- Se asumirá que la variable respuesta y_i, $i = 1, \ldots, n$, $y_i \in \{0, 1\}$
- Se cuenta además con un conjunto de predictores $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})$
- El objetivo principal es modelar y estimar el efecto de los predictores sobre la probabilidad condicional

$$\pi_{x_i} = \mathbb{P}(y_i = 1 \mid x_i) = \mathbb{E}(y_i \mid x_i)$$

- Con dicha especificación, $y_i \mid \boldsymbol{x}_i \sim \text{Ber}(\pi_{x_i})$

uao

Modelos de regresión para datos binarios

- La función de enlace que nos permite relacionar la esperanza de la distribución Bernoulli con los predictores es:

$$g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- La transformación con la función exponencial nos queda:

$$\frac{\pi}{1-\pi} = \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_k x_k)$$

lo que implica efectos exponenciales multiplicativos sobre las chances

uao

Modelos de regresión para datos binarios

- **Modelo Probit:** utiliza la función de distribución acumulada

$$\pi = \Phi(\eta) = \Phi(\boldsymbol{x}_i'\beta)$$

una desventaja de este modelo es que requiere evaluación numérica de Φ en el proceso de estimación máximo verosímil de $\boldsymbol{\beta}$

- **Modelo log–log Complementario:** utiliza la función de distribución acumulada de la distribución de valores extremos mínimos

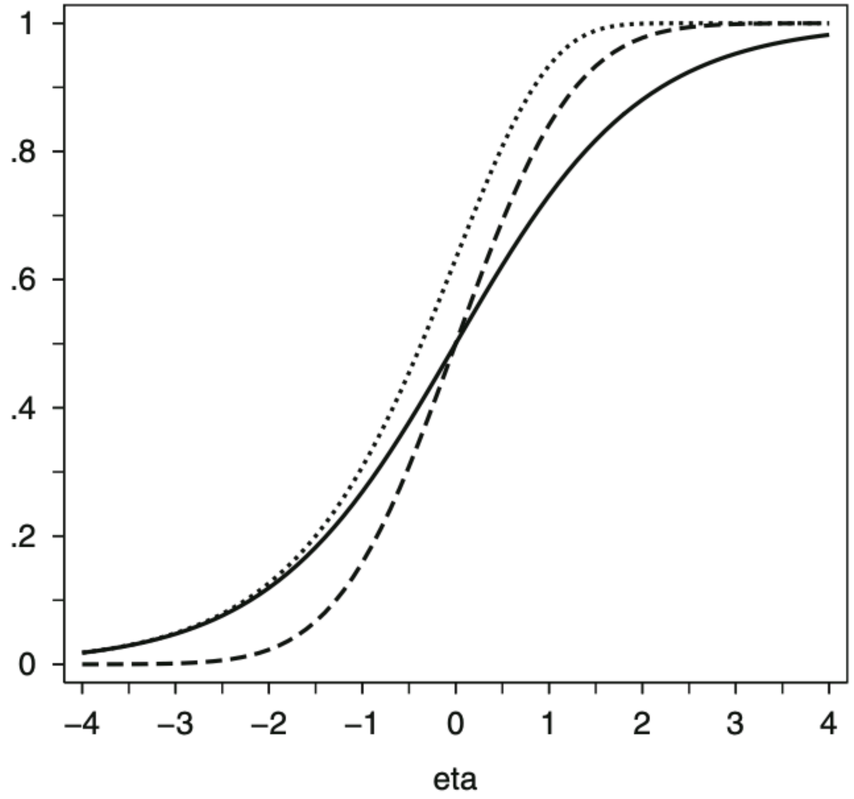
$$h(\eta) = 1 - \exp\left(-\exp(\eta)\right)$$

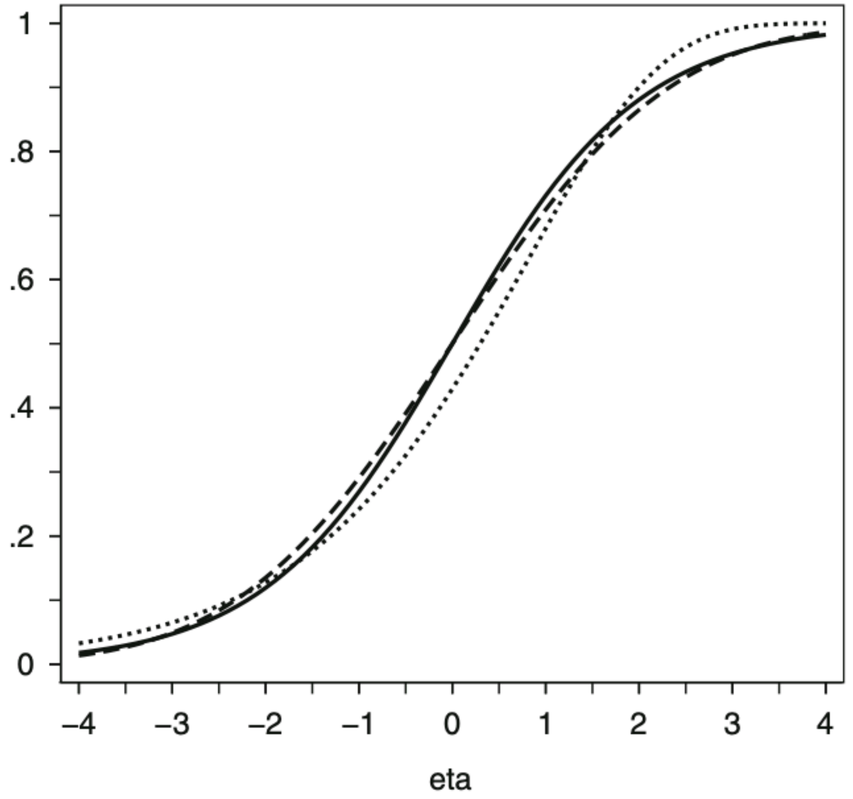
como función de respuesta, la cual tiene inversa

$$g(\pi) = \log\left(-\log\left(1-\pi\right)\right)$$

uao

Modelos de regresión para datos binarios





Modelos de regresión para datos binarios

Interpretación del modelo logit:

Basados en el predictor lineal

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

las chances

$$\frac{\pi_i}{1-\pi_i} = \frac{\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i)}{\mathbb{P}(y_i = 0 \mid \boldsymbol{x}_i)}$$

siguen un modelo multiplicativo

$$\frac{\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i)}{\mathbb{P}(y_i = 0 \mid \boldsymbol{x}_i)} = \exp(\beta_0)\exp(x_{i1}\beta_1)\cdots\exp(x_{ik}\beta_k)$$

uao

Modelos de regresión para datos binarios

Interpretación del modelo logit:

Si por ejemplo x_{i1} incrementa en una unidad a $x_{i1}+1$, los siguientes cambios aplican a la razón de chances

$$\frac{\mathbb{P}(y_i = 1 \mid x_{i1}, \ldots)}{\mathbb{P}(y_i = 0 \mid x_{i1}, \ldots)} \Big/ \frac{\mathbb{P}(y_i = 1 \mid x_{i1}+1, \ldots)}{\mathbb{P}(y_i = 0 \mid x_{i1}+1, \ldots)} = \exp(\beta_1)$$

$\beta_1 > 0 : \mathbb{P}(y_i = 1)/\mathbb{P}(y_i = 0)$ incrementa,
$\beta_1 < 0 : \mathbb{P}(y_i = 1)/\mathbb{P}(y_i = 0)$ decrece,
$\beta_1 = 0 : \mathbb{P}(y_i = 1)/\mathbb{P}(y_i = 0)$ se mantiene constante.

uao

Calculo numérico del estimador máximo verosímil EMV

- En algunas aplicaciones el EMV no puede ser determinado analíticamente
- EL sistema de ecuaciones obtenido después de igualar la función score a cero es no–lineal y no puede ser resuelto en forma cerrada
- Así, métodos numéricos para encontrar las raíces de la función score son requeridos
- Nos concentraremos en los algoritmos de Newton-Raphson y Fisher scoring

uao

Calculo numérico del estimador máximo verosímil EMV

Newton–Raphson

- El objetivo es calcular numéricamente las raíces de la función score, es decir la solución al sistema de ecuaciones

$$S(\theta) = 0$$

- Comenzando con un valor inicial $\theta^{(0)}$. En $\theta^{(0)}$ es aproximada por una linea recta tangente
- Una solución aproximada y mejorada $\theta^{(1)}$ es obtenida como la raíz de la recta tangente
- La recta tangente es obtenida mediante una expansión de primer orden en series de Taylor de $S(\theta)$ en $\theta^{(0)}$

$$g(\theta) = S(\theta^{(0)}) + S'(\theta^{(0)})(\theta - \theta^{(0)})$$

uao

Calculo numérico del estimador máximo verosímil EMV

- La raíz de la recta tangente provee una estimación mejorada

$$\theta^{(1)} = \theta^{(0)} - \frac{1}{S'(\theta^{(0)})} S(\theta^{(0)})$$

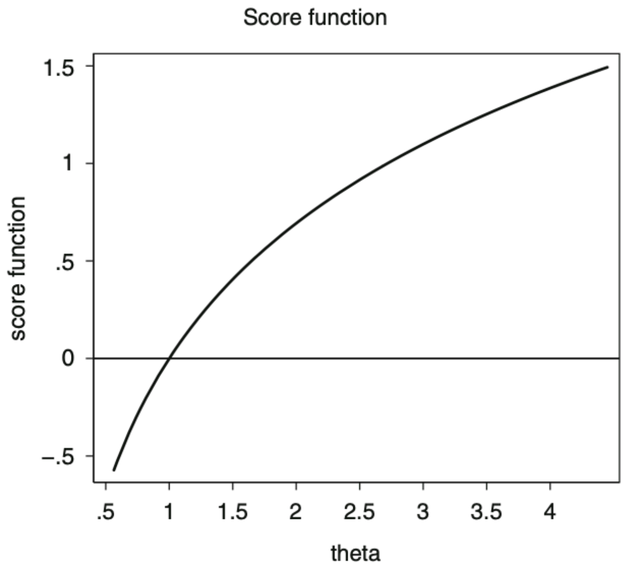
- Dado que $-S(\theta)$ es la matriz de información de Fisher observada $H(\theta)$, también podemos escribir

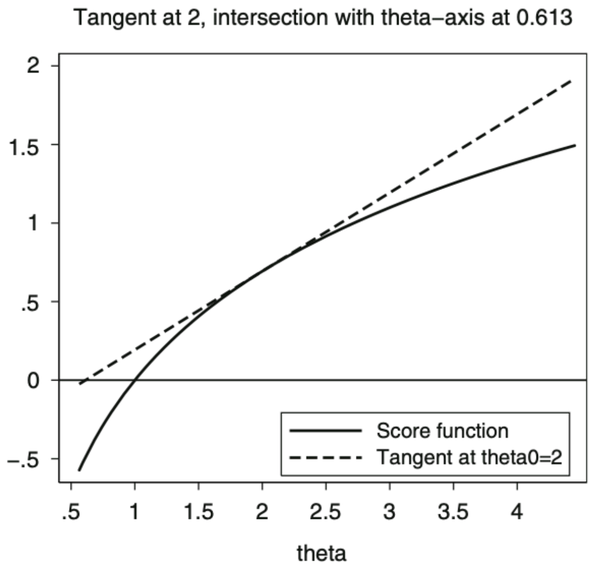
$$\theta^{(1)} = \theta^{(0)} + H(\theta^{(0)})^{-1} S(\theta^{(0)})$$

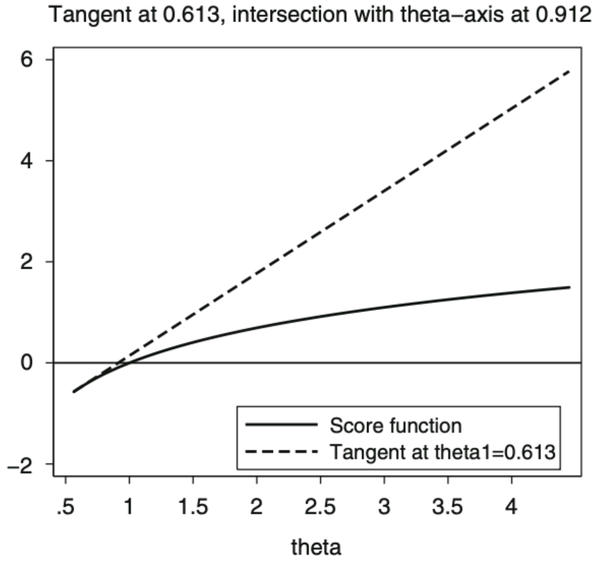
- Comenzando de $\theta^{(1)}$ se consigue una solución mejorada aproximada $\theta^{(2)}$ construyendo otra recta tangente a $S(\theta)$ en $\theta^{(1)}$ y calculando su raíz $\theta^{(2)}$
- El algoritmo continua iterativamente hasta que los cambios en las raíces son mínimos

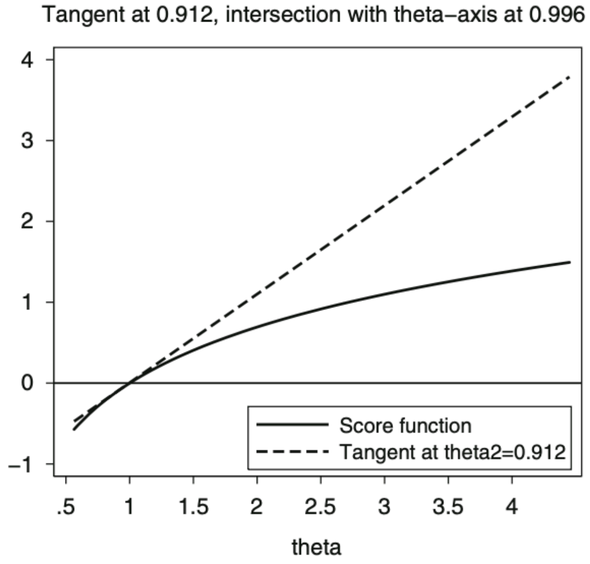
uao

Calculo numérico del estimador máximo verosímil EMV









Calculo numérico del estimador máximo verosímil EMV

La extensión del algoritmo al parámetro multivariado $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ es como sigue:

Sea $\boldsymbol{\theta}^{(t)}$ la solución aproximada de $S(\boldsymbol{\theta}) = \boldsymbol{\theta}$ en la iteración t. Una iteración mejorada está dada por:

$$\begin{aligned} \boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \left(\frac{\partial \boldsymbol{S}(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} \right)^{-1} \boldsymbol{S}(\boldsymbol{\theta}^{(t)}) \\ &= \boldsymbol{\theta}^{(t)} + \boldsymbol{H}(\boldsymbol{\theta}^{(t)})^{-1} \boldsymbol{S}(\boldsymbol{\theta}^{(t)}) \end{aligned}$$

El método de Fisher Scoring es obtenido si reemplazamos la matriz de información observada por la matriz de información esperada $\boldsymbol{F}(\boldsymbol{\theta}^{(t)})$

uao

Pruebas de Hipótesis Lineales

- Consideraremos hipótesis de la forma $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$ versus $H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}$,
- $\boldsymbol{C}$ es una matriz que tiene rango $r \leq p$
- Las pruebas más utilizadas son los test de razón de verosimilitud, test de score y el test de Wald.
- Todos estos test siguen una cierta distribución, pero de manera asintótica

uao

Pruebas de Hipótesis Lineales

- El test de Wald

$$w = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})^t [\boldsymbol{C}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}^t]^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$$

donde

$$\boldsymbol{F}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left(-\frac{\partial^2 l(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}} \partial \hat{\boldsymbol{\beta}}^t} \right)$$

- El test mide la distancia entre la estimación $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ y el valor hipotético $\boldsymbol{d}$ bajo H_0 ponderando por el inverso de la matriz de covarianza $\boldsymbol{C}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}^t$ de $\boldsymbol{C}\hat{\boldsymbol{\beta}}$

uao

Modelos para Datos de Conteo

Datos: La variable respuesta $y_i \in \{0, 1, 2, \ldots\}$ y se asumen condicionalmente independientes dadas las covariables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})$.

- **Modelo sin sobre-dispersión:**

$$y_i \sim \text{Po}(\lambda_i) \tag{13}$$

donde

$$\lambda_i = \exp\left(\boldsymbol{x}_i^t \boldsymbol{\beta}\right) \qquad \text{o} \qquad \ln \lambda_i = \boldsymbol{x}_i^t \boldsymbol{\beta}$$

- **Modelo con sobre-dispersión:**

$$\mathbb{E}(y_i) = \lambda_i = \exp\left(\boldsymbol{x}_i^t \boldsymbol{\beta}\right), \qquad \mathbb{V}ar(y_i) = \phi \lambda_i$$

uao

Modelos para Datos de Conteo

Estimación máximo verosímil: La estimación máximo verosímil se basa en la distribución Poisson,

$$f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{(\boldsymbol{x}_i^t \boldsymbol{\beta})^{y_i} \exp\left(-\boldsymbol{x}_i^t \boldsymbol{\beta}\right)}{y_i!} \tag{14}$$

Así, la log–verosimilitud de las n observaciones está dada por

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i(\boldsymbol{x}_i^t \boldsymbol{\beta}) - \exp\left(\boldsymbol{x}_i^t \boldsymbol{\beta}\right) \tag{15}$$

La función score está dada

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \exp\left(\boldsymbol{x}_i^t \boldsymbol{\beta}\right))$$

uao

Ejemplos

Ejemplo 1 (GLM)

Este es un diseño factorial fraccionado 2^{4-1} dirigido a procesos de sellado de parabrisas en la industria. Durante la operación de sellado aparecen daños en el producto, así, la variable respuesta es el número de partes buenas producidas en 1000 unidades; es lógico asumir que la distribución binomial es un modelo razonable. La siguiente tabla contiene la matriz diseño y la variable respuesta para este experimento (Lewis et al. 2001b).

Corrida	x_1	x_2	x_3	x_4	Partes buenas
1	1	1	1	1	338
2	1	-1	1	1	350
3	1	-1	-1	-1	647
4	1	1	-1	-1	826
5	-1	1	1	-1	917
6	-1	-1	1	-1	953
7	-1	-1	-1	-1	972
8	-1	1	-1	1	977

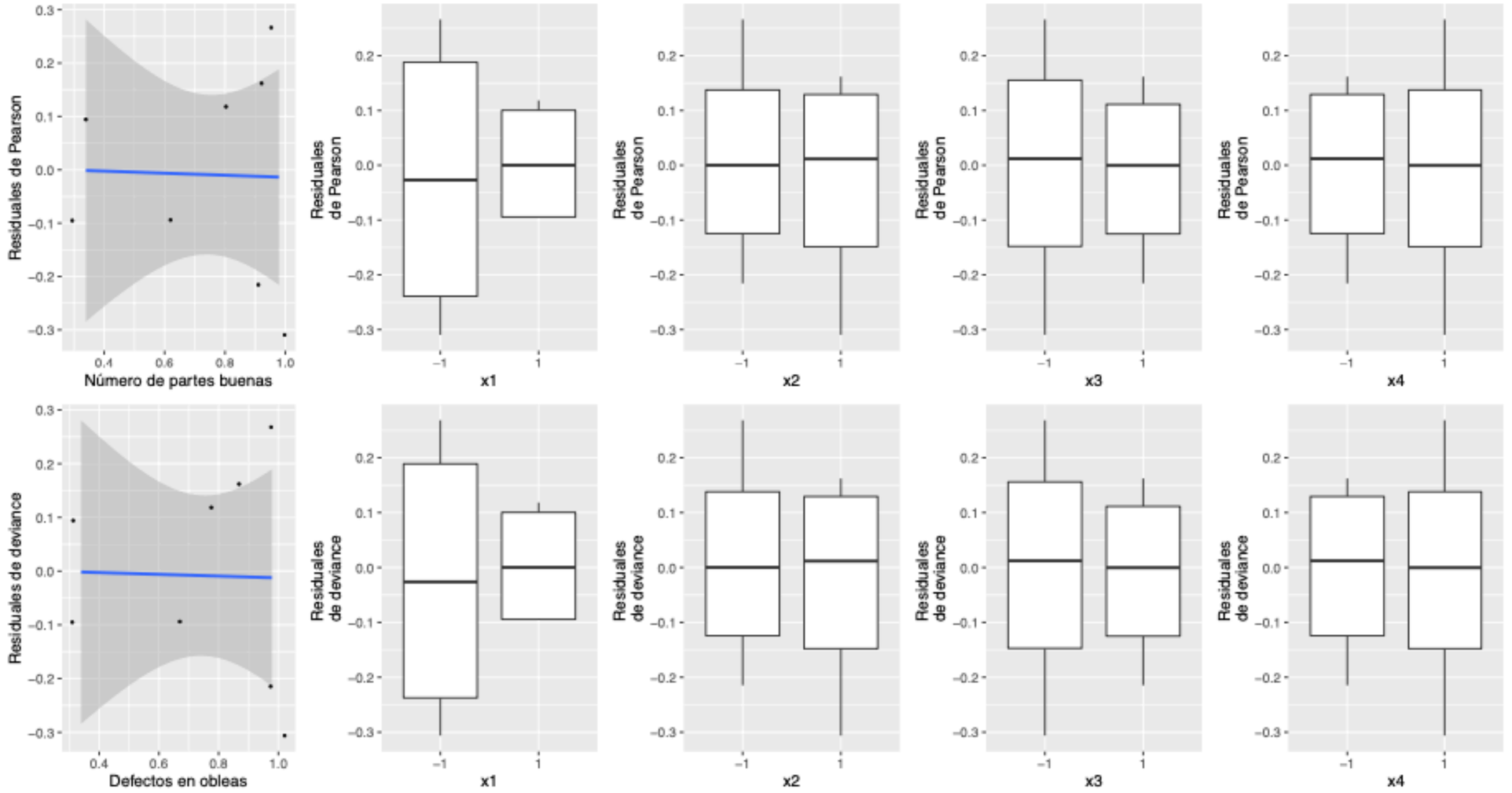
Ejemplo 1 (GLM)

Enlace	Deviance	Pseudo-R_1^2	Pseudo-R_2^2
probit	1.65	0.9997	0.9982
logit	**0.28**	**0.9999**	**0.9998**
cloglog	4.86	0.9988	0.9919
Cauchit	1.37	0.9999	0.9995
loglog	11.64	0.9952	0.9662

Ejemplo 1 (GLM)

Efecto	Coeficiente estimado	Error estándar	Wald	valor p	Intervalo de confianza
Intercepto	1.701	0.046	1364.85	0.00	(1.61, 1.79)
x_1	-1.483	0.046	1050.92	0.00	(-1.57, -1.39)
x_2	0.069	0.044	2.48	0.12	(-0.02, 0.17)
x_3	-0.667	0.047	205.63	0.00	(-0.76, -0.59)
x_4	-0.196	0.046	17.73	0.00	(-0.29, -0.11)
x_1x_2	0.154	0.044	12.58	0.00	(0.08, 0.24)
x_2x_3	-0.244	0.033	55.31	0.00	(-0.31, -0.18)

Ejemplo 1 (GLM)



Ejemplo 2 (GLM)

Los suelos tropicales, generalmente pobres en potasio (K), exigen fertilización con potasio cuando se cultivan con soja (*Glycine max* L.) para obtener rendimientos satisfactorios. La producción de soja se ve afectada por una larga exposición al déficit hídrico. Como el potasio es un nutriente implicado en el equilibrio hídrico de las plantas, por hipótesis, un buen aporte de potasio evita pérdidas de producción.

El objetivo de este experimento fue evaluar los efectos de las dosis de K y los niveles de humedad del suelo en la producción de soja. El experimento se realizó en invernadero, en macetas con dos plantas, que contenían 5 dm^3 de suelo. El diseño experimental fue de bloques completamente al azar con tratamientos en arreglo factorial 5 x 3. Las dosis de K fueron 0, 30, 60, 120 y 180 mg dm^{-3} , y la humedad del suelo osciló entre 35 y 40, 47,5 y 52,5 y 60 y 65% de la porosidad total (Serafim et al. 2012, para más detalles) .

Aplicaciones

¡Veámoslo en R!

uao

¿Preguntas?