Estadística e inferencia I

Segundo Cuatrimestre — 2024

Práctica 4: Modelos lineales generalizados

Modelos de dispersión exponencial

1. Determine cuáles de las siguientes distribuciones son EDMs identificando (en donde sea posible), el parámetro natural θ, la función cumulante $\kappa(\theta)$, el parámetro de dispersión ϕ y la función de normalización $a(y, \phi)$:

(*a*) Normal:

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

(*b*) Poisson

$$f(y; \mu) = \frac{e^{-\mu}\mu^y}{y!}$$

(*c*) Binomial. Considere $y \epsilon [0, 1]$ como la proporción de éxitos en n experimentos de Bernoulli con probabilidad de éxito μ

$$f(y; \mu, n) = \binom{n}{ny} \mu^{ny}(1-\mu)^{n-ny}$$

(*d*) Weibull

$$f(y; \alpha, \gamma) = \frac{\alpha}{\gamma}\left(\frac{y}{\gamma}\right)^{\alpha-1} e^{-(\frac{y}{\gamma})^\alpha}$$

(*e*) Weibull con $\alpha = 1$ y $\gamma = 1/\lambda$ (¿qué distribución es esta?)

(*f*) Geométrica

$$f(y; p) = p(1-p)^{y-1}$$

(*g*) Beta

$$p(y; \alpha, \beta) = \mathrm{B}[\alpha, \beta] y^{\alpha-1}(1-y)^{\beta-1}$$

(*h*) Gamma

$$f(y; \alpha, \lambda) = \frac{1}{\Gamma(\alpha)} e^{-\lambda y}\lambda^\alpha y^{\alpha-1}$$

(ayuda: realice el cambio de variable $\lambda = \alpha/\mu$ y haga los cálculos para la PDF $f(y; \alpha, \mu)$)

(*i*) Gaussiana inversa

$$f(y; \mu, \phi) = \frac{1}{\sqrt{2\pi y^3 \phi}} e^{-\frac{(y-\mu)^2}{2\phi y \mu^2}}$$

2. A partir de los resultados encontrados de θ, $\kappa(\theta)$ y ϕ para cada una de las distribuciones EDM del ejercicio anterior, calcule el valor medio $\mu = \mathrm{E}[y] = \frac{d\kappa(\theta)}{d\theta}$, la función varianza $V(\mu) = \frac{d^2\kappa(\theta)}{d\theta^2}$ y la varianza $\mathrm{var}[y] = \phi V(\mu)$ de cada distribución. Verifique que $\mathrm{E}[y]$ y $\mathrm{var}[y]$ calculadas de esta forma coinciden con las esperanzas y varianzas conocidas de las distribuciones.

3. Verifique que la aproximación de punto de ensilladura es exacta para las distribuciones normal y gaussiana inversa.

4. Calcule la función de devianza para las distribuciones normal y gamma. Muestre que si se reemplaza cada y_i por $100y_i$ (por ejemplo, un cambio de unidades de metros a centímetros) el valor numérico de la devianza para la distribución gamma no cambia pero para la distribución normal sí.

Modelos lineales generalizados

5. Muestre que las funciones de enlace canónicas son las siguientes:

(*a*) $g(\mu) = \mu$ cuando $Y \sim \mathrm{Normal}(\mu, \sigma^2)$

(*b*) $g(\mu) = \log(\mu)$ cuando $Y \sim \mathrm{Poisson}(\lambda)$

(*c*) $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ cuando $Y \sim \mathrm{Binomial}(p, n)$

6. Se le propone a un grupo de niños y niñas una actividad: construir torres lo más altas posible usando bloques cúbicos y cilíndricos. Se registran en https://github.com/franciscokordon/estainferencia/blob/main/data/blocks.csv las edades de los infantes, el número de bloques y el tiempo empleado. En este problema, solo consideramos el número de bloques utilizados y las edades.

(*a*) Graficar el número de bloques y vs las edades x. A partir del gráfico, proponer un MLG.

(*b*) Ajustar este MLG, es decir, encontrar estimaciones puntuales para los parámetros de regresión β.

(*c*) Determinar el error estándar de cada parámetro de regresión.

7. Nambe Mills es una fábrica de vajillas. En https://github.com/franciscokordon/estainferencia/blob/main/data/nambeware.csv figuran, entre otros datos, el diámetro y el precio de la vajilla producida durante un período de tiempo fijo.

(*a*) Graficar el precio y vs el diámetro x. Justificar por qué un MLG de Gamma es una opción razonable

(*b*) Ajustar este MLG, es decir, encontrar estimaciones puntuales para los parámetros de regresión β.

(*c*) Determinar el error estándar de cada parámetro de regresión.

8. Se busca modelar la cantidad de likes diarios que recibe un posteo en Instagram los 30 días posteriores a su publicación. Para eso, consideramos muestras $(x_1, y_1), \ldots, (x_n, y_n)$ de las variables aleatorias X, los días desde el posteo, que toma valores entre 0 y 30, y Y, la cantidad de likes por día, que es un proceso de Poisson que debe reflejar de manera razonable el fenómeno (por ejemplo, los primeros días se dan muchos likes y después van decayendo en el tiempo).

(*a*) Generar muestras sintéticas $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ para $n = 100$. Graficar y vs x.

(*b*) Proponer el Modelo Lineal Generalizado de Poisson y encontrar estimaciones puntuales para los parámetros de regresión β.

(c) Determinar el error estándar de cada parámetro de regresión. Obtener un intervalo de confianza de 0.95 para cada uno. Graficar el GLM encima de los datos incluyendo un intervalo de confianza para el modelo, usando el intervalo de confianza para β_1 (el parámetro que acompaña a x en la función lineal).