Estadística Bayesiana

Clase 19: Modelos Lineales Generalizados

Isabel Cristina Ramírez Guevara

Escuela de Estadística
Universidad Nacional de Colombia, Sede Medellín

Medellín, 6 de noviembre de 2020

Modelos Lineales Generalizados (GLM)

De acuerdo a la teoría de los modelos de regresión se tiene que la variable de respuesta Y es continua y se puede modelar mediante una distribución Normal. En algunos casos no se cumple esta condición, por eso es necesario el uso de un modelo lineal generalizado. El modelo de regresión lineal es un caso particular de estos modelos. El GLM generaliza la regresión lineal al permitir que el modelo lineal esté relacionado con la variable de respuesta a través de una función de enlace. Lo que hacen los GLM es establecer esa relación lineal no entre la media de la variable respuesta y los predictores, sino entre una función de la media de variable respuesta (función de enlace) y los predictores.

Modelos Lineales Generalizados (GLM)

Un GLM tiene tres componentes básicos:

- **Componente aleatoria:** identifica la variable respuesta y su distribución de probabilidad. Se supone que cada resultado Y de las variables dependientes se genera a partir de una distribución particular en la familia exponencial.

  En muchas aplicaciones, las observaciones de Y son binarias y se identifican como éxito y fracaso. Aunque de modo más general, cada Y_i indica el número de éxitos en un número fijo de ensayos, y se modeliza como una distribución binomial.

  En otras ocasiones cada observación es un conteo, con lo que se puede asignar a Y una distribución de Poisson.

Modelos Lineales Generalizados (GLM)

- **Componente sistemática:** especifica las variables explicativas (independientes o predictoras) utilizadas en la función predictora lineal, es decir las x_j. Estas se relacionan mediante:

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Esta combinación lineal de variables explicativas se denomina predictor lineal. Alternativamente, se puede expresar como un vector $(\eta_1, \cdots, \eta_n)$ tal que:

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

donde x_{ij} es el valor del j-ésimo predictor para el i-ésimo individuo, con $i = 1, \cdots, n$ y $j = 1, \cdots, k$.

Modelos Lineales Generalizados (GLM)

- **Función de enlace (link):** es una función del valor esperado de Y, $E(Y) = \mu$, como una combinación lineal de las variables predictoras. Suponga que la función de enlace es $g(.)$ y está dada por:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Por lo tanto la función de enlace relaciona las componentes aleatoria y sistemática.
Para $i = 1, \cdots, n$ se tiene:

$$\mu_i = E(Y_i) \quad \eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

La función de enlace más simple es $g(\mu) = \mu$, esto es, la identidad que da lugar al modelo de regresión lineal clásico.

Modelos de Regresión Poisson

Las variables aleatorias Poisson representan el número de eventos por unidad de tiempo. Si $Y_i \sim$ Poisson (θ_i), E(Y_i)=V(Y_i)=θ_i. En el modelo GLM se usa habitualmente el logaritmo de la media para la función de enlace, de modo que el modelo log-lineal es:

$$ln(\theta_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

Interpretación de los parámetros

En los modelos log-lineales de Poisson, el efecto de cada covariable X es lineal con la media logarítmica de Y, lo que resulta en un efecto exponencial de X, sobre la media de Y.

Suponga el caso donde se tiene una única variables explicativa, en ese caso la media de Y se puede expresar como:

$$ln(\theta_i) = \beta_0 + \beta X_i$$

$$\theta_i = e^{\beta_0} e^{\beta X_i} = B_0 B_1^{x_i}$$

donde $B_j = e^{\beta_j}$ para $j = 0, 1$, donde B_0 denota el número esperado de conteos (o Y) cuando la covariable es igual a cero ($X = 0$). La interpretación de β_1 es ligeramente diferente de la correspondiente en modelos normales ya que las diferencias relativas a la media se consideran en el caso de Poisson.

Interpretación de los parámetros

Denotemos por $\theta(x) = \mathrm{E}\,(Y|\ X = x)$, entonces

$$ln(\theta(x+1)) - ln(\theta(x)) = \beta_1$$

resultando en

$$\theta(x+1) = B_1\theta(x) = e^{\beta_1}\theta(x)$$

Por lo tanto, cuando la covariable X aumenta en una unidad, entonces el valor esperado de Y se vuelve igual a B_1 multiplicado por el valor esperado de Y para $X = x$. La interpretación puede basarse en el cambio porcentual esperado de Y dado por (B_1 -1)x100 cuando X aumenta en una unidad.

Interpretación de los parámetros

La interpretación de los parámetros para el caso del modelo de regresión Poisson múltiple es similar. La diferencia aquí es que para interpretar un cambio de una sola variable explicativa (digamos, X_j), las otras covariables deben permanecer constantes.

En la inferencia Bayesiana, la estimación puntual de los parámetros del modelo se hacen con la media o mediana de la distribución posterior.

Ejemplo

Se tiene una base de datos con el número de daños en aeronaves durante 30 misiones de ataque en la guerra de Vietnam. Por lo tanto se tienen 30 observaciones con las siguientes cuatro variables:

- **damage:** el número de ubicaciones dañadas de la aeronave.
- **type:** variable binaria que indica el tipo de avión (0 para A4; 1 para A6).
- **bombload:** la carga de bombas del avión en toneladas.
- **airexp:** los meses totales de experiencia de la tripulación aérea

Ejemplo

Estructura del modelo:

$$\text{damage}_i \sim \text{Poisson}(\lambda_i)$$

$$ln\lambda_i = \beta_1 + \beta_2 \text{type}_i + \beta_3 \text{bombload}_i + \beta_4 \text{airexp}_i$$

para $i = 1, \cdots, 30$. Con distribuciones a priori,

$$\beta_j \sim \text{Normal}(0, 0.001).$$

Ejemplo

Se obtienen los siguientes resultados:

	2.5 %	50 %	97.5 %
beta1	-2.22502500	-0.46310	1.249000
beta2	-0.45492250	0.57130	1.558000
beta3	0.03638875	0.17120	0.308000
beta4	-0.03039025	-0.01387	0.001982
B1	0.108000	0.6293	3.486
B2	0.634485	1.7710	4.747
B3	1.037000	1.1870	1.361
B4	0.970100	0.9862	1.002

El modelo estimado es:

$$ln\hat{\lambda}_i = -0.46310 + 0.57130\text{type}_i + 0.17120\text{bombload}_i - 0.01387\text{airexp}_i$$

Ejemplo

Al observar los intervalos posteriores al 95 % para los β_j´s el único que no contiene el cero es el que corresponde al coeficiente de la variable bombload. Por lo tanto solamente vamos a interpretar B_3, cada tonelada de bombas que se carga al avión aumenta el número esperado de daños en las aeronaves en un 18 %, cuando los otra covariables, tipo de avión y meses de experiencia de la tripulación, permanecen constantes.