El modelo lineal generalizado de la familia Tweedie y su aplicación en Salud Pública

RODRIGO VILLEGAS[*]

PROGRAMA DE BIOESTADÍSTICA
GRUPO DE CIENCIA DE DATOS PARA LA SALUD PÚBLICA
ESCUELA DE SALUD PÚBLICA, FACULTAD DE MEDICINA, UNIVERSIDAD DE CHILE

GRUPO TRANSDISCIPLINARIO PARA LA OBESIDAD DE POBLACIONES
UNIVERSIDAD DE CHILE

Resumen

En este artículo se revisa un tipo de modelo lineal generalizado que se basa en la distribución de Tweedie. Es una herramienta unificada para manejar datos semicontinuos (cuando se tiene un exceso de ceros y asimetría positiva). Se explora la estructura Poisson-Gamma compuesta, los métodos de estimación por máxima verosimilitud y se realiza una comparación crítica con modelos de dos partes, destacando la capacidad del enfoque Tweedie sobre otro tipo de modelos basados en las distribuciones clásicas (Normal, Binomial o Poisson).

1. Introducción: el desafío de la semicontinuidad

En la investigación aplicada en salud, a menudo la variable respuesta presenta una estructura híbrida denominada semicontinua (Min & Agresti, 2022). Esta se caracteriza por la coexistencia de una masa puntual en el valor cero y una distribución continua sesgada a la derecha para los valores positivos. Es decir, las respuestas tienen dos características problemáticas simultáneas:

- Exceso de ceros verdaderos ya que hay una proporción sustancial de individuos que no experimenta el evento de interés.
- Una distribución continua sesgada para valores positivos, es decir, entre los individuos que sí experimentan el evento, los valores son continuos, positivos y con asimetría a la derecha, frecuentemente similar a una distribución Gamma.

Ejemplos típicos que se observan en salud,

- Ingesta de alimentos: al evaluar la ingesta de pescado y mariscos una considerable proporción de encuestados podrían declarar cero consumo y algunos otros mostrar una distribución del consumo de tipo sesgada (Figura 1).
- Visitas al médico: algunos sujetos no acudieron al médico en el último año versus otros que acudieron varias veces.
- Gasto en medicamentos: algunos pacientes que no requieren medicación (cero) versus el gasto variable entre aquellos que sí lo requieren.

Los métodos tradicionales, como la regresión lineal ordinaria, fracasan ante estos datos debido a la violación de los supuestos de normalidad y homocedasticidad. Asimismo, las transformaciones logarítmicas *ad-hoc* ($\log(Y+1)$) introducen sesgos o pérdida de información y dependen de constantes arbitrarias que comprometen la validez de la inferencia estadística (C. Feng et al., 2014). Por otra parte los modelos basados en la distribución de Poisson, que es adecuada para conteos, no es capaz de modelar en forma correcta la magnitud de los eventos positivos al haber un exceso de ceros y en consecuencia una violación al supuesto de equidispersión.

2. Fundamentos matemáticos: la familia Tweedie

La distribución Tweedie, nombrada en honor al estadístico británico Maurice Tweedie (1919-1996), quien la estudió en 1984, y formalizada por Bent Jørgensen en 1987 (Jørgensen, 1987), pertenece a la clase de Modelos de Dispersión Exponencial (MDE). Esta clase de modelos es parte de una familia más amplia de distribuciones cuya forma es,

$$f(y|\theta,\phi) = a(y,\phi)\exp\left[\frac{y\theta-\kappa(\theta)}{\phi}\right],$$

donde las funciones $a(\cdot)$ y $\kappa(\cdot)$ son conocidas, θ es el parámetro natural y $\phi>0$ es el parámetro de dispersión. La media μ y la varianza de la variable aleatoria Y en un MDE está dado por $E(Y)=\mu=\kappa'(\theta)$ y $Var(Y)=\kappa''(\phi)$, respectivamente.

[*] rvillega@uchile.cl.

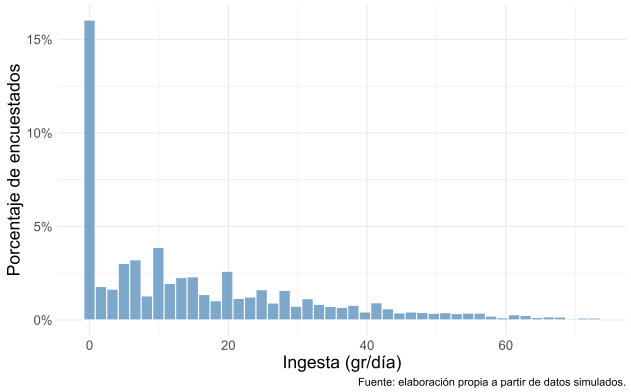

Figura 1. Ingesta diaria de pescados y mariscos (en gramos).

La familia de distribuciones Tweedie corresponde a un caso especial donde la relación entre la media y la varianza viene dada por

$$Var(\mu) = \phi\mu^p, \quad p \notin (0,1),$$

donde p es el *índice de Tweedie*. La naturaleza de esta distribución varía según el valor de p:

- Si $p = 0$ equivale a una distribución Normal.
- Si $p = 1$, equivale a una distribución de Poisson (representa el número de eventos que suceden durante un intervalo de tiempo dado o región específica).
- Si $1 < p < 2$, corresponde a una distribución compuesta Poisson-Gamma (datos semicontinuos).
- Si $p = 2$, equivale a una distribución Gamma (representa el tiempo hasta que se produce α veces un determinado suceso).
- Si $p = 3$ se aproxima a una distribución Gaussiana Inversa.

Dado que las distribuciones de Tweedie también pertenecen a la familia de distribuciones exponenciales, se pueden utilizar en el marco de los modelos lineales generalizados (McCullagh & Nelder, 1989).

2.1 La distribución compuesta Poisson-Gamma

Para el intervalo $1 < p < 2$, el modelo Tweedie asume que la variable observada Y es el resultado de la suma de N eventos aleatorios, donde cada evento tiene una intensidad X_i:

$$Y = \sum_{i=1}^{N} X_i, \quad N \sim \text{Poisson}(\lambda), X_i \sim \text{Gamma}(\alpha, \gamma)$$

Esta transformación permite modelar simultáneamente la probabilidad de ocurrencia (proceso de conteo) y la severidad del evento (proceso continuo), integrando ambos en una única función de verosimilitud. Se dice que la composición de esta distribución se da cuando el parámetro λ de la distribución de Poisson es a su vez una variable aleatoria que sigue una distribución Gamma.

3. Comparativa metodológica

La literatura identifica tres estrategias principales para el manejo de estos datos, resumidas en el Cuadro 1. A diferencia de los modelos de dos partes, como los modelos cero inflados o *Hurdle* (C. X. Feng, 2021), que requieren estimar un modelo basado en la función *logit* y otro modelo Gamma por separado, el modelo Tweedie permite una interpretación directa mediante una función de enlace (habitualmente logarítmica), esto es,

$$\log[E(Y|\mathbf{X})] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

El coeficiente $\exp(\beta)$ representa el efecto multiplicativo sobre la *media global esperada*, incorporando de forma inherente tanto el cambio en la probabilidad de tener un cero como el cambio en la magnitud de los valores positivos.

Este modelo, además de modelar en forma exacta los ceros y los valores continuos, tiene su atractivo en ejemplos como estimar el total de visitas al médico o la ingesta total de algún alimento ya que en este ultimo caso la variable Y representa la ingesta total en un período dado, N el número de eventos en los cuales se incurrió en la ingesta y X_i el valor de la ingesta para el i-ésimo evento.

4. Estimación e implementación

La estimación de los parámetros se realiza mediante el método de máxima verosimilitud (Zhang, 2013). Dado que el índice p no es conocido *a priori*, los algoritmos modernos (como `tweedie.profile` de la librería `tweedie` en R) (Dunn, 2022) evalúan la verosimilitud a lo largo de un rango de valores para $p \in (1,2)$, para así hallar el ajuste óptimo que minimice la desviación del modelo.

5. Consideraciones finales

El modelo lineal generalizado Tweedie constituye una solución parsimoniosa y matemáticamente rigurosa para la modelación de variables semicontinuas con una alta presencia de ceros. Su principal ventaja reside en la capacidad de generar predicciones a nivel poblacional y con ello ayudar en la planificación presupuestaria o la evaluación del impacto económico de alguna política alimentaria, evitando las complicaciones algebraicas que tiene considerar una mixtura de modelos fragmentados. Sin embargo, si el interés es desentrañar los determinantes de la incidencia (cero vs. positivo) separadamente de los determinantes de la severidad (magnitud dado positivo), un modelo en dos partes sigue siendo más informativo para realizar inferencias.

Cuadro 1: Comparación de estrategias para datos semicontinuos.

Modelo	Manejo de ceros	Supuesto de varianza	Aplicación principal
Tobit	Censura latente	Homocedasticidad	Datos truncados
Dos partes (*Hurdle*)	Modelo binario separado	Flexible (2 etapas)	Procesos independientes
Tweedie GLM	**Integrado (Poisson)**	**Ley de potencia μ^p**	**Predicción unificada**

Referencias

Dunn, P. K. (2022). *Tweedie: Evaluation of Tweedie Exponential Family Models* [R package version 2.3.5].

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications in statistical analysis. *Shanghai Archives of Psychiatry*, *26*(2), 105-109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Feng, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distributions and Applications*, *8*, 8. https://doi.org/10.1186/s40488-021-00121-4

Jørgensen, B. (1987). Exponential Dispersion Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *49*(2), 127-145. https://doi.org/10.1111/j.2517-6161.1987.tb01685.x

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models* (2nd). Chapman; Hall/CRC. https://doi.org/10.1007/978-1-4899-3242-6

Min, Y., & Agresti, A. (2022). Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, *1*(1-2), 7-33.

Zhang, Y. (2013). Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. *Stat Comput*, *23*, 743-757. https://doi.org/10.1007/s11222-012-9343-7