Implementación de la distribución Burr Hatke para evaluar el proceso de estimación de parámetros en datos sobre dispersos *

Valentina Hurtado Sepúlveda *Universidad Nacional de Colombia - sede Medellín*
Freddy Hernández Barajas *Universidad Nacional de Colombia - sede Medellín*

Resumen: En áreas del conocimiento como la medicina o ingeniería, existen conjuntos de datos que no se pueden modelar con distribuciones tradicionales. En modelación de datos discretos, lo usual es emplear la distribución Poisson, sin embargo, al tener varianza igual a su media (equidispersión), podría representar un limitante a la hora de analizarlos. Se estudia la distribución Burr Hatke (DBH), con varianza mayor a su media, propiedad útil para modelar datos excesivamente dispersos. En el presente estudio, se implementó Burr-Hatke (DBH) en el paquete DiscreteDists de R y se llevó a cabo un estudio de simulación para evaluar el proceso estimación de parámetros. Se realizó una aplicación con datos reales, donde se ponen a competir, resultando que el mejor ajuste fue el de la distribución DBH frente a la PO. Es así como su implementación resultaría de gran importancia para llevar a cabo posteriores estudios académicos.

Keywords: Distribución Burr-Hatke, Simulación estadística, Estimación de parámetros, Lenguaje de programación R

1. Introducción

Los datos se manifiestan de múltiples formas, y la estadística es el área encargada de crear y fomentar nuevas herramientas para el tratamiento e interpretación de datos. Desde la academia es importante estar al tanto de las nuevas distribuciones que van surgiendo. El tratamiento de datos discretos típicamente se ha modelado con la distribución Poisson (PO), sin embargo, al ser equidispersa, resulta más conveniente hacer uso de nuevas distribuciones alternas. La distribución Burr-Hatke (DBH) es adecuada para modelar datos sobredispersos, sin embargo, para fines prácticos no era posible usarla, es por esto que en este trabajo se implementó dicha distribución en el paquete DiscreteDists de R con el fin de tenerla a disposición en posteriores investigaciones. Adicionalmente se realizó un estudio de simulación para evaluar el proceso de estimación de parámetros y un pequeño ejemplo con datos reales donde se usó los modelos gamlss por su gran flexibilidad. Este documento está estructurado de la siguiente manera: en la sección 2, revisión de la distribución discreta Burr-Hatke y su implementación en R a través del paquete DiscreteDists. La sección 3 presenta el estudio de simulación, en la sección 4 los resultados, en la sección 5, se ilustra la aplicación de la distribución DBH vs PO con el modelo gamlss en un contexto específico, con el fin de destacar su desempeño en situaciones prácticas.

2. Metodología e implementación

- Crear funciones para implementar la distribución bajo estudio (DBH) en R

*Autor de contacto: vhurtados@unal.edu.co.

- Analizar el proceso de estimación de parámetros a través de dos modelos: con covariables y sin covariables.
- Comparar dos modelos gamlss con datos reales.

Distribución Burr-Hatke

La DBH fue introducida por Maniu & Voda, 2008. Tiene un parámetro μ, y su función de masa de probabilidad (f.m.p) y función de densidad acumulada (f.d.a) está dada por

$$f(x;\mu) = \left(\frac{1}{x+1} - \frac{\mu}{x+2}\right)\mu^x$$

$$F(x;\mu) = 1 - \frac{\mu^{x+1}}{x+2}$$

donde, $0 < \mu < 1$ y $x \in \mathbb{N}_0$

Implementación

Para llevar a cabo la implementación de la DBH se crearon las siguienets funciones: función de densidad de probabilidad, función de densidad acumulada, función quantil y la generadora de números aleatorios. Se pueden encontar de la siguiente manera:

```
dDBH()  # Densidad
pDBH()  # P(X<=x)
qDBH()  # Cuantil
rDBH()  # Generador de números aleatorios
 DBH()  # Familia
```

Ejemplo

```
library(DiscreteDists)  # To use the DBH family
library(gamlss)    # To use gamlss

set.seed(190)
y <- rDBH(n=1000, mu=0.74)

mod <- gamlss(y~1, family=DBH,
              control=gamlss.control(n.cyc=500, trace=FALSE)) # To fit the model

inv_logit <- function(x) exp(x) / (1+exp(x)) # using the inverse link function
inv_logit(coef(mod, what='mu')) # Extracting the fitted values for mu
```

```
## (Intercept)
##   0.7429384
```

3. Estudio de Simulación

Se llevó a cabo un estudio de simulación Monte Carlo donde se consideraron dos casos: "con covariables" y "sin covariables". Las medidas para evaluar el rendimiento del procedimiento de estimación de un parámetro θ, fueron el valor medio y el error cuadrático medio (MSE) del estimador $\hat{\theta}$, definidos respectivamente

$$Mean = \frac{\sum_{i=1}^{i=k} \hat{\theta}_i}{k}$$

$$MSE = \frac{\sum_{i=1}^{i=k} (\hat{\theta}_i - \theta)^2}{k}$$

donde k representa el núero de estimaciones para θ.

La metodología para los casos analizados en el estudio de simulación es la siguiente.

- Caso 1: Simulación sin covariables

  En este caso, se analizó el comportamiento asintótico del parámetro estimado $\hat{\mu}$ simulando datos de una distribución $DBH(\mu = 0.416)$ considerando diferentes tamaños de muestra $n = 10, 20, ..., 300$.

$$y_i \sim DBH(\mu)$$
$$\mu = 0.416$$

- Caso 2: Simulación con covariables

  En el segundo caso, se analizó el comportamiento asintótico de los coeficientes estimados β_0 y β_1 del modelo de regresión DBH, considerando diferentes tamaños de muestra $n = 10, 20, ..., 300$. Los datos se generan del siguiente modelo:

$$y_i \sim DBH(\mu_i)$$
$$logit(\mu_i) = \beta_0 + \beta_1 \times X_{1i}$$
$$X_1 \sim U(0, 1)$$

  El vector de parámetros en este caso fue $\theta = (\beta_0 = 1.064, \beta_1 = -0.7)^\top$. Del modelo (4) se tiene que $X_1 \sim U(1, 0)$, lo cual implica que $E(X_1) = 0.5$, entonces $\mu = \exp(0.335 - 0.012(0.5))/(1 + \exp(0.335 - 0.012(0.5)) = 0.416$.

4. Resultados

Se presenta la media y el MSE con el fin de explorar la evolución del parámetro estimado. La figura 1 muestra los reultados del modelo sin covariables, y las figuras 2 y 3 las del modelo con covariables:

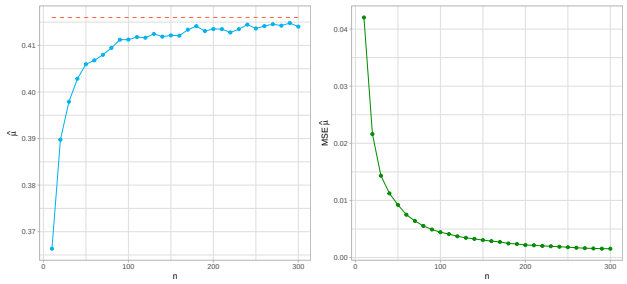

Figure 1: A la izquierda, el promedio de los parámetros estimados $\hat{\mu}$ versus n. La línea roja corresponde al valor real $\mu = 0.416$. A la derecha, el error cuadrático medio para los parámetros estimados $\hat{\mu}$ versus n.

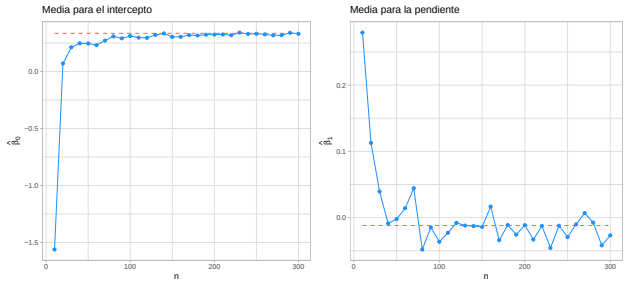

Figure 2: Promedio de los parámetros estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ versus n. La línea roja corresponde al valor objetivo.

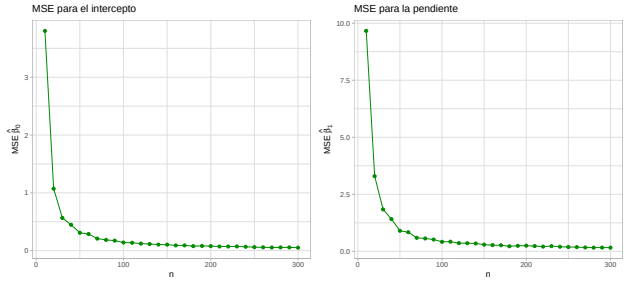

Figure 3: Error cuadrático medio para la estimación de parámetros β_0 y β_1 versus n.

5. Aplicación

En esta sección se presentan resultados de la implementación de modelos GAMLSS con la base de datos "Doctor'' del paquete `Ecdat` para explicar el número de visitas al médico (y) usando como covariables children (número de niños en el hogar) y health (estado de salud) para dos modelos: DBH y PO. Se utiliza como medida de comparación el criterio de información Akaike (AIC) (Akaike, 1973), para medir la calidad de los modelos y además se analizan los residuales.

La Tabla 2 muestra valores AIC para cada modelo, se prefiere el modelo con el menor AIC.

Table 1: Modelos DBH vs. PO.

Dist	Modelo	AIC
DBH	`gamlss(y∼children+health,family=DBH())`	1616.209
PO	`gamlss(y∼children+health,family=PO())`	2213.905

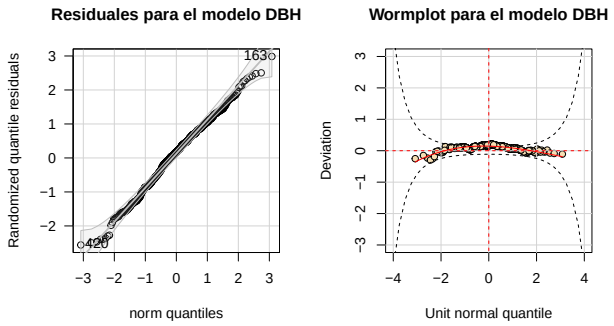

Figure 4: Estimaciones para DBH

Table 2: Modelos DBH vs. PO.

Model	Estimate	Std.Error	t value	P-value
Intercept	2.7256	0.3855	7.070	5.45e-12
Children	-0.3313	0.1214	-2.729	0.00659
Health	0.6612	0.1999	3.308	0.00101

6. Conclusiones

Se encontró que la estimación de parámetros tiende a ser más precisa a medida que incrementa el tamaño de muestra n. Al comparar los modelos gamlss implementando la DBH frente a la PO, resultó ser el modelo de la DBH el que mejor se ajustó a los datos. Es así como la implementación de esta nueva distribución resulta ser una gran alternativa y demostró su capacidad para modelar sobredispersión.

7. Referencias

1. **R: A Language and Environment for Statistical Computing**

   - *Autores:* R Core Team
   - *Organización:* R Foundation for Statistical Computing
   - *Dirección:* Vienna, Austria
   - *Año:* 2019
   - *URL:* https://www.R-project.org/

2. **Discrete Burr-Hatke Distribution With Properties, Estimation Methods and Regression Model**

   - *Autores:* El-Morshedy, M., Eliwa, M. S., Altun, Emrah
   - *Revista:* IEEE Access
   - *Año:* 2020
   - *Volumen:* 8
   - *Número:*
   - *Páginas:* 74359-74370
   - *DOI:* 10.1109/ACCESS.2020.2988431

3. **Generalized Additive Models for Location, Scale and Shape**

   - *Autores:* Rigby, R.A., Stasinopoulos, D.M.
   - *Revista:* Journal of the Royal Statistical Society: Series C (Applied Statistics)
   - *Volumen:* 54
   - *Páginas:* 507–554
   - *Año:* 2005