MODELADO DE CUANTILES MEDIANTE EL MODELO DE REGRESIÓN LINEAL LOG-SKEW-NORMAL

Anlly Daniela Giraldo Melo

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2021

Modelado de cuantiles mediante el modelo de regresión lineal log-skew-normal

Anlly Daniela Giraldo Melo

Trabajo de grado presentado como requisito parcial para optar al título de Matemática

Orientador:
Raúl Alejandro Morán Vásquez

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2021

Índice general

Introducción

El análisis de regresión es una de las técnicas estadísticas más populares para modelar la relación entre una variable respuesta y un conjunto de variables explicativas. Esta metodología ha sido ampliamente aplicada en diversas áreas del conocimiento, como la economía, ingeniería, física, química, biología, ciencias sociales y de la salud, entre otras. El modelo de regresión lineal normal (MRLN) es comúnmente utilizado para describir la relación entre una variable respuesta y un conjunto de variables explicativas. Sin embargo, cuando nos acercamos al estudio de este tema descubrimos que las suposiciones del MLRN rara vez se cumplen, debido a que, por ejemplo, es muy común que las observaciones asociadas a variables de interés provengan de distribuciones asimétricas o de colas pesadas.

Los primeros enfoques para enfrentar este problema consistieron en aplicar algún tipo de trasformación a los datos con el fin de alcanzar la normalidad, luego emplear los procedimientos habituales y las propiedades bien conocidas de la distribución normal. Sin embargo, este enfoque tiene la desventaja de que los parámetros del modelo solo son interpretables en términos de las variables transformadas y no en términos de las variables originales. Por esta razón, desde finales del siglo XIX se inició un creciente interés en la búsqueda de distribuciones no normales que permitieran un modelamiento más flexible y que facilitaran la interpretación de los resultados. Por ejemplo, Galton usó en 1879 la transformación logarítmica dando inicio al estudio de la distribución log-normal, aunque fue Edgeworth [21] quien formalizó este concepto alrededor del año 1898. Durante el siglo XX surgieron muchas propuestas para el modelamiento flexible donde se buscaban algunas propiedades deseables para los modelos como por ejemplo, obtener parámetros interpretables e incluir parámetros de asimetría y de colas pesadas.

Una familia de distribuciones flexibles que ha tenido un importante rol en los últimos años es la distribución skew-normal. Las primeras ideas sobre esta distribución se remontan a comienzos del siglo XX cuando Fernando de Helguero publicó en 1909 dos artículos con una formulación para construir distribuciones no normales bajo un

enfoque diferente al que había sido utilizado por Edgeworth y Pearson, el cual se basa en un mecanismo de selección [8]. De esta manera, el trabajo de Helguero fue un precursor de la idea actual de la distribución skew-normal, muy utilizada para modelar datos asimétricos. La familia de distribuciones skew-normal se dio a conocer por Azzalini mediante un artículo publicado en 1985 [3].

La distribución skew-normal es una extensión de la distribución normal con la ventaja de que involucra un parámetro adicional que controla la asimetría de la distribución, permitiendo tener en cuenta la asimetría de los datos. Estudios sobre la distribución skew-normal aparecen en Azzalini [9], Genton [17] y Arellano-Valle y Azzalini [2]. La importancia de esta distribución puede evidenciarse en su flexibilidad para modelar de manera conjunta varias características poblacionales, lo cual es llamativo en el modelado estadístico de datos. Existe una extensa literatura sobre estudios relacionados con la distribución skew-normal, en particular estudios sobre extensiones multivariadas de esta distribución se pueden encontrar en Azzalini y Dalla Valle [7]. Además, Branco y Dey [11] proponen las distribuciones skew-t, para el caso univariado y multivariado, extendiendo las respectivas distribuciones skew-normal. La distribución skew-t es útil en problemas de regresión cuando la distribución de errores exhibe asimetría y colas pesadas. Adicionalmente, se ha demostrado que la distribución skew-normal es eficaz en muchas aplicaciones que incluyen series de tiempo [18], análisis Bayesiano [24], estadística espacial [1], modelos gráficos [12], entre muchas otras.

Si bien la distribución skew-normal es adecuada para modelar datos asimétricos, esta distribución no considera la naturaleza positiva de los datos que pueden ocurrir con frecuencia en diversos fenómenos, lo mismo ocurre con la distribución normal que tiene soporte en todos los reales y tiene el supuesto de simetría. Una metodología alternativa para modelar datos positivos es considerar la transformación de Box-Cox [10] de la variable respuesta, aunque este enfoque tiene la desventaja de que los parámetros solo son interpretables en términos de la variable trasformada. Otro enfoque consiste en considerar la distribución log-normal, que tiene soporte en $\mathbb{R}_+$ y que además permite modelar asimetría. Una extensión de la distribución log-normal se presenta a través de la distribución log-skew-normal [23, 25], la cual tiene soporte en $\mathbb{R}_+$ e incluye un parámetro adicional que permite controlar la asimetría de la distribución.

En este trabajo derivamos algunas propiedades de la distribución log-skew-normal que permiten la interpretación de algunos de sus parámetros con cuantiles de la variable respuesta extendiendo algunas técnicas propuestas por Morán-Vásquez et al. [26] en el ámbito univariado. Además, proponemos y estudiamos el modelo de regresión

lineal log-skew-normal, el cual permite estudiar la manera en que los cuantiles de la variable respuesta son afectados por un conjunto de variables explicativas, teniendo en cuenta la posible asimetría de la respuesta. Este modelo es una alternativa al modelo de regresión cuantílica univariada [20]. La utilidad del modelo propuesto para modelar cuantiles para datos positivos, posiblemente asimétricos, se ilustra mediante un análisis a datos reales sobre nacidos vivos en el Hospital Manuel Uribe Ángel del municipio de Envigado, Antioquia, Colombia.

Capítulo 1

Preliminares

En este capítulo estudiamos la familia de distribuciones skew-normal y algunas de sus propiedades. Después, presentamos la distribución log-skew-normal, que es la base del modelo de regresión que desarrollamos en el Capítulo 2.

1.1. Distribución skew-normal

La función de densidad de probabilidad (FDP) y la función de distribución acumulada (FDA) de una variable aleatoria con distribución normal estándar, denotadas por φ y Φ, respectivamente, son dadas por

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \text{ y } \Phi(z) = \int_{-\infty}^{z} \varphi(u)\mathrm{d}u, \quad z \in \mathbb{R}.$$

Es posible mostrar que $\varphi(z) = \varphi(-z)$ y $\Phi(z) + \Phi(-z) = 1$, las cuales son propiedades de simetría de la distribución normal estándar.

En la Proposición 1.1.1 presentamos un esquema para generar un conjunto de distribuciones. La familia de distribuciones skew-normal se puede obtener como un caso particular del esquema presentado, teniendo como base la FDP y la FDA de la distribución normal estándar.

Proposición 1.1.1. *Sea f_0 una FDP sobre $\mathbb{R}$, $G_0(\cdot)$ una FDA continua en los reales, y $w(\cdot)$ una función real-valuada, tal que*

$$f_0(-z) = f_0(z), \quad w(-z) = -w(z), \quad G_0(-z) = 1 - G_0(z), \tag{1.1}$$

para todo $z \in \mathbb{R}$. Entonces

$$f(z) = 2f_0(z)G_0\{w(z)\} \tag{1.2}$$

es una FDP en $\mathbb{R}$.

Demostración. Consideremos la función $g(z)$ definida por $g(z) = 2\left[G_0\{w(z)\} - \frac{1}{2}\right] f_0(z)$, $z \in \mathbb{R}$. Por las condiciones dadas en (1.1) tenemos que

$$\begin{aligned} g(-z) &= 2\left[G_0\{w(-z)\} - \frac{1}{2}\right] f_0(-z) \\ &= 2\left[G_0\{-w(z)\} - \frac{1}{2}\right] f_0(z) \\ &= 2\left[1 - G_0\{w(z)\} - \frac{1}{2}\right] f_0(z) \\ &= -2\left[G_0\{w(z)\} - \frac{1}{2}\right] f_0(z) \\ &= -g(z). \end{aligned}$$

Por lo tanto, $g(z)$ es una función impar. Además, como f_0 es integrable y G_0 es una FDA, entonces, $|2G_0\{w(z)\} - 1| \leq 1$ luego, $|g(z)| = |2G_0\{w(z)\} - 1|\ |f_0(z)| \leq |f_0(z)| = f_0(z)$, es decir, $g(z)$ es integrable. De lo anterior se sigue que

$$\begin{aligned} 0 &= \int_{\mathbb{R}} g(z)\mathrm{d}z \\ &= \int_{\mathbb{R}} 2G_0\{w(z)\}f_0(z)\mathrm{d}z - \int_{\mathbb{R}} f_0(z)\mathrm{d}z \\ &= \int_{\mathbb{R}} 2G_0\{w(z)\}f_0(z)\mathrm{d}z - 1. \end{aligned}$$

En consecuencia, $\int_{\mathbb{R}} 2G_0\{w(z)\}f_0(z)\mathrm{d}z = \int_{\mathbb{R}} f(z)\mathrm{d}z = 1$. Lo anterior prueba que $f(z)$ es una FDP con soporte en $\mathbb{R}$.

Si seleccionamos $f_0 = \varphi$, $G_0 = \Phi$ y $w(z) = \lambda z$, para cualquier constante $\lambda \in \mathbb{R}$, las cuales cumplen las condiciones dadas en (1.2), obtenemos la FDP

$$\varphi(z;\lambda) = 2\varphi(z)\Phi(\lambda z), \quad z \in \mathbb{R}. \tag{1.3}$$

Definición 1.1.1. *Decimos que una variable aleatoria $Z \in \mathbb{R}$ tiene distribución skew-normal estándar con parámetro de asimetría λ, denotada por $Z \sim \mathrm{SN}(\lambda)$, si su FDP es dada por* (1.3).

En la Figura 1.1 ilustramos diferentes formas de $\varphi(z;\lambda)$ para varios valores de λ. Observamos que el parámetro λ controla la magnitud y dirección de la asimetría de la distribución. A medida que $|\lambda|$ se hace más grande, la asimetría de la FDP es más pronunciada. Si $\lambda < 0$, la FDP muestra una asimetría a izquierda. Si $\lambda > 0$, la asimetría es a derecha. Cuando $\lambda = 0$, la FDP es simétrica y coincide con la FDP de la distribución

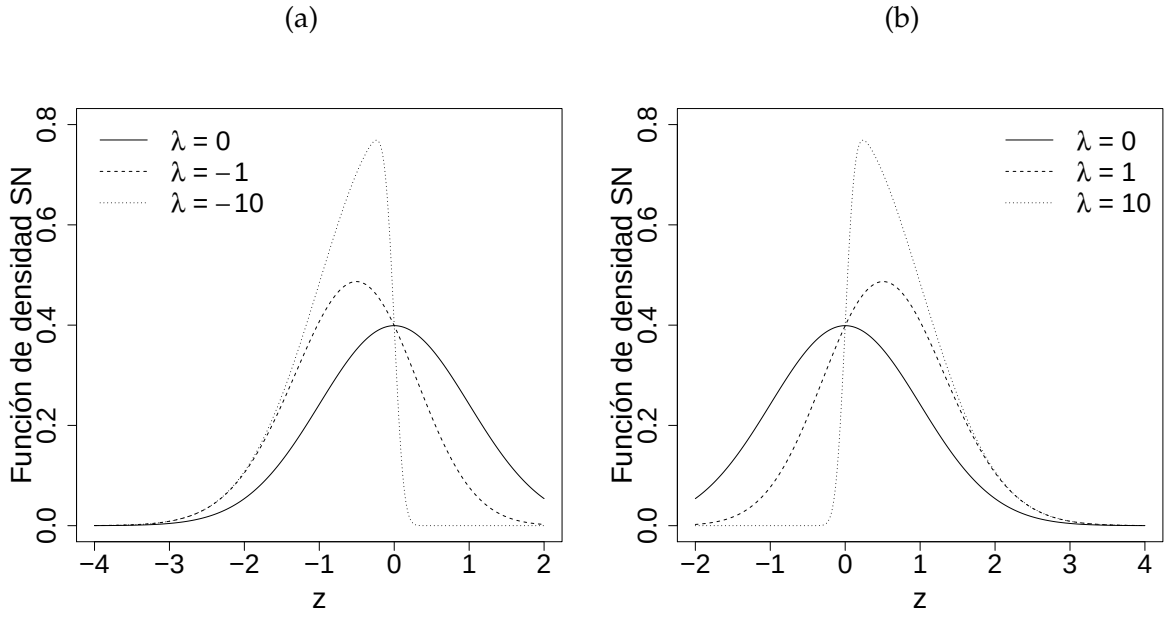

Figura 1.1: FDP de la distribución skew-normal estándar para (a) $\lambda = 0, -1, -10$, (b) $\lambda = 0, 1, 10$

normal estándar. Lo anterior indica que la distribución normal es un caso particular de la distribución skew-normal.

En el Teorema 1.1.1 introducimos parámetros de localización y de escala para la distribución skew-normal.

Teorema 1.1.1. *Sean ξ, ω números reales con $\omega > 0$. Si $X = \xi + \omega Z$, donde $Z \sim \text{SN}(\lambda)$, entonces la FDP de X es dada por*

$$\varphi(x; \xi, \omega^2, \lambda) = \frac{2}{\omega}\varphi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\lambda\frac{x-\xi}{\omega}\right), \quad x \in \mathbb{R}. \tag{1.4}$$

Demostración. Consideremos la transformación $Z = (X - \xi)/\omega$, donde $Z \sim \text{SN}(\lambda)$. El Jacobiano de la transformación es dado por $J(z \to x) = 1/\omega$. Por lo tanto, a partir de (1.3), tenemos que la FDP de X es dada por

$$\begin{aligned}\varphi(x; \xi, \omega^2, \lambda) &= \frac{1}{\omega}\varphi((x-\xi)/\omega; \lambda) \\ &= \frac{2}{\omega}\varphi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\lambda\,\frac{x-\xi}{\omega}\right).\end{aligned}$$

En el Teorema 1.1.1, el parámetro ξ es un parámetro de localización y el parámetro ω es un parámetro de escala, los cuales determinan una familia de localización-escala

para la distribución de X cuando λ es fijo.

Definición 1.1.2. *Decimos que la variable aleatoria $X \in \mathbb{R}$ tiene distribución skew-normal con parámetro de localización $\xi \in \mathbb{R}$, parámetro de escala $\omega > 0$ y parámetro de asimetría $\lambda \in \mathbb{R}$ si su FDP es dada por (1.4). Escribimos $X \sim \text{SN}(\xi, \omega^2, \lambda)$.*

Como consecuencia de la Definición 1.1.2, cuando $\xi = 0$ y $\omega = 1$ obtenemos la FDP (1.3). Recíprocamente, si $X \sim \text{SN}(\xi, \omega^2, \lambda)$, entonces $Z = (X - \xi)/\omega \sim \text{SN}(\lambda)$.

En el lema 1.1.1 presentamos un resultado acerca de la distribución normal estándar que nos permitirá obtener la función generadora de momentos (FGM) para una variable aleatoria con distribución skew-normal.

Lema 1.1.1. *Si $U \sim \text{N}(0, 1)$ entonces*

$$\text{E}(\Phi(hU + k)) = \Phi\left(\frac{k}{\sqrt{1 + h^2}}\right), \quad h, k \in \mathbb{R}. \tag{1.5}$$

Demostración. Sea $U \sim \text{N}(0, 1)$ y $h, k \in \mathbb{R}$. Note que

$$\begin{aligned} \Phi(hU + k) &= P(Z_0 \leq hU + k) \\ &= P(Z_0 - hU \leq k), \end{aligned}$$

donde $Z_0 \sim \text{N}(0, 1)$. Debido a que $Z_0 - hU \sim \text{N}(0, 1 + h^2)$, entonces

$$\begin{aligned} P(Z_0 - hU \leq k) =& P\left(\frac{Z_0 - hU}{\sqrt{1 + h^2}} \leq \frac{k}{\sqrt{1 + h^2}}\right) \\ =& \Phi\left(\frac{k}{\sqrt{1 + h^2}}\right). \end{aligned}$$

De esta manera,

$$\begin{aligned} \text{E}(\Phi(hU + k)) =& \text{E}\left(\Phi\left(\frac{k}{\sqrt{1 + h^2}}\right)\right) \\ =& \Phi\left(\frac{k}{\sqrt{1 + h^2}}\right), \end{aligned}$$

donde la última igualdad se obtiene al calcular el valor esperado de una constante.

En el lema 1.1.2 mostramos la FGM de una variable aleatoria distribuida skew-normal.

Lema 1.1.2. *Si* $X \sim \text{SN}(\xi, \omega^2, \lambda)$, entonces la FGM de X es dada por

$$M_X(t) = 2\exp(\xi t + \omega^2 t^2/2)\Phi\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\,\omega t\right) \tag{1.6}$$

Demostración. Usando la definición de una FGM y la Definición 1.1.2, tenemos que

$$\begin{aligned} M_X(t) &= \text{E}(\exp(tX)) \\ &= \text{E}(\exp(t(\xi + \omega Z))), \end{aligned}$$

donde $Z \sim \text{SN}(\lambda)$. Calculando el valor esperado anterior obtenemos que

$$\begin{aligned} M_X(t) &= \exp(\xi t)\text{E}(\exp(\omega t Z)) \\ &= \exp(\xi t)\int_{\mathbb{R}} \exp(\omega z t) 2\varphi(z)\Phi(\lambda z)\text{d}z \\ &= 2\exp(\xi t)\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp(\omega z t)\exp(-z^2/2)\Phi(\lambda z)\text{d}z \\ &= 2\exp(\xi t)\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp(\omega z t - z^2/2 - \omega^2 t^2/2 + \omega^2 t^2/2)\Phi(\lambda z)\text{d}z \\ &= 2\exp(\xi t + \omega^2 t^2/2)\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp(-(z-\omega t)^2/2)\Phi(\lambda z)\text{d}z \\ &= 2\exp(\xi t + \omega^2 t^2/2)\int_{\mathbb{R}} \varphi(z-\omega t)\Phi(\lambda z)\text{d}z. \end{aligned}$$

Haciendo el cambio de variable $u = z - \omega t$ tenemos que

$$\begin{aligned} M_X(t) &= 2\exp(\xi t + \omega^2 t^2/2)\int_{\mathbb{R}} \varphi(u)\Phi(\lambda u + \lambda\omega t))\text{d}u \\ &= 2\exp(\xi t + \omega^2 t^2/2)\text{E}(\Phi(\lambda U + \lambda\omega t)), \end{aligned}$$

donde $U \sim \text{N}(0,1)$. Finalmente, usando el Lema 1.1.1 obtenemos el resultado.

En la Proposición 1.1.2 mostramos que transformaciones afines de una variable aleatoria con distribución skew-normal también presentan distribución skew-normal. Recordemos que para cualquier número real x la función signo, denotada por $\text{sgn}(x)$, se define por

$$\text{sgn}(x) = \begin{cases} 1 & \text{si } x > 0, \\ 0 & \text{si } x = 0, \\ -1 & \text{si } x < 0. \end{cases}$$

Proposición 1.1.2. *Sean a y b números reales con $a \neq 0$. Si $X \sim \text{SN}(\xi, \omega^2, \lambda)$*, entonces $aX + b \sim \text{SN}(a\xi + b, a^2\omega^2, \text{sgn}(a)\,\lambda)$.

Demostración. Sea $Y = aX + b$, entonces la FGM de Y es dada por

$$\begin{aligned} M_Y(t) &= \text{E}(\exp(tY)) \\ &= \text{E}(\exp(t(aX+b))) \\ &= \exp(bt)\text{E}(\exp(atX)) \\ &= \exp(bt)M_X(at). \end{aligned} \tag{1.7}$$

Note que por definición $\omega > 0$, entonces $\sqrt{\omega^2} = \omega$. Por lo tanto, la FGM (1.6) la podemos escribir como

$$M_X(t) = 2\exp(\xi t + \omega^2 t^2/2)\Phi\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\sqrt{\omega^2}t\right). \tag{1.8}$$

A partir de (1.7) y (1.8) obtenemos que

$$M_Y(t) = 2\exp((a\xi + b)t + (a\omega)^2 t^2/2)\Phi\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\sqrt{\omega^2}at\right).$$

Ahora, si $a > 0$, entonces

$$\begin{aligned} M_Y(t) &= 2\exp((a\xi + b)t + (a\omega)^2 t^2/2)\Phi\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\sqrt{a^2\omega^2}t\right) \\ &= 2\exp((a\xi + b)t + (a\omega)^2 t^2/2)\Phi\left(\frac{\text{sgn}(a)\lambda}{\sqrt{1+(\text{sgn}(a)\lambda)^2}}\sqrt{a^2\omega^2}t\right), \end{aligned}$$

para $a < 0$ tenemos que

$$\begin{aligned} M_Y(t) &= 2\exp((a\xi + b)t + (a\omega)^2 t^2/2)\Phi\left(\frac{-\lambda}{\sqrt{1+\lambda^2}}\sqrt{a^2\omega^2}t\right) \\ &= 2\exp((a\xi + b)t + (a\omega)^2 t^2/2)\Phi\left(\frac{\text{sgn}(a)\lambda}{\sqrt{1+(\text{sgn}(a)\lambda)^2}}\sqrt{a^2\omega^2}t\right). \end{aligned}$$

En ambos casos concluimos que $Y \sim \text{SN}(a\xi + b, a^2\omega^2, \text{sgn}(a)\,\lambda)$.

1.2. Algunas propiedades de la distribución skew-normal

En esta sección presentamos algunas propiedades básicas de la distribución skew-normal.

Proposición 1.2.1. *Sea $Z \sim$SN(λ) con FDP $\varphi(z;\lambda)$. Entonces*

1. $\varphi(z;0) = \varphi(z)$, *para todo* $z \in \mathbb{R}$;
2. $\varphi(0;\lambda) = \varphi(0)$, *para todo* $\lambda \in \mathbb{R}$;
3. $-Z \sim$SN($-\lambda$), *equivalentemente* $\varphi(-z;\lambda) = \varphi(z;-\lambda)$, *para todo* $z \in \mathbb{R}$;
4. *Si* $Z_0 \sim \mathrm{N}(0,1)$, *entonces* $|Z_0|$ *y* $|Z|$ *se distribuyen idénticamente;*
5. $Z^2 \sim \chi^2_1$, *para todo* λ.

Demostración.

1. $\varphi(z;0) = 2\varphi(z)\Phi(0) = 2\varphi(z)/2 = \varphi(z)$, para todo $z \in \mathbb{R}$.
2. $\varphi(0;\lambda) = 2\varphi(0)\Phi(0\times\lambda) = 2\varphi(0)/2 = \varphi(0) = 1/\sqrt{2\pi}$, para todo $\lambda \in \mathbb{R}$.
3. Debido a la simetría con respecto al cero de la distribución normal estándar, tenemos que $\varphi(z) = \varphi(-z)$, para todo z. Por lo tanto, $\varphi(-z;\lambda) = 2\varphi(-z)\Phi(\lambda(-z)) = 2\varphi(z)\Phi((-\lambda)z) = \varphi(z;-\lambda)$, así, $-Z \sim$SN($-\lambda$) para todo $z \in \mathbb{R}$.
4. Sea $Z_0 \sim \mathrm{N}(0,1)$, entonces para $z_0 > 0$

$$\begin{aligned} P(|Z_0| \le z_0) &= P(-z_0 \le Z_0 \le z_0) \\ &= \int_{-z_0}^{z_0} \varphi(u)\mathrm{d}u \\ &= \int_0^{z_0} 2\varphi(u)\mathrm{d}u, \end{aligned}$$

   donde la última igualdad se cumple por la simetría de la función φ con respecto al cero.

   Por otro lado, para cualquier $z > 0$ tenemos que

$$\begin{aligned} P(|Z| \le z) &= \int_{-z}^{z} 2\varphi(u)\Phi(\lambda u)\mathrm{d}u \\ &= \int_{-z}^{0} 2\varphi(u)\Phi(\lambda u)\mathrm{d}u + \int_0^z 2\varphi(u)\Phi(\lambda u)\mathrm{d}u. \end{aligned}$$

Haciendo el cambio de variable $t = -u$ en la primera integral tenemos que

$$\begin{aligned} P(|Z| \leq z) &= -\int_z^0 2\varphi(-t)\Phi(-\lambda t)\mathrm{d}t + \int_0^z 2\varphi(u)\Phi(\lambda u)\mathrm{d}u \\ &= \int_0^z 2\varphi(t)(1-\Phi(\lambda t))\mathrm{d}t + \int_0^z 2\varphi(u)\Phi(\lambda u)\mathrm{d}u \\ &= \int_0^z 2\varphi(t)\mathrm{d}t - \int_0^z 2\varphi(t)\Phi(\lambda t)\mathrm{d}t + \int_0^z 2\varphi(u)\Phi(\lambda u)\mathrm{d}u \\ &= \int_0^z 2\varphi(t)\mathrm{d}t \\ &= P(|Z_0| \leq z). \end{aligned}$$

Lo anterior muestra que las variables aleatorias $|Z_0|$ y $|Z|$ son idénticamente distribuidas.

5. A partir de la propiedad 4 tenemos que

$$P(Z^2 \leq z) = P(|Z| \leq \sqrt{z}) = P(|Z_0| \leq \sqrt{z}) = P(Z_0^2 \leq z),$$

lo cual muestra que Z^2 y Z_0^2 son idénticamente distribuidas. Sabemos que $Z_0^2 \sim \chi_1^2$, esto implica que $Z^2 \sim \chi_1^2$.

En la Proposición 1.2.1, la propiedad 1 muestra que la distribución normal estándar es un caso particular de la distribución skew-normal, si además consideramos la variable aleatoria $X \sim \mathrm{SN}(\xi, \omega^2, 0)$, un cálculo directo en (1.4) muestra que $\varphi(x; \xi, \omega^2, 0) = \exp(-(x-\xi)^2/2\omega^2)/\omega\sqrt{2\pi}$, así $X \sim \mathrm{N}(\xi, \omega^2)$. Si graficamos la FDP de $Z \sim$SN(λ), la propiedad 2 indica que $\varphi(0)$ es un punto de intersección de las gráficas para todos los valores de λ, este hecho lo podemos observar en la Figura 1.1 donde graficamos la FDP de la distribución skew-normal estándar para $\lambda = 0, -1, -10, 1, 10$. La propiedad 3 es una propiedad de reflexión, señala que cuando el signo de λ cambia, la FDP se refleja en el lado opuesto del eje $z = 0$, esto también se ilustra en la Figura 1.1; además, esta propiedad es un caso particular de la Proposición 1.1.2 tomando $a = -1$, $b = 0$, $\xi = 0$ y $\omega = 1$. La propiedad 5 presenta una importante relación entre la distribución skew-normal estándar y la chi-cuadrado ya que la distribución chi-cuadrado es usada frecuentemente en inferencia estadística.

1.3. Distribución log-skew-normal

En esta sección presentamos la definición de la distribución log-skew-normal (LSN) [23]. Posteriormente, mostramos algunos resultados que establecen la manera en que algunos parámetros se relacionan con cuantiles de la variable aleatoria. Estas relaciones nos permitirán interpretar los parámetros de la distribución LSN como características de la variable de interés, hecho que es muy importante en el modelado estadístico.

Definición 1.3.1. *Decimos que la variable aleatoria positiva Y tiene distribución log-skew-normal con parámetro de escala $\xi > 0$, parámetro de dispersión relativa $\omega > 0$ y parámetro de asimetría $\lambda \in \mathbb{R}$, si* $\log(Y) \sim \text{SN}(\log(\xi), \omega^2, \lambda)$. *Escribimos* $Y \sim \text{LSN}(\xi, \omega^2, \lambda)$.

En el Teorema 1.3.1 presentamos la FDP de una variable aleatoria con distribución LSN.

Teorema 1.3.1. *La FDP de* $Y \sim \text{LSN}(\xi, \omega^2, \lambda)$ *está dada por*

$$\frac{2}{y\,\omega}\varphi\left(\frac{\log(y/\xi)}{\omega}\right)\Phi\left(\lambda\,\frac{\log(y/\xi)}{\omega}\right), \quad y > 0. \tag{1.9}$$

Demostración. Usando la Definición 1.3.1, consideremos la transformación $X = \log(Y)$, donde $X \sim \text{SN}(\log(\xi), \omega^2, \lambda)$. El Jacobiano de la transformación es dado por $J(x \to y) = 1/y$. Por lo tanto, mediante (1.4) encontramos que la FDP de Y es dada por

$$\begin{aligned} f_Y(y) &= \frac{1}{y}\varphi(\log(y); \log(\xi), \omega^2, \lambda) \\ &= \frac{2}{y\,\omega}\varphi\left(\frac{\log(y) - \log(\xi)}{\omega}\right)\Phi\left(\lambda\,\frac{\log(y) - \log(\xi)}{\omega}\right) \\ &= \frac{2}{y\,\omega}\varphi\left(\frac{\log(y/\xi)}{\omega}\right)\Phi\left(\lambda\,\frac{\log(y/\xi)}{\omega}\right), \quad y > 0. \end{aligned}$$

En el Teorema 1.3.2 afirmamos que si $Y \sim \text{LSN}(\xi, \omega^2, \lambda)$, entonces todos los cuantiles de Y son proporcionales al parámetro ξ.

Teorema 1.3.2. *Si* $Y \sim \text{LSN}(\xi, \omega^2, \lambda)$, *entonces el α-cuantil y_α de Y, $\alpha \in (0, 1)$, satisface*

$$y_\alpha = \xi \exp(\omega\, q_\alpha), \tag{1.10}$$

donde q_α es el α-cuantil de $Z \sim \text{SN}(\lambda)$.

Demostración. Sea $Y \sim \text{LSN}(\xi, \omega^2, \lambda)$ y $\alpha \in (0,1)$. El α-cuantil y_α de Y es dado por

$$\begin{aligned} P(Y \leq y_\alpha) = \alpha &\Leftrightarrow P(\log(Y) \leq \log(y_\alpha)) = \alpha \\ &\Leftrightarrow P\left(\frac{\log(Y) - \log(\xi)}{\omega} \leq \frac{\log(y_\alpha) - \log(\xi)}{\omega}\right) = \alpha. \end{aligned} \tag{1.11}$$

Como $Y \sim \text{LSN}(\xi, \omega^2, \lambda)$, entonces $\log(Y) \sim \text{SN}(\log(\xi), \omega^2, \lambda)$ y $Z = (\log(Y) - \log(\xi))/\omega \sim \text{SN}(0, 1, \lambda)$. Por lo tanto, a partir de (1.11) tenemos que

$$P\left(Z \leq \frac{\log(y_\alpha) - \log(\xi)}{\omega}\right) = \alpha,$$

de lo cual concluimos que $(\log(y_\alpha) - \log(\xi))/\omega = q_\alpha$, donde q_α es el α-cuantil de Z. Así, $y_\alpha = \xi \exp(\omega\, q_\alpha)$.

El Teorema 1.3.2 motiva el estudio de regresión a través de la distribución LSN, ya que si dotamos al parámetro de escala ξ de una estructura de regresión, entonces todos los cuantiles de la variable respuesta Y serán afectados por un conjunto de variables explicativas. Por ejemplo, si suponemos que $\log(\xi) = \sum_{j=i}^{r} x_j \beta_j$, donde cada β_j es un parámetro de regresión desconocido y las x's son variables explicativas, tendríamos que $\exp(\beta_j)$ es el efecto multiplicativo por cambio unitario en x_j sobre los cuantiles de Y. En el Capítulo 3 profundizaremos en esta metodología.

Sea $Y \sim \text{LSN}(\xi, \omega^2, \lambda)$, un coeficiente de variación basado en cuantiles [33] para Y se define por

$$\text{CV}_Y = \frac{3}{4}\,\frac{y_{3/4} - y_{1/4}}{y_{1/2}}. \tag{1.12}$$

En el Lema 1.3.1 establecemos una función monótona creciente que nos facilitará la interpretación del parámetro ω como un parámetro de dispersión relativa de la distribución de Y.

Lema 1.3.1. *Sean a, b y c constantes tales que $a < b < c$. Para $x \in \mathbb{R}$, la función $h(x)$ definida por*

$$h(x) = \frac{3}{4}\,\frac{\exp(cx) - \exp(ax)}{\exp(bx)} \tag{1.13}$$

es monótona creciente.

Demostración. Sean a, b, y c constantes tales que $a < b < c$. Derivando (1.13) obtenemos que

$$\begin{aligned} h'(x) =& \frac{3}{4}\,\frac{\exp(bx)(c\exp(cx) - a\exp(ax)) - b\exp(bx)(\exp(cx) - \exp(ax))}{\exp(2bx)} \\ =& \frac{3}{4\,\exp(bx)}[c\exp(cx) - a\exp(ax) - b\exp(cx) + b\exp(ax)] \\ =& \frac{3}{4\,\exp(bx)}[(c-b)\exp(cx) + (b-a)\exp(ax)]. \end{aligned}$$

Como $a < b < c$, entonces $(c-b) > 0$ y $(b-a) > 0$; además $\exp(x) > 0$ para todo x. Por lo tanto, $h'(x) > 0$. Así, la función $h(x)$ es monótona creciente.

En el Teorema 1.3.3 establecemos la relación entre el CV_Y y la función (1.13) definida en el Lema 1.3.1.

Teorema 1.3.3. *Sea* $Y \sim \mathrm{LSN}(\xi, \omega^2, \lambda)$*, entonces el* CV_Y *definido en* (1.12) *se puede escribir como función de* ω *mediante la forma* (1.13)*.*

Demostración. Sea $Y \sim \mathrm{LSN}(\xi, \omega^2, \lambda)$ y q_α el α-cuantil de $Z \sim \mathrm{SN}(\lambda)$, $\alpha \in (0,1)$. Por lo tanto, reemplazando (1.10) en (1.12) obtenemos que

$$\mathrm{CV}_Y = \frac{3}{4}\,\frac{\exp(q_{3/4}\,\omega) - \exp(q_{1/4}\,\omega)}{\exp(q_{1/2}\,\omega)}. \tag{1.14}$$

Sabiendo que $q_{1/4} < q_{1/2} < q_{3/4}$, obtenemos el resultado.

Como consecuencia del Teorema 1.3.3 tenemos que para una variable aleatoria Y con distribución LSN, el CV_Y depende del parámetro ω a través de la función dada en (1.14), la cual es monótona creciente, es decir, al aumentar (o disminuir) el valor de ω aumenta (o disminuye) el valor de CV_Y. Por lo tanto, podemos interpretar a ω como un parámetro de dispersión relativa de la distribución de Y.

Capítulo 2

Modelo de Regresión Lineal log-skew-normal

En este capítulo definimos el modelo de regresión lineal log-skew-normal el cual nos permitirá modelar la relación entre los cuantiles de una variable aleatoria positiva y un conjunto de variables explicativas. Adicionalmente, describimos la estimación de los parámetros del modelo con base en el método de máxima verosimilitud y establecemos una interpretación para los coeficientes de regresión. Finalmente, definimos intervalos de confianza y pruebas de hipótesis para los coeficientes de regresión basados en la matriz de información observada.

2.1. Definición

Sean $Y_1, \ldots, Y_n$ variables aleatorias independientes que representan las mediciones de una variable aleatoria positiva Y en n individuos. El modelo de regresión lineal log-skew-normal se define como

$$\begin{cases} Y_i \overset{\text{ind}}{\sim} \text{LSN}(\xi_i, \omega^2, \lambda) \\ \log(\xi_i) = \boldsymbol{x}_i' \boldsymbol{\beta}, \end{cases} \tag{2.1}$$

para $i = 1, \ldots, n$, donde $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir})'$ es el vector que contiene las mediciones del i-ésimo individuo en las variables explicativas $x_1, \ldots, x_r$. De esta manera, x_{ij} representa el valor observado del i-ésimo individuo en la j-ésima variable explicativa, $j = 1, \ldots, r$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)'$ es el vector de los parámetros de regresión. Los parámetros a estimar en el modelo (2.1) son $\beta_1, \ldots, \beta_r, \omega$ y λ, para un total de $r + 2$ parámetros.

El modelo de regresión lineal (2.1) permite estudiar la relación entre cuantiles de

la variable respuesta Y y un conjunto de variables explicativas $x_1, \ldots, x_r$. Además, el Teorema 1.3.2 nos permite establecer una interpretación para los parámetros de regresión. En efecto, a partir de (1.10) tenemos que la forma funcional del α-cuantil y_α de Y, $\alpha \in (0,1)$, bajo el modelo (2.1), es dado por

$$y_\alpha = \exp\left(\sum_{j=1}^{r} \beta_j x_j + \omega\, q_\alpha\right), \tag{2.2}$$

donde q_α es el α-cuantil de la variable aleatoria skew-normal estándar $Z \sim \text{SN}(\lambda)$. Por lo tanto, $\exp(\beta_j)$ es el efecto multiplicativo por cambio unitario de x_j, cuando todas las demás variables explicativas se mantienen fijas. Esto quiere decir que todos los cuantiles de Y se ven afectados a través de otras variables.

De (2.2) observamos que el α-cuantil de Y varía de acuerdo a un valor común q_α ponderado por el parámetro de dispersión relativa. Por lo tanto, la estimación de $y_{\alpha_1}, \ldots, y_{\alpha_m}$ requiere las estimaciones de $\boldsymbol{\beta}$, ω y λ obtenidas en un solo ajuste del modelo (2.1) y el cálculo separado de $q_{\alpha_1}, \ldots, q_{\alpha_m}$.

En el modelo (2.1) tenemos que $Y_i \overset{\text{ind}}{\sim} \text{LSN}(\xi_i, \omega^2, \lambda)$, para $i = 1, \ldots, n$, donde $\log(\xi_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, lo cual es equivalente a

$$\log(Y_i) \overset{\text{ind}}{\sim} \text{SN}(\boldsymbol{x}_i'\boldsymbol{\beta}, \omega^2, \lambda), \tag{2.3}$$

de esta manera, el modelo (2.1) es equivalente a un modelo de regresión lineal skew-normal [9] con respuesta $\log(Y)$. Por lo tanto, los parámetros involucrados en el modelo de regresión lineal LSN pueden ser estimados a través de un modelo de regresión lineal skew-normal con respuesta $\log(Y)$.

2.2. Estimación de los parámetros

Denotemos por $\widehat{\boldsymbol{\beta}}, \widehat{\omega}$ y $\widehat{\lambda}$ a los estimadores de máxima verosimilitud de $\boldsymbol{\beta}$, ω y λ, respectivamente.

Sean $y_1, \ldots, y_n$ los valores observados de una muestra aleatoria $Y_1, \ldots, Y_n$, a partir de (2.3) tenemos que la función de log-verosimilitud es dada por $\ell = \sum_{i=1}^{n} \ell_i$, donde

$$\begin{aligned}
\ell_i &= \log(2) - \log(\omega) + \log\left\{\varphi\left(\frac{\log(y_i) - \boldsymbol{x}_i'\boldsymbol{\beta}}{\omega}\right)\right\} + \log\left\{\Phi\left(\lambda\,\frac{\log(y_i) - \boldsymbol{x}_i'\boldsymbol{\beta}}{\omega}\right)\right\} \\
&= \log(2) - \log(\omega) + \log\left\{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\log(y_i) - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\omega^2}\right)\right\} + \log\left\{\Phi\left(\lambda\,\frac{\log(y_i) - \boldsymbol{x}_i'\boldsymbol{\beta}}{\omega}\right)\right\} \\
&= \log(2/\sqrt{2\pi}) - \log(\omega) - \frac{(\log(y_i) - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\omega^2} + \log\left\{\Phi\left(\lambda\,\frac{\log(y_i) - \boldsymbol{x}_i'\boldsymbol{\beta}}{\omega}\right)\right\}.
\end{aligned} \tag{2.4}$$

Si tomamos $v_i = (\log(y_i) - \boldsymbol{x}_i'\boldsymbol{\beta})/\omega$, las componentes del vector score para un i fijo se pueden deducir de (2.4), y son dadas por las siguientes derivadas [9, Sec. 3.1]

$$\begin{aligned}\frac{\partial \ell_i}{\partial \boldsymbol{x}_i'\boldsymbol{\beta}} &= \frac{v_i}{\omega} - \frac{\lambda\varphi(\lambda v_i)}{\omega\Phi(\lambda v_i)}\\ \frac{\partial \ell_i}{\partial \omega} &= -\frac{1}{\omega} + \frac{v_i^2}{\omega} - \frac{\lambda v_i\varphi(\lambda v_i)}{\omega\Phi(\lambda v_i)}\\ \frac{\partial \ell_i}{\partial \lambda} &= \frac{v_i\varphi(\lambda v_i)}{\Phi(\lambda v_i)}.\end{aligned} \tag{2.5}$$

Para encontrar $\widehat{\boldsymbol{\beta}}, \widehat{\omega}$ y $\widehat{\lambda}$ igualamos a cero la suma correspondiente de términos (2.5), es decir

$$\begin{aligned}\sum_{i=1}^{n} v_i - \lambda\sum_{i=1}^{n}\varphi(\lambda v_i)/\Phi(\lambda v_i) &= 0\\ \sum_{i=1}^{n} v_i^2 - \lambda\sum_{i=1}^{n} v_i\varphi(\lambda v_i)/\Phi(\lambda v_i) &= n\\ \sum_{i=1}^{n} v_i\varphi(\lambda v_i)/\Phi(\lambda v_i) &= 0.\end{aligned} \tag{2.6}$$

La presencia del término no lineal $\varphi(\lambda v_i)/\Phi(\lambda v_i)$ en las ecuaciones (2.6) hace que no sea posible encontrar la solución explícita para estas ecuaciones, por lo tanto, se deben emplear procedimientos de búsqueda numérica implementados en paquetes computacionales. Por esta razón, el análisis del modelo de regresión lineal propuesto se llevará a cabo utilizando el software `R` [31].

Azzalini [9] propone considerar la maximización directa de la función de log-verosimilitud para encontrar $\widehat{\boldsymbol{\beta}}, \widehat{\omega}$ y $\widehat{\lambda}$. Para realizar este trabajo numérico usaremos el paquete `sn` [5] en `R`, este paquete tiene disponible varios métodos de optimización numérica entre los cuales se encuentran algunos métodos basados en el método cuasi-Newton [27, Cap. 6], el método Nelder-Mead [27, Sec. 9.5], un método de gradiente conjugado [27, Sec. 5.2], entre otros.

2.3. Intervalos de confianza y pruebas de hipótesis

Sea $\boldsymbol{\theta} = (\boldsymbol{\beta}', \omega, \lambda)'$ y $\widehat{\boldsymbol{\theta}}$ el estimador de máxima verosimilitud de $\boldsymbol{\theta}$, vamos a aproximar la distribución de $\widehat{\boldsymbol{\theta}}$ mediante una distribución normal con matriz de covarianza dada por la matriz inversa de la matriz de información observada, en [9, Sec. 3.1] podemos encontrar una expresión para esta matriz.

Sea $\mathcal{I}$ la matriz de información observada para $\boldsymbol{\theta}$, la cual es de orden $(r+2)\times(r+2)$. La teoría asintótica [13] asegura que $\widehat{\boldsymbol{\theta}} \sim \mathrm{N}_{r+2}(\boldsymbol{\theta}, \mathcal{I}^{-1})$, donde $\mathcal{I}^{-1}$ es la matriz inversa de $\mathcal{I}$. Por lo tanto, la matriz de covarianza asintótica de $\widehat{\boldsymbol{\theta}}$ es

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\theta}}) = \mathcal{I}^{-1} \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} . \tag{2.7}$$

La raíz cuadrada de los elementos en la diagonal principal de (2.7) proporciona la estimación del error estándar asintótico para $\widehat{\theta}_k$, $k = 1, 2, \ldots r+2$. En particular para $\widehat{\beta}_j$, $j = 1, \ldots, r$, que son útiles para procedimientos inferenciales sobre los coeficientes de regresión.

Sea $\alpha \in (0,1)$ y $\widehat{\mathrm{SE}}(\widehat{\beta}_j)$ el error estándar asintótico de $\widehat{\beta}_j$, $j = 1, \ldots, r$. Un intervalo asintótico del $100(1-\alpha)\,\%$ de confianza para β_j es dado por $\widehat{\beta}_j \pm z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\widehat{\beta}_j)$, donde $z_{1-\alpha/2}$ denota el percentil $100(1-\alpha/2)$ de la distribución normal estándar. Este intervalo de confianza nos permite cuantificar el efecto de la variable explicativa x_j sobre la variable respuesta Y. Por otro lado, la hipótesis nula $\mathrm{H}_0 : \beta_j = 0$ puede contrastarse con la hipótesis $\mathrm{H}_1 : \beta_j \neq 0$ usando el estadístico de Wald $W = (\widehat{\beta}_j/\widehat{\mathrm{SE}}(\widehat{\beta}_j))^2$, el cual sigue una distribución asintótica χ^2 con un grado de libertad.

El paquete `sn` [5] también calcula la matriz de covarianza asintótica de $\boldsymbol{\theta}$, a partir de la cual es posible calcular los errores estándar.

Capítulo 3

Aplicación a datos de recién nacidos

Las mediciones antropométricas evaluadas al nacer ayudan a predecir el crecimiento y a identificar el riesgo de patologías en los recién nacidos, que incluso pueden llevar a situaciones de mortalidad [29]. El peso es uno de los parámetros importantes ya que el peso bajo al nacer es un factor de riesgo determinante en la muerte neonatal [19]. Se han establecido varias tablas de referencia para la medida del peso que buscan evaluar el crecimiento de los recién nacidos, dichas tablas consisten en curvas de percetiles en función del sexo y la edad gestacional al nacer [16, 32]. Nuestro enfoque es construir curvas cuántilicas para el peso bajo asimetría usando el modelo de regresión lineal log-skew-normal.

3.1. Descripción de los datos

El conjunto de datos corresponde a los nacimientos reportados por el Hospital Manuel Uribe Ángel, de enero de 2018 hasta septiembre de 2019, en el municipio de Envigado, Antioquia, Colombia [14]. La variable respuesta en este estudio es el peso (en gramos) del recién nacido. Las variables explicativas utilizadas son: sexo (S; 0 para femenino, 1 para masculino), tiempo de gestación (TDG; en semanas), edad de la madre (EM; en años) y edad del padre (EP; en años). Descartamos algunos casos con valores negativos para la edad del padre y datos faltantes, obteniendo una muestra con 9824 mediciones. Nuestro objetivo es estudiar la relación entre el peso y las variables explicativas y determinar cómo varían los cuantiles del peso de acuerdo a las variables explicativas.

La Figura 3.1 proporciona un análisis descriptivo de los datos. En las Figuras 3.1(a) y (b) observamos que los cuantiles del peso se ven afectados por el sexo y el tiempo de gestación, mientras que la edad de la madre y del padre parece tener un efecto más débil en los cuantiles del peso, según lo indican las Figuras 3.1(c) y (d). Pa-

ra estudiar estas relaciones ajustamos el modelo de regresión lineal log-skew-normal $Y_i \overset{\text{ind}}{\sim} \text{LSN}(\xi_i, \omega^2, \lambda)$, donde

$$\log(\xi_i) = \beta_1 + \beta_2 \text{S}_i + \beta_3 \text{TDG}_i + \beta_4 \text{EM}_i + \beta_5 \text{EP}_i, \tag{3.1}$$

para $i = 1, \ldots, 9824$.

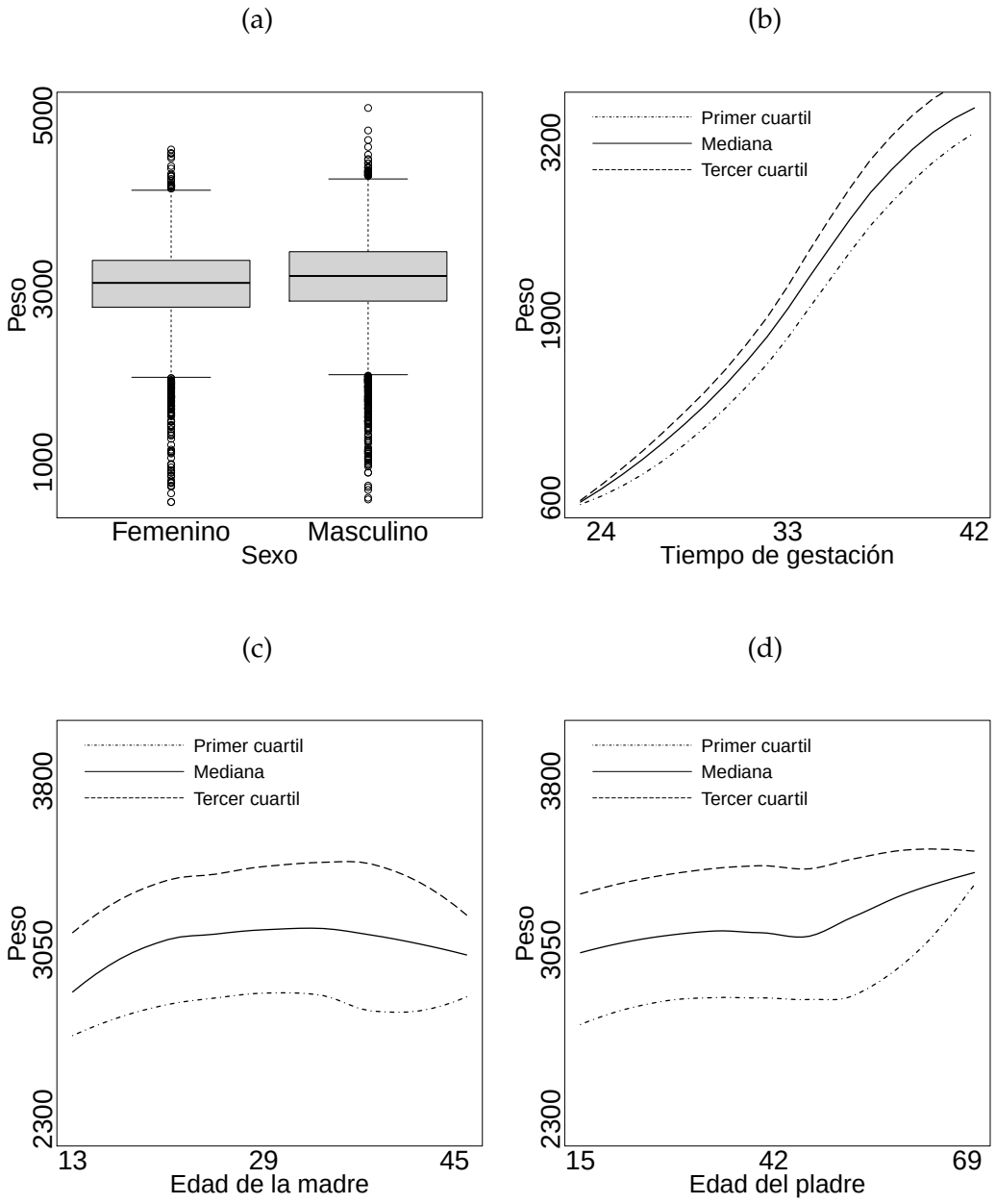

Figura 3.1: Análisis descriptivo: Boxplots comparativos (a) peso vs. sexo; líneas de cuartiles (b) peso vs. tiempo de gestación, (c) peso vs. edad de la madre, (d) peso vs. edad del padre.

En la tabla 3.1 presentamos los estimadores de máxima verosimilitud para los parámetros de regresión , los errores estándar asintóticos, intervalos de confianza del $99\,\%$ y el valor-p de la prueba de Wald para contrastar $\mathrm{H}_0: \beta_j = 0$ contra $\mathrm{H}_1: \beta_j \neq 0,\ j = 1,\ldots,5$ para el modelo (3.1). Notamos que para este modelo, el sexo, el tiempo de gestación y la edad de la madre son variables significativas, mientras que la edad del padre no lo es.

Eliminando la variable edad del padre, ajustamos el modelo $Y_i \overset{\text{ind}}{\sim} \mathrm{LSN}(\xi_i, \omega^2, \lambda)$, donde

$$\log(\xi_i) = \beta_1 + \beta_2 \mathrm{S}_i + \beta_3 \mathrm{TDG}_i + \beta_4 \mathrm{EM}_i, \tag{3.2}$$

para $i = 1,\ldots,9824$. La tabla 3.2 indica que las variables sexo, tiempo de gestación y edad de la madre son significativas, por lo tanto, consideramos el modelo (3.2) como el modelo final. Para el modelo final, la estimativa de máxima verosimilitud del parámetro de asimetría es $\widehat{\lambda} = -1{,}2912$ con un error estándar igual a $0{,}0644$ y la del parámetro de dispersión relativa es $\widehat{\omega} = 0{,}1604$ con un error estándar de $0{,}0025$; el valor de $\widehat{\lambda}$ indica el ajuste de un modelo asimétrico.

Tabla 3.1: Estimaciones, error estándar asintótico (SE), intervalos de confianza asintóticos del $99\,\%$ y valor-p de la prueba de Wald asociado al modelo (3.1)

Variable	**Peso**				
explicativa	**Estimación**	**SE**	**L. inf**	**L. sup**	**valor-*p***
Intercepto	5,3120	0,0312	5,2316	5,3925	0,0000
Sexo	0,0316	0,0025	0,0252	0,0380	0,0000
Tiempo de gest.	0,0714	0,0008	0,0694	0,0734	0,0000
Edad madre	0,0014	0,0002	0,0008	0,0021	0,0000
Edad padre	0,0004	0,0002	−0,0001	0,0010	0,0400

Tabla 3.2: Estimaciones, error estándar asintótico (SE), intervalos de confianza asintóticos del $99\,\%$ y valor-p de la prueba de Wald asociado al modelo final (3.2)

Variable	**Peso**				
explicativa	**Estimación**	**SE**	**L. inf**	**L. sup**	**valor-*p***
Intercepto	5,3163	0,0312	5,2360	5,3966	0,0000
Sexo	0,0316	0,0025	0,0252	0,0380	0,0000
Tiempo de gest.	0,0714	0,0008	0,0694	0,0734	0,0000
Edad madre	0,0018	0,0002	0,0012	0,0023	0,0000

3.2. Modelado de cuantiles

La ecuación (2.2) nos permitió establecer una interpretación de los parámetros de regresión para estudiar la manera en que las variables explicativas afectan los cuantiles de la variable respuesta. A partir del modelo final (3.2) tenemos que la forma funcional del α-cuantil del peso, y_α, $\alpha \in (0,1)$, es dado por

$$\widehat{y}_\alpha = \exp(5{,}3163 + 0{,}0316\,\mathrm{S} + 0{,}0316\,\mathrm{TDG} + 0{,}0018\,\mathrm{EM} + 0{,}1604\,\widehat{q}_\alpha), \tag{3.3}$$

donde $\widehat{q}_\alpha$ es el α-cuantil de $Z \sim \mathrm{SN}(-1{,}2912)$. En la figura 3.2 presentamos las curvas cuantílicas ajustadas para el peso vs. el tiempo de gestación para los percentiles $0{,}5; 5; 25; 50; 75; 95$ y $99{,}5$ del modelo final. En este caso fijamos la edad de la madre en su promedio para cada nivel de la variable sexo, $26{,}18$ años para femenino y $26{,}17$ años para masculino.

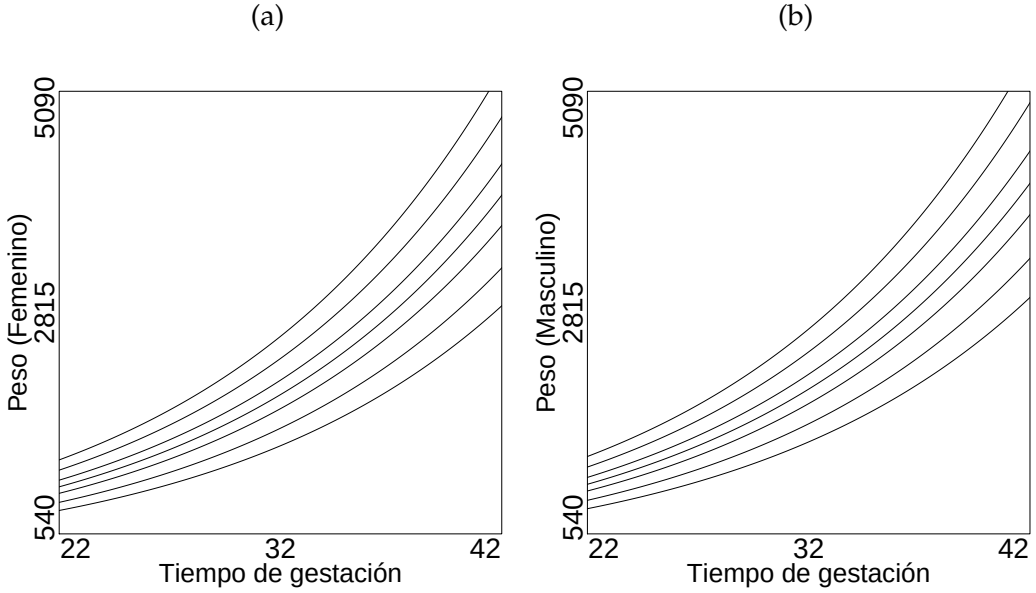

Figura 3.2: Curvas cuantílicas ajustadas (para los percentiles $0{,}5; 5; 25; 50; 75; 95$ y $99{,}5$ de abajo hacia arriba) para peso vs. tiempo de gestación, para la edad de la madre fijada en su promedio. (a) femenino; (b) masculino.

Capítulo 4

Conclusiones y sugerencias

En este trabajo estudiamos la familia de distribuciones skew-normal y presentamos algunas de sus propiedades básicas. Partiendo del estudio de esta familia, definimos la distribución log-skew-normal que se basa en aplicar una transformación logarítmica a una variable aleatoria con distribución skew-normal. Mostramos que la distribución log-skew-normal es adecuada para modelar datos positivos y asimétricos. También estudiamos algunas propiedades importantes que nos permitieron establecer la interpretación de sus parámetros y su relación con los cuantiles de la variable de interés, lo cual hace que esta distribución sea llamativa para fines de modelado estadístico usando regresión. Asimismo, propusimos el modelo de regresión lineal log-skew-normal que permite estudiar la relación entre un conjunto de variables explicativas y su relación con cuantiles de la variable respuesta. Además, establecimos la relación de este modelo con el modelo de regresión lineal skew-normal y definimos intervalos de confianza y pruebas de hipótesis para los coeficientes de regresión basados en la matriz de información observada. Finalmente, presentamos una aplicación a datos de recién nacidos donde se discutió una alternativa para la construcción de curvas cuantílicas que son ampliamente utilizadas en medicina. Es de resaltar que el modelo de regresión lineal log-skew-normal es nuevo dentro de la literatura estadística.

En trabajos futuros buscamos desarrollar procedimientos de diagnóstico para el modelo de regresión lineal log-skew-normal y estudiar las distribuciones skew-normal y log-skew-normal multivariadas con el objetivo de proponer modelos de regresión que permitan considerar asociaciones entre las variables respuesta. Adicionalmente, abordaremos otras familias de distribuciones, como por ejemplo la log-skew-t y log-skew-slash y sus versiones multivariadas, que sean apropiadas para modelar datos positivos asimétricos y con presencia de valores atípicos. Igualmente, en futuras investigaciones compararemos nuestra metodología con regresión cuantilica clásica y otras metodologías paramétricas y no-paramétricas para el modelado de cuantiles.

Bibliografía

[1] D. Allard y P. Naveau. A new spatial skew-normal random field model. *Communications in Statistics. Theory and Methods*, 36(9):1821–1834, 2007.

[2] R.B. Arellano-Valle y A. Azzalini. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33(3):561–574, 2006.

[3] A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985.

[4] A. Azzalini. Further results on a class of distributions which includes the normal ones. *Statistica*, 46(2):199–208, 1986.

[5] A. Azzalini. *The R package* `sn`*: The Skew-Normal and Related Distributions such as the Skew-t (version 1.6-2).* Università di Padova, Italia, 2020. URL `http://azzalini.stat.unipd.it/SN`.

[6] A. Azzalini, T. Cappello, y S. Kotz. Log-skew-normal and log-skew-t distributions as models for family income data. *Journal of Income Distribution*, 11(3-4):12–20, 2002.

[7] A. Azzalini y A. Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

[8] A. Azzalini y G. Regoli. The work of Fernando de Helguero on non-normality arising from selection. *Chilean Journal of Statistics (ChJS)*, 3(2), 2012.

[9] A. Azzalini with the collaboration of A. Capitanio. *The skew-normal and related families*. Cambridge University Press, 2014.

[10] G. E. P Box y D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[11] M.D. Branco y D. Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79(1):99–113, 2001.

[12] A Capitanio, A Azzalini, y E. Stanghellini. Graphical models for skew-normal variates. *Scandinavian Journal of Statistics*, 30(1):129–144, 2003.

[13] M. Chiogna. A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution. *Statistical Methods and Applications*, 14(3):331–341, 2005.

[14] Portal de Datos Abiertos del estado Colombiano. Nacidos Vivos en Hospital Manuel Uribe Angel. URL https://www.datos.gov.co/Salud-y-Protecci-n-Social/Nacidos-Vivos-en-Hospital-Manuel-Uribe-Angel/udqu-ifxr.

[15] M. M. de Queiroz, R. H. Loschi, y R. W. C. Silva. Multivariate log-skewed distributions with normal kernel and their applications. *Statistics*, 50(1):157–175, 2016.

[16] T. R. Fenton y J. H. Kim. A systematic review and meta-analysis to revise the Fenton growth chart for preterm infants. *BMC pediatrics*, 13(1):1–13, 2013.

[17] M. G. Genton. *Skew-elliptical distributions and their applications: a journey beyond normality*. CRC Press, 2004.

[18] M. G. Genton y K. R. Thompson. Skew-elliptical time series with application to flooding risk. En *Time series analysis and applications to geophysical systems*, págs. 169–185. Springer, 2004.

[19] N. W. Gidi, M. Berhane, T. Girma, A. Abdissa, R. Lim, K. Lee, C. Nguyen, y F. Russell. Anthropometric measures that identify premature and low birth weight newborns in Ethiopia: a cross-sectional study with community follow-up. *Archives of disease in childhood*, 105(4):326–331, 2020.

[20] R. Koenker y G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, págs. 33–50, 1978.

[21] C. Ley. Flexible modelling in statistics: past, present and future. *Journal de la Société Française de Statistique*, 156(1):76–96, 2015.

[22] X. Liao, Z. Peng, y S. Nadarajah. Tail properties and asymptotic expansions for the maximum of the logarithmic skew-normal distribution. *Journal of Applied Probability*, 50(3):900–907, 2013.

[23] G. D. Lin y J. Stoyanov. The logarithmic skew-normal distributions are moment-indeterminate. *Journal of Applied Probability*, 46(3):909–916, 2009.

[24] B. Liseo y N. Loperfido. A bayesian interpretation of the multivariate skew-normal distribution. *Statistics & probability letters*, 61(4):395–401, 2003.

[25] Y. Marchenko y M. Genton. Multivariate log-skew-elliptical distributions with applications to precipitation data. *Environmetrics*, 21(3-4):318–340, 2010.

[26] R. A. Morán-Vásquez, M. A. Mazo-Lopera, y S. L. P Ferrari. Quantile modeling through multivariate log-normal/independent linear regression models with application to newborn data. *Biometrical Journal*, 2021.

[27] J. Nocedal y S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[28] G. A. Paula. *Modelos de regressão: com apoio computacional*. IME-USP, São Paulo, 2004.

[29] C. B. Paulsen, B. B. Nielsen, O. A. Msemo, S. L. Møller, J. R. Ekmann, T. G. Theander, I. C. Bygbjerg, J. P. A. Lusingu, D. T. R. Minja, y C. Schmiegelow. Anthropometric measurements can identify small for gestational age newborns: a cohort study in rural Tanzania. *BMC pediatrics*, 19(1):1–10, 2019.

[30] M. Pourahmadi. Construction of skew-normal random variables: Are they linear combinations of normal and half-normal? 2007.

[31] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

[32] A. A. Rashidi, O. Kiani, M. Heidarzadeh, B. Imani, M. Nematy, A. Taghipour, M. Sadr, y A. Norouzy. Reference curves of birth weight, length, and head circumference for gestational age in Iranian singleton births. *Iranian Journal of Pediatrics*, 28(5), 2018.

[33] R. A. Rigby y D. M. Stasinopoulos. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, 6(3):209–229, 2006.