Anales Científicos, 77 (2): 319-328 (2016)
ISSN 0255-0407 (Versión impresa)
ISSN 2519-7398 (Versión electrónica)

DOI: http://dx.doi.org/10.21704/ac.v77i2.809

Presentado:18/10/2016
Aceptado: 22/12/2016

UNA APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL CON ERRORES AJUSTADOS BAJO LA DISTRIBUCIÓN SKEW-NORMAL

AN APPLICATION OF THE LINEAR REGRESSION MODEL WITH ERRORS ADJUSTED UNDER DISTRIBUTION SKEW-NORMAL

[1]Carlos López de Castilla Vásquez & [2]Paul Antonio Loarte Laos

Resumen

El presente trabajo de investigación tiene como objetivo principal mostrar que el modelo de regresión lineal múltiple *skew-normal*, desde el marco de la estimación por máxima verosimilitud, representa apropiadamente el fenómeno de asimetría en la distribución de los errores que podría estar presente en diferentes áreas de investigación: médico, experimental, actuarial, económico, etc. La aplicación del modelo propuesto se desarrolla en un estudio relativo al índice de masa corporal (*BMI*) usando dos conjuntos de datos provenientes del Instituto Australiano del Deporte. Se estimaron el modelo de regresión lineal múltiple *skew-normal* y el modelo de regresión lineal clásico. Estos modelos fueron comparados usando el Criterio de Información de Akaike (AIC) y el Logaritmo de la Función de Verosimilitud (LogVerosimilitud) obteniendo mejores resultados con la regresión *skew-normal* dado el comportamiento asimétrico de los errores.

Palabras Clave: *Regresión, AIC, skew-normal, BMI.*

Summary

The present research has as main objective to show that the skew-normal multiple linear regression model, from the framework of estimation by maximum likelihood, represents appropriately the phenomenon of asymmetry in the distribution of errors that might be present in different research areas: medical, experimental, actuarial, economic, etc. The application of the proposed model is developed in a study on the body mass index (BMI) using two sets of data from the Australian Institute of Sport. The skew-normal multiple linear regression model and the classic linear regression model was estimated. These models were compared using the Akaike Information Criterion (AIC) and the logarithm of the likelihood function (Log Verosimilitud), getting best results with the skew-normal regression given the asymmetric behavior in the errors.

Keywords: *Regression, AIC, skew-normal, BMI.*

1. Introducción

En el desarrollo de muchas de las metodologías estadísticas es necesario que los datos cuantitativos se ajusten a la distribución normal (distribución simétrica). Sin embargo, en muchas áreas de estudio los datos suelen presentar, de manera natural, un comportamiento *asimétrico*. Una distribución de probabilidad que contempla esta asimetría es la denominada *skew-normal* ampliamente explorada por Azzalini (1985), donde la distribución normal es un caso particular de dicha distribución. Dentro de la regresión lineal clásica se asume que los errores se ajustan a la distribución normal. Sin embargo, en numerosas situaciones el comportamiento de los errores también suele ser asimétrico. He aquí la necesidad de estudiar un modelo de regresión lineal asumiendo una distribución con características matemáticas similares a la distribución normal y capaz de reproducir el fenómeno de asimetría presentado por los errores del modelo.

En este trabajo de investigación se presenta el modelo

[1]Docente del Departamento Académico de Estadística Informática de la Universidad Nacional Agraria La Molina, Lima, Perú. E-mail: clopez@lamolina.edu.pe
[2]Ingeniero Estadístico Informático de la Universidad Nacional Agraria La Molina, Lima, Perú.

de regresión *skew-normal*, introducido por Azzalini y Capitanio (1999), desde el enfoque de estimación por máxima verosimilitud, en el que los errores del modelo de regresión se ajustan a una distribución que contempla la simetría y no asimetría de los mismos, teniendo en cuenta una reparametrización para la distribución *skew-normal* de Sahu, Dey y Branco (2003).

En la aplicación se compara el comportamiento del modelo de regresión *skew-normal* frente al modelo de regresión lineal clásico usando el Criterio de Información de Akaike (AIC) y el criterio del Logaritmo de la Función de Verosimilitud para identificar el mejor modelo usando la librería SN implementada en el software estadístico R desarrollado por Azzalini sobre un conjunto de datos del Instituto Australiano del Deporte.

2. Materiales y métodos

Tipo de investigación

La presente investigación es de tipo exploratoria, descriptiva y correlacional.

Diseño de investigación

Es una investigación no experimental del tipo transversal, dado que en esta investigación los datos se colectaron en un determinado espacio y tiempo.

Formulación de hipótesis

El modelo de regresión lineal *skew-normal* asume una distribución capaz de reproducir el fenómeno de asimetría, presentado por los errores del modelo, a diferencia de la regresión lineal clásica.

Población y muestra

Los conjuntos de datos provienen de muestras de atletas de élite, entrenados en el Instituto Australiano del Deporte, 102 de sexo masculino y 100 de sexo femenino. El conjunto de datos se encuentra en la librería *sn* del programa R desarrollado por Adelchi Azzalini bajo la denominación *ais* (Australian Institute of Sport data).

Metodología aplicada

La estimación de los parámetros del modelo de regresión, skew-normal, se realiza vía máxima verosimilitud en el programa R Project implementado por Adelchi Azzalini. Cada uno de los grupos estudiados en esta investigación será analizado con el siguiente procedimiento:

1. Análisis descriptivo de los datos.
2. Estimación de los coeficientes del modelo de regresión lineal clásico para los dos conjuntos, antes mencionados.
3. Pruebas de normalidad para los errores de los dos conjuntos de datos.
4. Estimación de los coeficientes del modelo de regresión skew-normal para el o los conjuntos de datos cuyos errores no se ajusten a una distribución normal.
5. Estimación de los parámetros de la distribución skew-normal para los errores del modelo al cual se ha aplicado la regresión skew-normal.
6. Comparación del modelo de regresión lineal clásico frente al modelo skew-normal con diferentes criterios estadísticos (AIC y LogVerosimilitud).

Distribución Skew-Normal univariada

La distribución skew-normal ha sido introducida por Adelchi Azzalini en 1985. A partir de entonces, se ha detallado la distribución skew-normal en estudios como en Azzalini y Dalla Valle (1996), Sahu's (2003) y Azzalini y Capitanio (1999-2014).

A partir de la modelación de la simetría, se denota a f_0 como una función de densidad de probabilidad definido en $\Re^d$, $G_0(.)$ como una función de distribución continua en la recta real y $w(.)$ como una función de valor real en , tal como se muestra a continuación, $f_0(-x) = f_0(x)$, $w(-x) = w(x)$ y $G_0(-y) = 1 - G_0(y)$, para todo . Entonces se construye la función de densidad $f(x) = 2 f_0(x) G_0\{w(x)\}$ en . Ahora se considera $f_0 = \phi,\ G_0 = \Phi$, como la función de densidad normal estándar y la función de distribución acumulada respectivamente; $w(x) = \delta x$ para cualquier valor de δ, esto genera la siguiente función de densidad:

$$f(x;\delta) = 2\phi(x)\Phi(\delta x), \qquad (-\infty < x < \infty) \quad (1)$$

donde δ denota el parámetro de asimetría, para cualquier valor definido en $\Re^1$, $\phi(.)$ y $\Phi(.)$ representan la función de densidad de probabilidad y la función de distribución acumulada de la distribución normal estándar, respectivamente.

Dada la variable aleatoria Y, se tiene que:

$$Y = \mu + \sigma Z, \qquad (\mu \in \Re, \sigma \in \Re^+) \quad (2)$$

entonces la variable aleatoria continua Y tiene distribución skew-normal $Y \sim SN(\mu, \sigma^2, \delta)$ con parámetro de locación μ, parámetro de escala σ y

parámetro de asimetría δ cuya función de densidad es:

$$f(y) = 2\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\delta\left(\frac{y-\mu}{\sigma}\right)\right) \quad y \in \Re \tag{3}$$

donde $\phi(.)$ y $\Phi(.)$ representan la función de densidad de probabilidad y la función de distribución acumulada de la distribución normal estándar, respectivamente.

Distribución Skew-Normal univariada según Sahu

La parametrización de Sahu's et al.(2003) es equivalente a la usada en Azzalini y Capitanio (1999), y puede ser implementada usando el algoritmo EM (Esperanza y Maximización). La distribución *skew-normal* de Sahu facilita la implementación de algoritmos que son usados para obtener los estimadores de máxima verosimilitud con varianza de distribución finita.

Una variable aleatoria Y tiene distribución *skew-normal* $Y \sim SN(\mu, \sigma^2, \delta)$, con parámetro de locación μ, parámetro de escala σ^2 y parámetro de forma δ, si su función de densidad está dada por:

$$f_Y(y \mid \mu, \sigma^2, \delta) = \frac{2}{\sqrt{\sigma^2+\delta^2}}\phi\left(\frac{y-\mu}{\sqrt{\sigma^2+\delta^2}}\right)\Phi\left(\frac{\delta}{\sigma}\frac{y-\mu}{\sqrt{\sigma^2+\delta^2}}\right) \tag{4}$$

Si $\delta = 0$, entonces la función de distribución de probabilidad corresponde a la función de distribución normal.

Regresión lineal Skew-Normal

Sea el modelo de regresión $Y_j = \beta_1 + \sum_{i=2}^{p}\beta_i x_{ji} + \delta z_j + \varepsilon_j$ según Sahu's et al.(2003) ajustado a la distribución *skew-normal*, donde $\varepsilon_j \sim N(0, \sigma^2)$ y $z_j \sim HN_1(0, \sigma^2)$ tiene una distribución univariada estandarizada *half-normal* para $j = 1, \ldots, n$ independientes. Además, $\delta z_j + \varepsilon_j \sim SN(0, \sigma^2, \delta)$ tiene distribución skew-normal con parámetro de locación 0, parámetro escala σ^2 y parámetro de forma δ.

Si $Y_j \sim SN(x_j^T\boldsymbol{\beta}, \sigma^2, \delta)$ y se tienen n observaciones independientes para la variable respuesta Y_j unidimensional, $\boldsymbol{x}_j^T = (1, x_{j2}, \ldots, x_{jp})$ y $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$. La función de densidad de probabilidad de Y_j viene dada por:

$$f_{Y_j}(y_j \mid \boldsymbol{\theta}) = \frac{2}{\sqrt{\sigma^2+\delta^2}}\phi\left(\frac{y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}}{\sqrt{\sigma^2+\delta^2}}\right)\Phi\left(\frac{\delta}{\sigma}\frac{y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}}{\sqrt{\sigma^2+\delta^2}}\right) \tag{5}$$

donde $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2, \delta)$. El logaritmo de la función verosimilitud para $\boldsymbol{\theta}$ en la muestra observada $Y_1, \ldots, Y_n$ está dada por $l(\boldsymbol{\theta}) = \sum_{j=1}^{n} l_j(\boldsymbol{\theta})$, donde:

$$l_j(\boldsymbol{\theta}) = \log 2 - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2+\delta^2) - \frac{1}{2}A_j + \log\Phi(B_j) \tag{6}$$

con $A_j = \frac{1}{\sigma^2+\delta^2}\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}\right)^2$ y $B_j = \frac{\delta}{\sigma\sqrt{\sigma^2+\delta^2}}\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}\right)$.

Reescribiendo la función log-verosimilitud:

$$l(\boldsymbol{\theta}) = n\log 2 - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2+\delta^2) - \frac{1}{2(\sigma^2+\delta^2)}Q(\boldsymbol{\beta}) + \sum_{j=1}^{n}\log\Phi(B_j) \tag{7}$$

donde $Q(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})$.

A partir del logaritmo de la función de verosimilitud se define la función *score* para $\boldsymbol{\theta}$, la cual, está definida como la primera deriva de la función logaritmo de verosimilitud, para la estimación de los parámetros del modelo de regresión, tal como se muestra a continuación:

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(U(\boldsymbol{\beta})^T, U(\sigma^2), U(\delta)\right)^T \tag{8}$$

donde:

$$U(\boldsymbol{\beta}) = \frac{1}{\sigma^2+\delta^2}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X\beta}) - \frac{\delta}{\sigma(\sigma^2+\delta^2)}\boldsymbol{X}^T\boldsymbol{a} \tag{9}$$

$$U(\sigma^2) = -\frac{1}{2}\frac{n}{\sigma^2+\delta^2} + \frac{1}{2(\sigma^2+\delta^2)}Q(\boldsymbol{\beta}) - \frac{1}{2}\frac{2\sigma^2+\delta^2}{\sigma^2(\sigma^2+\delta^2)^{3/2}}\frac{\delta}{\sigma}(\boldsymbol{Y} - \boldsymbol{X\beta})^T\boldsymbol{a} \tag{10}$$

$$U(\delta) = -\frac{n\delta}{\sigma^2+\delta^2} + \frac{\delta}{(\sigma^2+\delta^2)^2}Q(\boldsymbol{\beta}) + \frac{\sigma}{(\sigma^2+\delta^2)^{3/2}}(\boldsymbol{Y} - \boldsymbol{X\beta})^T\boldsymbol{a} \tag{11}$$

y además $\boldsymbol{a} = (a_1, \ldots, a_n)^T$, donde $a_j = W_\Phi(B_j) = \phi(B_j)/\Phi(B_j)$, $j = 1, \ldots, n$. Ahora mediante un proceso iterativo es posible obtener los estimadores de máxima verosimilitud para $\boldsymbol{\beta}$, σ^2 y δ cuyas expresiones están dadas por:

$$\beta^{(m+1)} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left[\boldsymbol{Y} - \frac{\delta^{(m)}}{\sigma^{(m)}}\left(\sigma^{2(m)} + \delta^{2(m)}\right)^{1/2}\boldsymbol{a}\left(\boldsymbol{\theta}^{(m)}\right)\right] \tag{12}$$

$$\left(\sigma^{2(m+1)}, \delta^{(m+1)}\right)^T = \arg\max_{\sigma^2,\delta}\left[l\left(\beta^{(m+1)}, \sigma^2, \delta\right)\right] \tag{13}$$

para $m = 0,1,2,3\ldots$. Al llevar a cabo la maximización con la última ecuación (13) se considera un método multivariante secante (Dennis and Schanabel, 1996), de modo que las funciones de score usadas con esta maximización del algoritmo están dadas en (10) y (11).

Se requieren valores iniciales $\boldsymbol{\beta}^{(0)}$, $\sigma^{2^{(0)}}$ y $\delta^{(0)}$ para los procedimientos dados en las expresiones (12) y (13). Está reparametrización propuesta por Sahu et al. (2003) es equivalente a la propuesta por Azzalini and Capitanio (1999), entonces también se puede usar el algoritmo EM en el modelo estudiado.

Prueba de Shapiro-Wilk

La prueba de Shapiro-Wilk, propuesta en 1965, calcula un estadístico W que comprueba si una muestra aleatoria proviene específicamente de una distribución normal. Los valores pequeños de W son evidencia de desviaciones de la normalidad y puntos porcentuales para el estadístico W, obtenidos vía simulaciones de Monte Carlo. Esta prueba muestra buenos resultados en comparación con otras pruebas de bondad de ajuste.

Criterio de Información de Akaike (AIC)

Según Marcerau (2012), el Criterio de Información de Akaike (AIC) admite que el modelo *verdadero*; es decir, un modelo con la verdadera distribución de los datos es desconocido, y que dentro de los modelos que están siendo validados, ninguno es considerado el que realmente describe la variable en estudio, Akaike (1974). El AIC intenta escoger dentro de los modelos que están disponibles, aquel que minimice la divergencia de Kullback – Leibler (KL). El modelo con menor valor AIC es un modelo que mejor se ajusta a los datos.

Estado Nutricional

Según la Organización Mundial de la Salud (OMS), el estado nutricional se define como el estado de crecimiento o el nivel de micronutrientes de un individuo. Según la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO, siglas de Food and Agriculture Organization), el estado nutricional es la condición física que presenta una persona, como resultado del balance entre sus necesidades e ingesta de energía. Estudios de Klause & Mahan (1995) citado por Nascimento OV y Alencar FH (2007), definen el estado nutricional como la condición en la que el organismo asume un gasto metabólico de energía, donde la sumatoria de la interacción de elementos somáticos y funcionales es la responsable de la absorción de nutrientes y adecuación de las necesidades fisiológicas del individuo.

Valor nutricional

El estado nutricional refleja en cada momento si la ingestión, absorción y utilización de los nutrientes son adecuadas a las necesidades del organismo. Sobre todo en deportistas de alta competencia se tiene necesidad de una dieta que le garantice salud y bienestar a través de un correcto equilibrio nutricional y que no le provoque molestias en el transcurso de entrenamientos y competiciones.

Evaluación nutricional

Los sistemas de evaluación del estado nutricional utilizan una variedad de métodos para caracterizar las diferentes etapas de una deficiencia nutricional. La evaluación del estado nutricional utilizando mediciones antropométricas se enmarcan en la denominada antropometría. Entre las variables antropométricas, la estatura y el peso corporal son las variables más empleadas en la evaluación del estado nutricional, al ser muy simple su evaluación en el contexto del resto de las mediciones.

Índice de Masa Corporal

Se interpreta como una medida en la cual la masa del individuo se distribuye por unidad de área corporal (peso Kg/cuadrado de la estatura m^2). Los valores del IMC son un reflejo de las reservas corporales de energía. Esta afirmación se evidencia por su alta correlación con la grasa corporal estimada por métodos válidos como la densitometría, y por su alta correlación con los pliegues cutáneos que son predictores de la grasa corporal, en ese sentido el estado nutricional también depende de la evaluación de los indicadores hematológicos y bioquímicos.

Recuento de Glóbulos Rojos

Es un examen de sangre que mide la cantidad de glóbulos rojos que tiene una persona. La cantidad de oxígeno que

los tejidos corporales reciben depende de la cantidad de glóbulos rojos y su ausencia puede significar una enfermedad como la anemia.

Recuento de Glóbulos Blancos

Es un examen que mide el número de glóbulos blancos en la sangre. Los glóbulos blancos, también llamados leucocitos, ayudan a combatir infecciones. Un alto número de glóbulos blancos se denomina leucocitosis y puede deberse a anemias o reacciones infecciosas.

Hematocrito

Es un examen de sangre que mide el porcentaje del volumen de toda la sangre que está compuesta de glóbulos rojos. Esta medición depende del número de glóbulos rojos y de su tamaño. Valores anormales de esta prueba pueden estar asociados a una desnutrición.

Hemoglobina

La hemoglobina es una proteína en los glóbulos rojos que transporta oxígeno. El nivel bajo de hemoglobina puede deberse a anemia de varios tipos y nutrición deficiente.

Concentración de Hierro en plasma

Mide la cantidad de hierro que hay en la sangre. Los niveles superiores a los normales pueden significar hemólisis y anemia debido a que los glóbulos rojos se destruyen con mucha rapidez o deficiencia de vitamina B12 y de vitamina B6, entre otros.

Pliegues Cutáneos

La medida de su espesor permite estimar con bastante aproximación la cantidad de grasa subcutánea, que constituye el 50% de la grasa corporal. Con los pliegues cutáneos, se cuantifica la cantidad de tejido adiposo subcutáneo.

Porcentaje de grasa corporal

El porcentaje de grasa se obtiene dividiendo el total de masa grasa (aquella formada por grasas estructurales y de depósito) entre el total de masa corporal (peso total). El porcentaje de grasa corporal depende de muchas otras variables como el nivel de agua, tejido muscular, tejido óseo, órganos, etc; por esta razón, es que éste porcentaje no depende solamente del peso actual de la persona.

3. Resultados y discusión

Los datos del Instituto Australiano de Deporte provienen de atletas hombres y mujeres, de los cuales se han recolectado indicadores clínicos (biométricos, hematológicos y bioquímicos) tales como el índice de masa corporal (BMI), evaluado en función al conteo de glóbulos rojos (RCC), conteo de glóbulos blancos (WCC), hematocritos (Hc), hemoglobina (Hg), concentración de hierro en plasma (Fe), total de pliegues dérmicos (SSF) y el porcentaje de grasa corporal (Bfat). En la Tabla 1 se presentan las estadísticas descriptivas para los datos de los atletas separados por género (hombres y mujeres).

Tabla1. Estadísticas descriptivas de las variables biométricas, hematológicas y bioquímicas para los atletas hombres y mujeres.

		Media	Desv. Estand.	Mínimo	Máximo	Asimetría	Kurtosis	Error estándar
ATLETAS MUJERES (n_1=100)	RCC	4.40	0.32	3.80	5.33	0.68	0.24	0.03
	WCC	6.99	1.70	3.30	13.30	0.74	0.92	0.17
	Hc	40.48	2.62	35.90	47.10	0.26	-0.70	0.26
	Hg	13.56	0.92	11.60	15.90	0.09	-0.86	0.09
	Fe	56.96	30.96	12.00	182.00	1.33	2.46	3.10
	BMI	21.99	2.64	16.75	31.93	0.68	1.09	0.26
	SSF	86.97	33.85	33.80	200.80	0.76	0.56	3.39
	Bfat	17.85	5.45	8.07	35.52	0.34	-0.15	0.55
ATLETAS HOMBRES (n_2=102)	RCC	5.03	0.35	4.13	6.72	0.91	4.58	0.03
	WCC	7.22	1.90	3.90	14.30	0.85	1.49	0.19
	Hc	45.65	2.57	40.30	59.70	1.47	7.17	0.25
	Hg	15.55	0.93	13.50	19.20	0.96	2.21	0.09
	Fe	96.40	52.66	8.00	234.00	0.86	0.07	5.21
	BMI	23.90	2.77	19.63	34.42	1.39	2.87	0.27
	SSF	51.42	18.85	28.00	113.50	1.37	1.70	1.87
	Bfat	9.25	3.18	5.63	19.94	1.51	1.98	0.32

Fuente: Elaboración propia

Se observa que la concentración de hierro en plasma (Fe) y el total de pliegues dérmicos (SSF) presentan valores altos en relación a su asimetría para el grupo de atletas mujeres; análogamente, para el grupo de atletas hombres estos valores se presentan elevados en los hematocritos (Hc) y porcentaje de grasa corporal (Bfat); en líneas generales todos los indicadores presentan asimetría positiva.

En la Tabla 2 se presentan los resultados de un análisis de regresión lineal clásico para estudiar la relación entre el índice de masa corporal y sus variables biométricas, hematológicas, bioquímicas.

Ambos modelos son significativos (** al 1% y * al 5%), en ese sentido, la hemoglobina (Hg) y el total de pliegues dérmicos (SSF) son variables significativas para los dos grupos en estudio. En la Figura 1, se muestran los histogramas para los residuales correspondientes a los modelos de regresión analizados y sus respectivos gráficos de normalidad Q-Q plots.

Tabla 2. Estimación de coeficientes e indicadores del modelo de regresión lineal clásico.

	ATLETAS MUJERES		ATLETAS HOMBRES	
Intercepto	9.69	**	9.36	*
RCC	-1.10		-0.97	
WCC	0.05		-0.13	
Hc	-0.18		0.02	
Hg	1.42	**	0.92	.
Fe	0.00		0.01	
SSF	0.06	*	0.11	*
Bfat	-0.03		-0.08	
R^2	0.54		0.50	
P valor	3.8E-13		5.7E-12	

Fuente: Elaboración propia

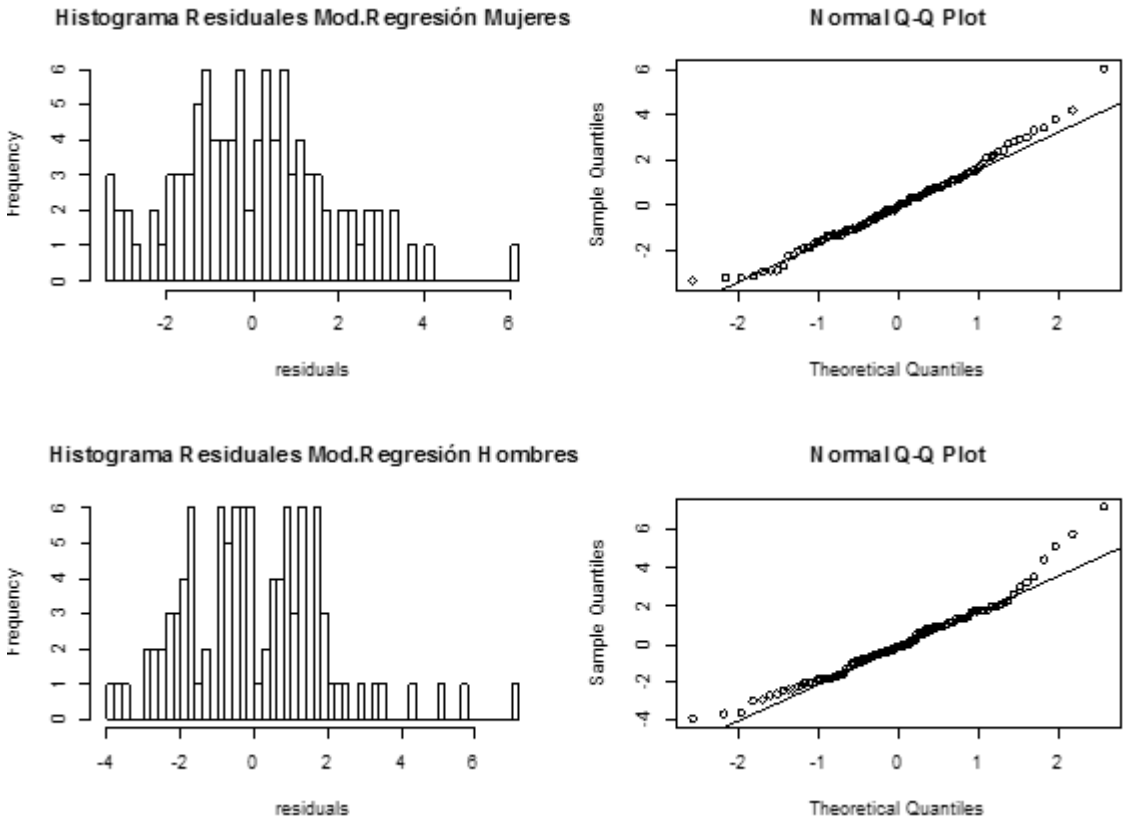

Figura 1. Gráficos descriptivos para los residuales del modelo de regresión lineal clásico para atletas mujeres.
Fuente: Elaboración propia

En los gráficos de normalidad se observa, que en el modelo correspondiente a los hombres, los errores no tienen un buen ajuste a la distribución normal; caso contario, para el grupo de las mujeres que presentaría un mejor ajuste de sus errores a la distribución normal. La Tabla 3 presenta la prueba de normalidad "Shapiro-Wilk" para los errores de los modelos analizados en mujeres y hombres.

Tabla 3. Prueba de Normalidad Shapiro-Wilk.

	Atletas Mujeres	Atletas Hombres
P valor	0.20	0.01

Fuente: Elaboración propia

Como resultado de la prueba anterior solo en el grupo de hombres existe una fuerte evidencia estadística para rechazar que los errores de su modelo de regresión tienen una distribución normal lo que nos llevaría a aplicar un modelo de regresión *skew-normal*. La estimación de los coeficientes de regresión se realiza bajo el método de máxima verosimilitud cuyos valores se muestran en la Tabla 4.

Tabla 4. Coeficientes estimados en el modelo de regresión skew-normal para el grupo de los atletas hombres.

Coeficientes	ATLETAS HOMBRES	
Intercepto	10.25	**
RCC	-0.81	
WCC	-0.07	
Hc	0.02	
Hg	0.82	.
Fe	0.00	
SSF	0.09	*
Bfat	-0.09	

Fuente: Elaboración propia

Se observa que tanto la hemoglobina (Hg) y el total de pliegues dérmicos (SSF) son variables significativas (** al 1% y * al 5%) en el modelo de regresión skew-normal. Sin embargo, es importante analizar los residuales del modelo estudiado. En la Figura 2, se presenta la distribución de estos residuales y su gráfico QQ-plot. Los gráficos a continuación muestran los residuales del modelo evaluado en el grupo de hombres.

Los errores presentan asimetría y se observa gráficamente que no se ajustan a una distribución normal, la cual está contemplada en este tipo de regresión especial, *skew-normal*.

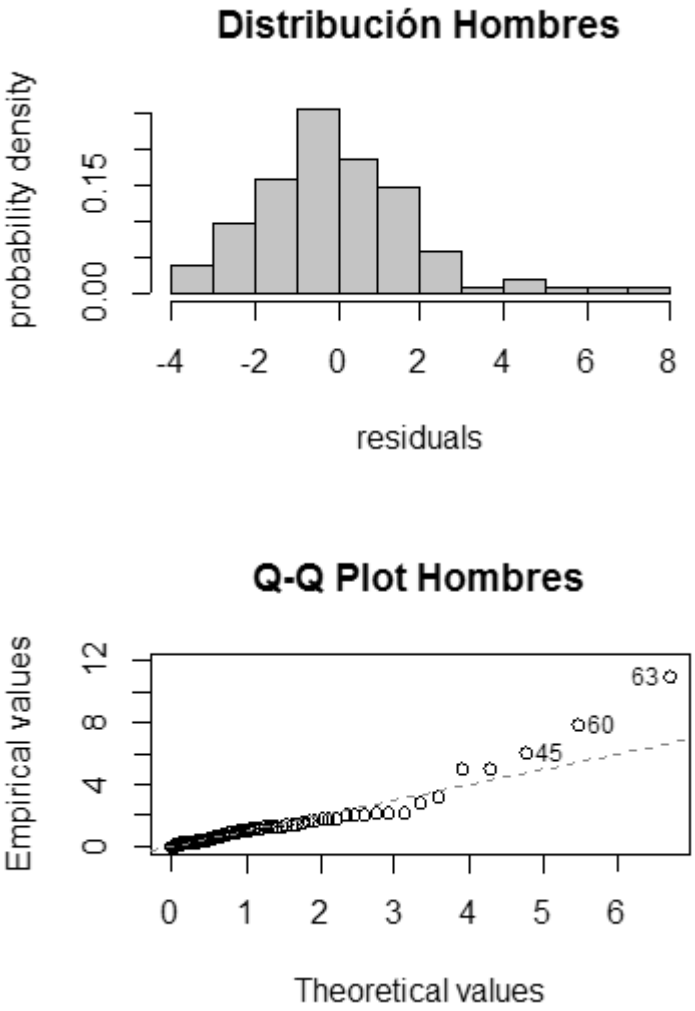

Figura 2. Gráficos para los errores del modelo de regresión skew-normal en atletas hombres.

Fuente: Elaboración propia

La Tabla 5 presenta los parámetros estimados de la distribución skew-normal de los errores del modelo de regresión en el conjunto de atletas hombres.

Tabla 5. Parámetros estimados de la distribución skew-normal para los errores en el grupo de los atletas hombres.

SKEW-NORMAL	
Parámetros	**Atletas Hombres**
Locación	1.2E-12
Escala	1.94
Forma	0.65

Fuente: Elaboración propia

Se estiman los parámetros; y en este caso en particular, el parámetro de locación tiende a cero (acorde a la teoría). La estimación de parámetros se realizó mediante el

método de máxima verosimilitud usando la librería *sn* del programa estadístico R. En consecuencia, es válida la aplicación de la regresión skew-normal para el conjunto de datos correspondiente a los hombres.

La Tabla 6, presenta dos criterios para elegir un modelo de regresión, para los modelos de regresión lineal clásica y skew-normal, los cuales son el centro de estudio de esta investigación. Según el criterio del logaritmo de la función de verosimilitud, se elige el modelo con el menor valor absoluto mientras que con el Criterio de Información de Akaike (AIC) el mejor modelo es el que presenta el valor más bajo.

El resultado obtenido en la Tabla 6 muestra que el modelo de regresión skew-normal es más adecuado que el modelo de regresión lineal clásico, para el conjunto de datos analizados. Al analizar el AIC en el modelo de regresión skew-normal para hombres, este valor resulta ser menor en comparación con el obtenido para el modelo de regresión lineal clásico; en consecuencia, indica que el modelo de regresión skew-normal es el que mejor trabaja con errores distribuidos asimétricamente, para el conjunto de atletas hombres analizados.

Si se considera el criterio del logaritmo de la función de verosimilitud, también se comprueba, por sus valores bajos (absolutos), que el modelo skew-normal es un modelo de regresión adecuado cuando estos presentan errores asimétricos en su distribución.

Se añade un análisis gráfico de residuales de los modelos, en los hombres, en la Figura 3, para el modelo de regresión lineal clásico (izquierda) y skew-normal (derecha).

Tabla 6. Criterio del Logaritmo de la función de Verosimilitud, Criterio de Información de Akaike (AIC) y el número de variables significativas.

Modelos	Indicadores	ATLETAS HOMBRES
Normal	N° Variables significativas	2
	Logaritmo de la función de Verosimilitud	-212.54
	AIC	443.09
Skew Normal	N° Variables significativas	2
	Logaritmo de la función de Verosimilitud	-208.36
	AIC	436.72

Fuente: Elaboración propia

En la Figura 3, se muestra que con la dispersión de los residuales, se cumple con el supuesto de homocedasticidad tanto para el modelo skew-normal y lineal clásico de regresión. Así mismo, se observa que los residuales del modelo skew-normal presentan una ligera mayor dispersión que en el modelo de regresión lineal clásico.

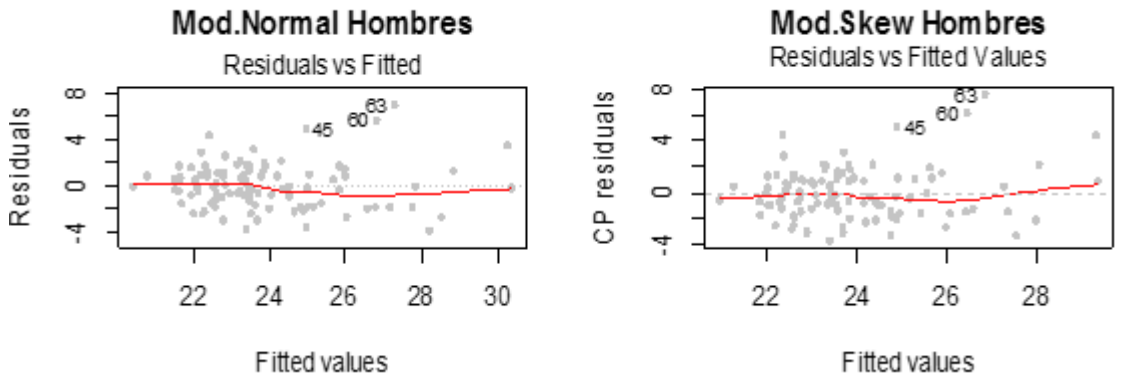

Figura 3. Gráficos para los residuales en los modelos de los atletas hombres.

Fuente: Elaboración propia

4. **Conclusiones**

Se comprobó para el conjunto de datos correspondiente a los hombres, que el modelo de regresión skew-normal, puede ajustar a su metodología la asimetría presentada por los errores; sin llegar a transformar los datos para la estimación de los coeficientes del modelo de regresión.

Los índices usados (Logverosimilitud, AIC) para validar los modelos en los hombres, reflejaron que la regresión skew-normal es apropiada para este tipo de datos, en comparación al modelo de regresión lineal clásico, dado que trabaja con distribuciones asimétricas.

En referencia al índice de masa corporal, se encontró

que el grupo de atletas hombres se ajusta mejor con un modelo de regresión skew-normal, dado la no normalidad de sus errores, mientras que para las mujeres, este indicador biométrico, se representa mejor con un modelo de regresión lineal clásico.

Se comprobó, dentro la estimación de los parámetros de la distribución skew-normal de los errores del modelo de regresión en el caso de los hombres, que el parámetro de locación se ajusta a lo esperado por la teoría.

5. Recomendaciones

Dado que, en el enfoque de esta investigación, se estiman los coeficientes de regresión skew-normal desde el marco de máxima verosimilitud y se observa el comportamiento de sus errores; para futuras investigaciones, se podría realizar el proceso completo de inferencia en este tipo de regresión.

A la fecha, para seleccionar entre un modelo u otro de la regresión skew-normal se utilizan los índices como el AIC, Logaritmo de Verosimilitud. Sin embargo, es necesario considerar que un modelo tenga estimaciones más precisas y con el menor número de variables posibles. En ese sentido, sería importante construir un estadístico de prueba para evaluar la parsimonia del modelo.

Para profundizar la investigación, sería importante realizar un análisis de multicolinealidad para el modelo de regresión skew-normal; dado que por ahora se cuenta con poca referencia bibliográfica para el tema en mención.

Para futuras investigaciones, se podría realizar una inferencia bayesiana sobre el parámetro de forma, para las distribuciones sesgadas y su respectiva aplicación a sus modelos de regresión lineal múltiple.

6. Literatura citada

Arellano-Valle, RB. y Azzalini, A. 2013. The centred parameterization and related quantities of the skew-t distribution. *Journal of Multivariate, Analysis*, 113, 73-90.

Arellano-Valle, RB., Bolfairine, H, y Lachos, VH. 2005. Skew-normal Linear Mixed Models. *Journal of Data Science*, 3, 415-438.

Arellano-Valle, RB., Ozan, S., Bolfarine, H. y Lachos, V. H. 2003. *Skew normal measurement error models*. To appear, Journal of Multivariate Analysis.

Azzalini, A. 1985. A class of distribution which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171-178.

Azzalini, A. and Capitanio, A. 1999. Statistical applications of the multivariate skew normal distribution. *J Roy Statist Soc*, B, 61, 579-602. Recuperado de http://arXiv.org/abs/0911.2093.

Azzalini, A. and Dalla Valle, A. 1996. The multivariate skew-normal distribution. *Biometrika,* 83 (4),715–726.

Azzalini, A with the collaboration of Capitanio, A. 2014. *The Skew-Normal and Related Families*. Cambridge: Cambridge University Press, IMS Monographs series.

Biblioteca Nacional de Medicina de los EE.UU/ Medlineplus. 2015. <https://www.nlm.nih.gov/medlineplus/spanish/> [Consultada: 05 de Julio de 2015].

Branco, MD; Dey, DK. 2001. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis,* 79 (1), 99-113.

Dalla Valle, A. 2004. The Skew-Normal Distribution. En M. G. Genton(editor), *Skew-Elliptical Distribution and their Applications: a journey beyond normality*, cap. I, pp 3-24 USA: Chapman & Hall/CRC.

Esteban González, MV. 2012. *Estadística Actuarial y Análisis de Regresión*. Leioa: Departamento de Economía Aplicada III-Econometría y Estadística de la Facultad de Ciencias Económicas y Empresariales de la Universidad del País Vasco/Euskal Herriko Unibertsitatea.

Ferreira, CS., Bolfarine, H. y Lachos, VH. 2011. Skew scale mixtures of normal distributions: Properties and estimation. *Statistical Methodology*, 8,(2), 154-171.

Ferreira, CS., Bolfarine, H. y Vilca, FE. 2008. *Diagnostics analysis for skew normal regression models*. Campinas: Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas.

Macerau, WM. 2012. *Comparação das distribuições* α *-estável, normal, t de student e Laplace assimétricas*. Dissertação de Mestrado em Estatística. Universidade Federal de São Carlos, Brasil.

Monterrey Gutiérrez, P; **Porrata Maury, C. 2001.** Procedimiento gráfico para la evaluación del estado nutricional de los adultos según el índice de masa corporal. *Rev Cubana Aliment Nutr*, 15 (1), 62-67. Recuperado de http://www.bvs.sld.cu/revistas/ali/vol15_1_01/ali09101.htm

Nascimento, OV; Alencar, FH. 2007. Perfil do estado nutricional do atleta adulto. *Fit Perf J*,6 (4), 241-246.

Organización de las Naciones Unidas para la Alimentación y la Agricultura. ***Papel de la FAO en la nutrición.*** <http://www.fao.org/nutrition/es/>

[Consultada: 14 de agosto de 2015].

Sahu, SK. and Chai, HS. 2008. A new skew-elliptical distribution and its properties. *Calcutta Statistical Association Bulletin*, 61 (20), 197-225.

Sahu, SK., Dey, DK. y Branco, MD. 2003. A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics*, 31 (2), 129-150.

Shapiro, SS. y Wilk, MB. 1965. An analysis of variance test for normality (complete samples), *Biometrika*, 52 (3/4), 591-611. Recuperado de http://sci2s.ugr.es/keel/pdf/algorithm/articulo/shapiro1965.pdf

Toma Inafuko, J. y Rubio Donet, J. L. 2008. *Estadística aplicada: segunda parte*. Lima: Centro de Investigación de la Universidad del Pacífico.