Revista Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 3 N°1 enero-junio 2014 • ISSN 0121-747X
Páginas 56 a 69

# PRUEBA PARA NORMALIDAD SESGADA EN EL MODELO LINEAL MIXTO CON INTERCEPTO ALEATORIO

# SKEW NORMALITY TEST IN A LINEAR MIXED MODEL WITH RANDOM INTERCEPT

MARISOL VALENCIA CÁRDENAS [a], JUAN CARLOS SALAZAR URIBE [b], JUAN CARLOS CORREA MORALES [c]

Recibido 27-03-14, aceptado 17-06-14, versión final 20-06-14.
Artículo Investigación

**RESUMEN**: Los modelos lineales mixtos se basan en el supuesto de que los efectos aleatorios y los errores son independientes y se distribuyen normalmente; sin embargo, este supuesto no siempre se satisface. Este trabajo propone una prueba para detectar normalidad sesgada de los residuales y efectos aleatorios de un modelo lineal mixto. Para esto se presenta un método gráfico usando simulación y se ilustra con unos datos reales. Se detectan casos lógicos donde los residuales del modelo lineal mixto se comportan acorde con las distribuciones Normal Sesgada y T Sesgada con la herramienta propuesta, proporcionando los análisis para mejorar la estimación.

**PALABRAS CLAVE**: Modelos lineales mixtos, Distribución normal sesgada, Datos longitudinales.

**ABSTRACT**: Linear mixed models are based on the assumption that both the random effects and the errors are independent and normally distributed. In literature, analytical and graphical methods have been proposed to validate such assumption; nevertheless, this assumption is not always satisfied. This work proposes a test to identify skew normality in residuals. According to this a graphical method is presented, using simulation, and it is illustrated with real data. Logical cases are detected showing adjustment to skew- and t- distributions with the proposed test, enabling analysts to improve estimation.

**KEYWORDS**: Linear mixed models, skew normal distribution, longitudinal data.

[a]Ingeniera Industrial, Magister en Estadística. Ph. D.(c) en Ingeniería, Industria y Organizaciones, Universidad Nacional. mvalencia@unal.edu.co

[b]Ph. D. en Estadística, University of Kentucky, Docente Universidad Nacional de Colombia, Sede Medellín. jcsalaza@unal.edu.co

[c]Ph. D. en Estadística, University of Kentucky, Docente Universidad Nacional de Colombia, Sede Medellín. jccorrea@unal.edu.co

# 1. INTRODUCCIÓN

El modelo lineal mixto permite estudiar asociaciones en datos cuya estructura es correlacionada, lo cual ocurre al introducir el efecto aleatorio al modelo lineal. Por ejemplo, las medidas repetidas a lo largo del tiempo por individuo, o medidas en espacios cercanos.

Autores como Lange & Ryan (1989) han analizado los supuestos de normalidad de los efectos aleatorios, ellos proponen un método gráfico para la detección de normalidad. Otro trabajo que ilustra los problemas de este supuesto es el de Verbeke (1996), quién describe cómo los efectos aleatorios son estimados deficientemente cuando provienen de mezclas de distribuciones gaussianas. Las técnicas para detectar desvíos de normalidad en efectos aleatorios y el error puro han recibido poca atención.

Es frecuente encontrar fenómenos cuya distribución de probabilidad refleja un comportamiento normal, especialmente cuando los tamaños de muestra aumentan. Sin embargo, y también con mucha frecuencia, se presentan otras situaciones que reflejan no normalidad, por ejemplo, en diversos casos en el análisis de los residuales de un modelo lineal y del modelo lineal mixto.

Algunos autores han estudiado la necesidad de emplear el modelo lineal mixto con componentes aleatorios y el error con distribución no normal, (Arellano, 2005, Zhou & He, 2007) mostrando otras formas de estimación del modelo bajo estas circunstancias. Otros (Gurka et al., 2006), recomiendan el uso de transformaciones sobre la variable respuesta, metodología que funciona en escenarios limitados.

El estudio del comportamiento distribucional del efecto aleatorio es importante ya que, además de explicar una variabilidad extra que no es tenida en cuenta por un modelo de sólo efectos fijos, afecta la distribución del error y además permite la "individualización" del modelo, aspecto que no puede estudiarse con un modelo lineal de sólo efectos fijos. Es necesario entonces establecer estimaciones más robustas en el modelo lineal mixto con un tipo de distribución diferente a la normal, por ejemplo, una con característica sesgada de la componente aleatoria. Una familia que reúne estas características es la normal sesgada (Azzalini & Capitanio, 1999). Por ello, es importante determinar los efectos que puede generar la procedencia de distribuciones diferentes de la normal, para el error y el efecto aleatorio en las estimaciones de este modelo lineal mixto.

En este artículo se propone una prueba para evaluar si los residuales de un modelo lineal mixto pueden pertenecer a una clase de distribuciones sesgadas, en especial, la normal o t sesgadas, posterior a la estimación de un tipo de modelo lineal mixto con componente de simetría compuesta y con intercepto aleatorio.

## 2. MODELO LINEAL MIXTO

La forma general del modelo lineal mixto está dada por:

$$\underset{(T\times 1)}{y} = \underset{(T\times p)}{X}\ \underset{(p\times 1)}{\beta} + \underset{(T\times Nr)}{Z}\ \underset{(Nr\times 1)}{b} + \underset{(T\times 1)}{\varepsilon}.$$

El vector $y$ tiene como componentes a $y_{ij}$ con $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, donde $n_i$ es la cantidad de medidas por sujeto y $N$ es el total de sujetos (Valencia, 2010).

Además:
$p$ es el número de covariables de la matriz de diseño $X$ más 1, $r$ es el número de covariables de la matriz de diseño $Z$ más 1, de forma que hay $r$ variables asociadas al componente aleatorio, incluyendo el intercepto.

- $X$ es la matriz de covariables de diseño que relaciona el vector de parámetros $\beta$ con la variable respuesta $y$.
- $Z$ es la matriz de diseño de los efectos aleatorios que relaciona al vector $b$ con la variable respuesta $y$, más precisamente, $Z = diag(Z_1, Z_2, \ldots, Z_N)$ y $b = (b_1', b_2', \ldots, b_N')'$ , $dim(b_j) = r\times 1$, con $j = 1, \ldots, r$.
  $\varepsilon \sim N(0, R)$
  $b \sim N(0, B)$.
- Los $b_j$ son independientes entre sí e independientes de los $\varepsilon_i$.
- $R = \sigma^2 I$ matriz diagonal, varianza del error.
- $B$ matriz diagonal, varianza del efecto aleatorio.

El objetivo al implementar un análisis basado en modelos lineales mixtos es la estimación de los $\beta$'s y la predicción de los valores de los efectos $b$'s, además de estimar las componentes de varianza (Lange & Ryan, 1989).

### 2.1. Varianza en el modelo lineal mixto, considerando la simetría compuesta

En esta sección se presentará la descomposición de la varianza total del modelo, describiendo la parte relativa al efecto aleatorio y el error del modelo. Cuando se asume que el componente de varianza es el de simetría compuesta, la composición de la varianza está especificada como:

$$\begin{aligned} \sigma_b^2 &: \text{Varianza entre medidas del mismo sujeto} \\ \sigma_s^2 &: \text{Covarianza entre sujetos} \\ \sigma^2 = \sigma_b^2 + \sigma_s^2 &: \text{Componente de covarianza} \end{aligned}$$

Con estas componentes se formula el coeficiente de correlación intraclase (ICC: Intraclass Correlation Coefficient):

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_s^2}. \tag{1}$$

El ICC representa la correlación entre dos medidas repetidas para el mismo sujeto. En este trabajo se considera un modelo de intercepto aleatorio. Al respecto, Verbeke & Molenberghs (2001), citan lo siguiente:

"The corresponding covariance matrix, with constant variance and constant correlation, is often called compound symmetry".

### 2.2. Familia de distribuciones asimétricas

La forma de la función densidad de las distribuciones asimétricas parte de la siguiente definición, formulada en principio por Azzalini (1985):

$$f(x;\lambda) = 2g(x)G(\lambda x), \qquad \lambda \in R. \tag{2}$$

Donde $g$ es una función densidad de una variable aleatoria continua $d$-dimensional con simetría centrada en 0, y $G$ es una función de distribución escalar, llevando a una distribución sesgada donde $\lambda$ es el parámetro que controla el sesgo o la asimetría, es decir, asume el papel de parámetro de forma. Esto conduce a la definición de distribuciones normales sesgadas (skew-Normal, SN) y distribuciones $t$ asimétricas (skew-T, ST), con algunas características, muy propias de estas familias. Es de anotar que tanto la normal multivariada como la distribución $t$ pertenecen a estas familias (Azzalini & Dalla Valle, 1996).

## 3. PRUEBA PROPUESTA PARA AJUSTE

El procedimiento para la prueba se describe con relación a la prueba propuesta en Valencia (2010); y fue desarrollado en R (R Core Team, 2014), usando funciones de las distribuciones SN y ST.

1. Estimar p: proporción acumulada de cada uno de los cuantiles reales de la variable de interés.

2. Obtener el cuantil teórico a partir de dichas proporciones acumuladas:

$$Qt = F^{-1}(p) \tag{3}$$

3. Transformar el cuantil con la resta de cuantiles reales $Qr$ y teóricos $Qt$:

$$Q = Qr - Qt \tag{4}$$

   $Q$ son los cuantiles a graficar de los datos, en una escala diferente a la original.

4. Obtener bandas de confianza bajo las dos distribuciones que serán comparadas. Sean $Li$: Límite inferior, $Ls$: Límite superior.

   **Corolario 1.** *Sea* $0 < p < 1$. *Si* $F$ *tiene una función densidad* $f$ *en una vecindad de* $\xi_p$ *y* $f$ *es positivo y continuo para* $\xi_p$, *entonces:*

   $$\widehat{\xi_{pn}} \sim AN\left(\xi_p, \frac{p(1-p)}{nf^2(\xi_p)}\right) \tag{5}$$

   Es decir, su estimación se distribuye asintóticamente normal con dicha media y varianza (Serfling, 1980). Este resultado es importante para la estimación de la desviación estándar del cuantil, en la construcción de las bandas de confianza de la prueba.

5. Calcular la proporción de puntos que caen por fuera de las bandas de la distribución bajo la hipótesis.

6. Si la proporción de puntos es inferior a una cota establecida por el investigador, por ejemplo el 10 %, el ajuste a la distribución es adecuado a la distribución hipotética.

## 4. ESTUDIO DE SIMULACIÓN

El comportamiento de la distribución Normal sesgada depende en gran medida del parámetro de sesgo, $\lambda$, valor que determina la forma de la distribución, pero no constituye el mismo coeficiente de asimetría muestral. En la figura 1, puede apreciarse la densidad de dicha distribución para diferentes valores del parámetro $\lambda$; fijando aquí la media en 0 y la desviación en 1. A medida que el valor de $\lambda$ se acerca a cero, se asemeja más a la distribución normal.

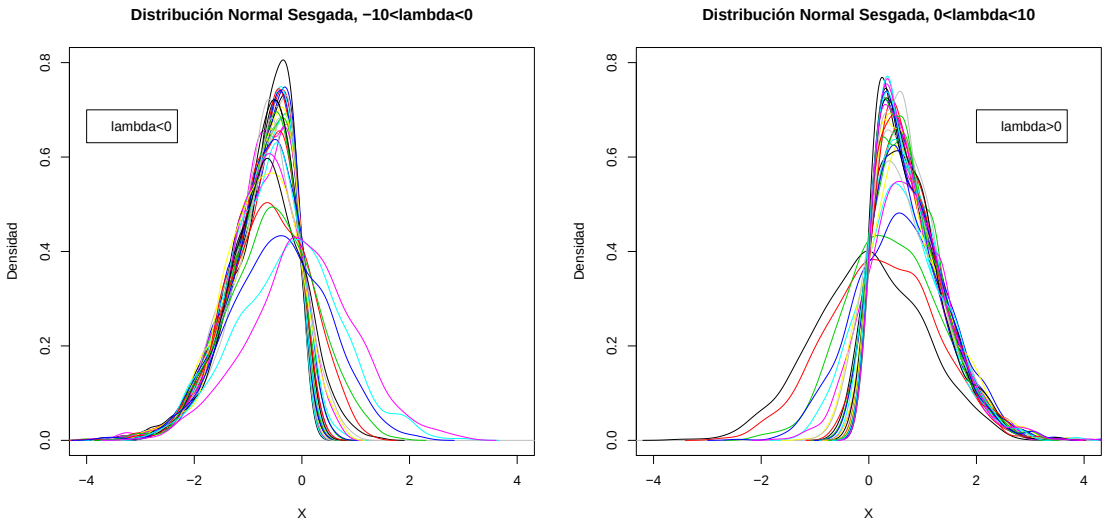


Figura 1: Distribución normal sesgada, para diferentes valores del parámetro de sesgo $\lambda$.

A mayor valor del parámetro $\lambda$, las desviaciones de los valores de $x$ son mayores, en relación a su media, 0. La distribución $t$ sesgada tiene un comportamiento muy similar a medida que aumenta el sesgo fijado.

## Simulaciones de la prueba

A continuación se describen los escenarios bajo los cuales se llevaron a cabo las simulaciones de la prueba propuesta. En un primer caso, fijando la distribución normal sesgada, con valor de sesgo $\lambda = 10$ para el efecto aleatorio y la distribución normal para el error, con $n = 100$ sujetos, cuyo vector de parámetros fijos teórico: $\beta = (50, 50)$ con varianza del efecto aleatorio 0.5 y una correlación intraclase de 0.05, no hay evidencias para rechazar la normalidad en todas las variables del efecto y el error estimados, además, se encuentra que la proporción de puntos por fuera de las bandas normales (0.023) es menor que la de los puntos por fuera de las bandas normales sesgadas (0.1467), con $p < 0{,}001$. La figura 2 (a la izquierda) muestra las bandas estimadas con la distribución normal sesgada para el efecto aleatorio del caso descrito, obtenido con el estimador de Bayes empírico (Verbeke & Lesaffre, 1996). Son pocos los puntos por fuera de las bandas, lo cual puede indicar buen ajuste a la distribución normal sesgada.

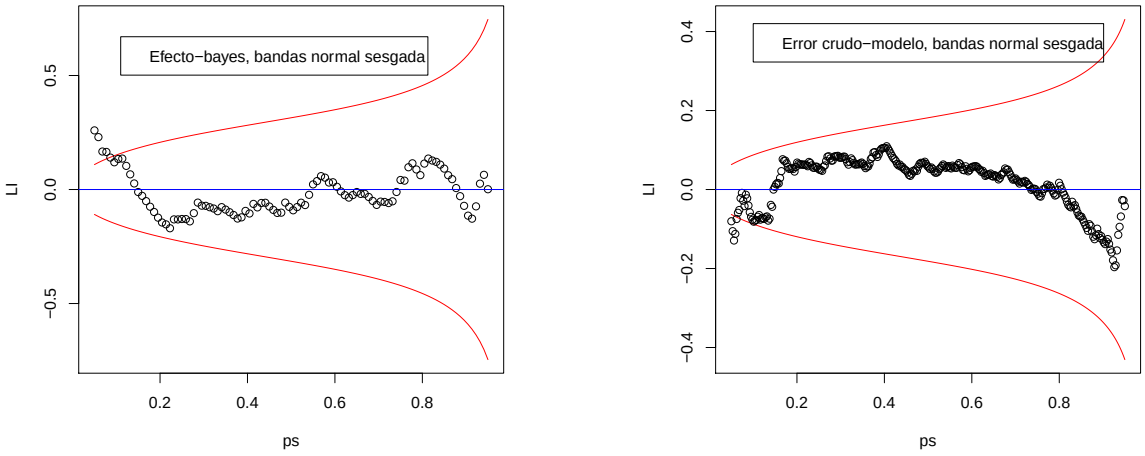


Figura 2: Gráfico de probabilidad con bandas de la normal sesgada para el efecto aleatorio (izquierda) y para el error crudo (derecha).

La figura 2 (a la derecha) muestra las bandas estimadas con la distribución normal sesgada para los cuantiles del error crudo estimado por el estimador de Bayes empírico, en la cual tampoco son muchos los puntos por fuera de éstas.

La figura 3 (a la izquierda) muestra las bandas estimadas con la distribución normal para el error

total estimado, la cual muestra también una cobertura aparentemente buena de los puntos; a la derecha, muestra las bandas de los puntos para el error total estimado, cuando las bandas graficadas provienen de la distribución normal sesgada con parámetro de sesgo 10.

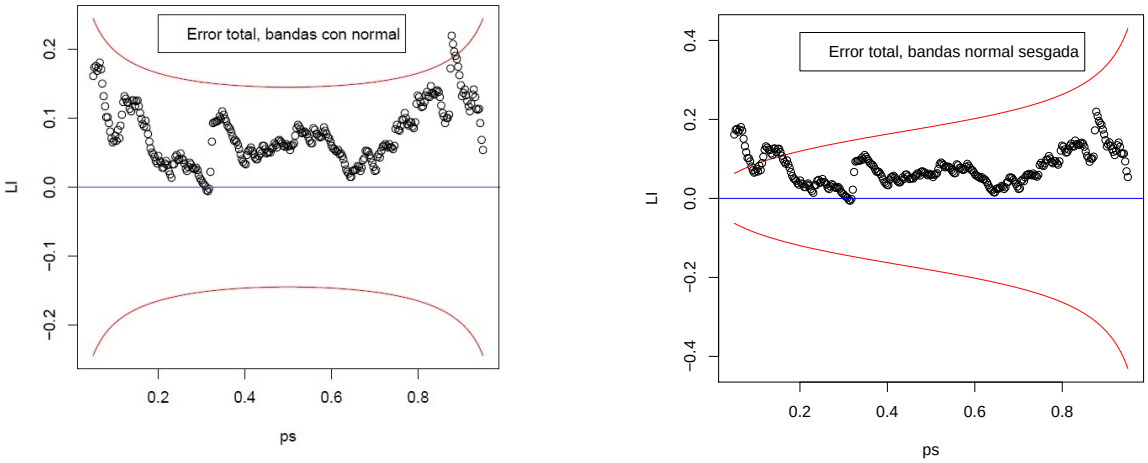


Figura 3: Gráfico de probabilidad con bandas de la normal para el error total (izquierda). Bandas de la normal sesgada para el error total (derecha).

De la figura 3, se aprecia a la izquierda una menor proporción de puntos por fuera de las bandas bajo la distribución normal, que los que están por fuera de las bandas bajo la distribución normal sesgada. Cuando este mismo escenario se repite 3000 veces, los resultados numéricos muestran la proporción de veces en que los puntos que caen por fuera de las bandas normales es menor que la proporción de puntos fuera de las bandas provenientes de la distribución normal sesgada (Tabla 1).

Tabla 1: Prueba de diferencia de proporciones con $\lambda = 10$.

| dist.b$_i$ | dist.e | $\rho_i$ | $\sigma_e^2$ | $\sigma_b^2$ | $\beta_o$ (sd) | sesgob0 | $P1$ | Perm | Perb | Pfm | Pfb | PropSN | PropN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.05 | 9.5 | 0.5 | 49.66 (0.03) | 0.01 | 0.9 | 0.9 | 0.95 | 0.94 | 0.94 | 0.02 | 0.82 |
| SN | Normal | 0.10 | 9.0 | 1.0 | 49.66 (0.03) | 0.01 | 0.9 | 1 | 0.95 | 0.96 | 0.96 | 0.02 | 0.81 |
| | | 0.75 | 2.5 | 7.5 | 49.68 (0.02) | 0.01 | 0.2 | 0.3 | 0.58 | 0.36 | 0.36 | 0.59 | 0.06 |

En la tabla 1, la columna correspondiente a $P1$, es la frecuencia de veces en que se encuentran insuficientes evidencias para rechazar normalidad para el error total, Perm y Perb, frecuencias para el error crudo estimado con el modelo y cuando se usa el estimador de Bayes para el efecto aleatorio, Pfm y Pfb frecuencias para el efecto estimado con el modelo y con Bayes y las dos últimas columnas son la proporción de casos en que hay más puntos dentro de las bandas de confianza normal sesgada: PropSN : PSN < PN, esto es, más puntos por fuera de las bandas normales (no normalidad), y en el segundo, más puntos dentro de las bandas de confianza normales: PropN : PS N > PN (no se encuentran evidencias para el rechazo de normalidad). En la primera línea, cuando el valor de la

varianza del efecto aleatorio es 0.5 ó 1, la proporción de veces en que se encuentran insuficientes evidencias para rechazar normalidad, obtenida con Shapiro Wilks, es adecuada ($P1$ a Pfb) y los valores de PropSN muestran que en un 82 % y 81 % de los casos hay más puntos contenidos en las bandas normales, lo que coincide con el resultado de la prueba de Shapiro Wilks, contrario a lo que ocurre con el resultado de la proporción para rechazar normalidad cuando la varianza del efecto aleatorio es 7.5, donde se detecta mayor la proporción de casos con distribución normal sesgada.

En la Tabla 2 se extienden las simulaciones a otros casos. En sus primeras 3 franjas, se fijan valores para $\sigma_e^2$ :9.5,9 y 8.5 y se mantienen fijos todos los demás parámetros y el escenario ST-SN, variando $\sigma_b^2$ así: 0.5, 1 y 7.5. En las siguientes 3, se fijan de manera inversa, $\sigma_b^2$: 9.5, 9 y 8.5, variando $\sigma_e^2$ así: 0.5, 1 y 7.5.

Tabla 2: Escenarios para la prueba de diferencia de proporciones.

| $\beta_0$ = 50, $n$ = 100, $\lambda$ = 100 | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| dist.b$_i$ | dist.e | $\rho_i$ | $\sigma_e^2$ | $\sigma_b^2$ | $\beta_o$ (sd) | sesgob0 | $P1$ | Perm | Perb | Pfm | Pfb | PropSN | PropN |
| | | 0.06 | 8.5 | 0.5 | 51.2 (0.49) | 0.02 | 0.6 | 0.34 | 0.42 | 0.61 | 0.39 | 0.29 | 0.19 |
| ST | SN | 0.11 | 8.5 | 1 | 51.43 (0.53) | 0.03 | 0.48 | 0.18 | 0.19 | 0.51 | 0.21 | 0.4 | 0.21 |
| | | 0.47 | 8.5 | 7.5 | 54 (0.95) | 0.08 | 0.1 | 0.02 | 0 | 0.05 | 0 | 0.69 | 0.07 |
| | | 0.05 | 9 | 0.5 | 51.11 (0.53) | 0.02 | 0.49 | 0.44 | 0.6 | 0.67 | 0.42 | 0.21 | 0.18 |
| ST | SN | 0.1 | 9 | 1 | 51.52 (0.52) | 0.03 | 0.43 | 0.17 | 0.2 | 0.37 | 0.12 | 0.47 | 0.19 |
| | | 0.45 | 9 | 7.5 | 53.98 (0.72) | 0.08 | 0.09 | 0.02 | 0 | 0.04 | 0 | 0.74 | 0.05 |
| ST | SN | 0.05 | 9.5 | 0.5 | 51.21 (0.63) | 0.02 | 0.55 | 0.3 | 0.5 | 0.59 | 0.37 | 0.24 | 0.21 |
| | | 0.10 | 9.5 | 1.0 | 51.58 (0.55) | 0.03 | 0.40 | 0.16 | 0.22 | 0.33 | 0.16 | 0.42 | 0.16 |
| | | 0.44 | 9.5 | 7.5 | 54.08 (1.02) | 0.08 | 0.13 | 0.02 | 0.01 | 0.05 | 0.01 | 0.75 | 0.09 |
| ST | SN | 0.94 | 0.5 | 8.5 | 54.05 (0.69) | 0.08 | 0.02 | 0.01 | 0.00 | 0.02 | 0.00 | 0.89 | 0.01 |
| | | 0.89 | 1.0 | 8.5 | 54.09 (0.72) | 0.08 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.96 | 0.00 |
| | | 0.53 | 7.5 | 8.5 | 54.21 (0.99) | 0.08 | 0.03 | 0.00 | 0.00 | 0.02 | 0.00 | 0.81 | 0.05 |
| | | 0.95 | 0.5 | 9.0 | 54.14 (0.63) | 0.08 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.91 | 0.01 |
| ST | SN | 0.90 | 1.0 | 9.0 | 54.28 (0.73) | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.94 | 0.00 |
| | | 0.55 | 7.5 | 9.0 | 54.38 (0.92) | 0.09 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.73 | 0.08 |
| | | 0.95 | 0.5 | 9.5 | 54.25 (0.69) | 0.08 | 0.00 | 0.02 | 0.00 | 0.01 | 0.00 | 0.95 | 0.00 |
| ST | SN | 0.90 | 1.0 | 9.5 | 54.42 (0.85) | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.98 | 0.00 |
| | | 0.56 | 7.5 | 9.5 | 54.58 (1.05) | 0.09 | 0.09 | 0.01 | 0.00 | 0.04 | 0.00 | 0.72 | 0.10 |
| $\beta_0$ = 50, $n$ = 100, $\lambda$ = 10 | | | | | | | | | | | | | |
| | | 0.94 | 0.5 | 8.5 | 54.09 (0.82) | 0.08 | 0.012 | 0.014 | 0.00 | 0.02 | 0.00 | 0.91 | 0.01 |
| ST | SN | 0.90 | 1.0 | 9.0 | 54.27 (1.13) | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.97 | 0.00 |
| | | 0.56 | 7.5 | 9.5 | 54.50 (1.17) | 0.09 | 0.08 | 0.00 | 0.00 | 0.02 | 0.00 | 0.73 | 0.11 |

En los resultados (tabla 2) se detectan proporciones altas de casos donde el error se distribuye normal sesgado (PropSN es mayor), que parece que aumentan a medida que la correlación intraclase fijada es mayor.

Situación similar ocurre en la última franja, dejando ver una clara tendencia de la mezcla de variables con estas distribuciones hacia una distribución normal sesgada. Lo anterior indica aciertos en la detección de normalidad o de normalidad sesgada con la prueba propuesta, generando una alternativa para mostrar el comportamiento del residual total del modelo mixto estudiado, pero

también puede extenderse en cada uno de sus componentes.

La prueba tiene un potencial para identificar una presencia de sesgo, tal como indica una simulación presentada en la tabla 3.

Tabla 3: Proporción de veces que detecta la distribución sesgada respectiva

| $\lambda$ | t sesgada | Normal sesgada |
|---|---|---|
| 0-0.1 | 100 | 100 |
| 1 | 100 | 100 |
| 1.5 | 98.95 | 88.99 |
| 2 | 91.99 | 50.01 |

Fuente: Elaboración propia

En esta tabla se observa que la prueba permite detectar en un alto porcentaje de veces, cuando una variable presenta un comportamiento bajo una de estas dos distribuciones al variar el parámetro de forma $\lambda$. Específicamente, bajo las distribuciones normal y $t$ sesgada con sesgo inferior a 1.5, detecta la distribución en un porcentaje de 98.5 % y 88.99 % de las veces. Cuando se simula bajo la normal sesgada con un sesgo de 2, la detecta en un 50 % de las veces, en cuyos casos, se observa una tendencia a encontrar en éstas, la $t$ sesgada. Sin embargo, este parámetro $\lambda$ es diferente del estadístico que mide la asimetría en un conjunto de datos.

En la siguiente subsección se presenta esta prueba usando datos reales, para los residuales de un modelo lineal mixto.

## 4.1. Aplicación

Se usarán datos presentados en Muñoz (2004), quien realiza un diseño experimental para determinar el comportamiento de unas especies de microalgas donde se extraen diferentes variables respuesta. Una de ellas fue la densidad celular, medida como el promedio del número de células ($x$ 106/ml), acorde con lo presentado por el mismo autor (Muñoz, 2004). En el experimento se tomaron mediciones a 36 muestras durante 5 días, esto es, longitudinalmente, en distintos medios de crecimiento. Uno de sus resultados fue que la densidad se relaciona directamente con la biomasa algal y por ello fue de gran interés su medición.

En este trabajo se estima un modelo lineal mixto con intercepto aleatorio, que considera el efecto que causan las medidas en el tiempo de la densidad celular, mediciones no independientes. La covariable $x_1$ será el tiempo en el crecimiento algal. Los resultados se muestran a continuación.

Estimación de un modelo lineal mixto donde los residuales se comportan de forma no normal.

Al estimar un modelo lineal mixto con 5 medidas repetidas, donde la variable respuesta es densidad celular y la variable explicativa es el día, se encuentra que su valor $P$ = 0, lo cual indica su alta significancia o efecto sobre el crecimiento celular, sin embargo, no se encuentra normalidad para los residuales del modelo estimado, cuyas estimaciones se presentan en la tabla 4.

Tabla 4: Coeficientes fijos del modelo lineal mixto para densidad celular.

| | Value | Std. Error | DF | t-value | p-value |
|---|---|---|---|---|---|
| (Intercept) | -1.856222 | 1.3948201 | 143 | -1.330797 | 0.1854 |
| $X_1$ | 5.390444 | 0.3006689 | 143 | 17.928175 | 0.0000 |

Los parámetros estimados toman valores: $\beta_0$=−1.85, vp=0.185, no significativo, y $\beta_1$ = 5.39 con $vp$=0, significativo al nivel 0.05 (tabla 3). Es importante mostrar las pruebas de ambos parámetros, dado que en este modelo lineal mixto se predice también el comportamiento del efecto aleatorio para la $i$-ésima muestra, esto es, un efecto conjunto: $\beta_0 + b_{0i}$, donde $i = 1, ..., 36$.

Tabla 5: Asimetría para los residuales del modelo de densidad celular

| | |
|---|---|
| Día 1: | 1.077633 |
| Día 2: | 1.061791 |
| Día 3: | 1.370898 |
| Día 4: | 0.998884 |
| Día 5: | 0.880008 |

Una medida de asimetría muestral en los residuales puede visualizarse al medir los indicadores de Fisher, usando la función skewness del paquete *fBasics* de R. Para cada día se encuentran las medidas de asimetría mostrados en la tabla 5. Los valores positivos indican que puede existir una asimetría a derecha para los residuales, evaluados por día. Esta información permite inducir la existencia de un sesgo en la variable, por día, sin embargo, los datos aquí presentados no tienen una relación directa con respecto a los valores del parámetro $\lambda$ en las distribuciones sesgadas; esto sólo refleja una aparente desviación en esta variable.

Se realizó la prueba de Shapiro Wilks para los residuales transformados (normalizados) de este modelo, indicando con una probabilidad valor $p < 0{,}001$, que hay evidencias para rechazar normalidad, al nivel de significancia de 0,05. En este ejemplo, puede apreciarse como el tiempo de crecimiento puede generar un efecto significativo que incorpora sesgo en el comportamiento de los residuales, lo que puede representar un sesgo en la estimación de los parámetros.

A continuación se muestra una aplicación de la prueba de normalidad sesgada propuesta para este caso. Para sus residuales, las bandas muestran como es posible una semejanza del comportamiento de éstos a una distribución normal sesgada con parámetro de sesgo alto ($\lambda$ = 200), donde la cola derecha es la que parece encontrarse más cubierta por las bandas de confianza normal sesgada que

la de la Normal, como se observa en la Figura 4.

La prueba de diferencia de proporciones arroja un intervalo de confianza al 90 % para la diferencia: $(Psn - Pn)$ =(−0.38, −0.015), lo cual indica que la proporción de puntos por fuera de las bandas normales es mayor que la de normales sesgadas.

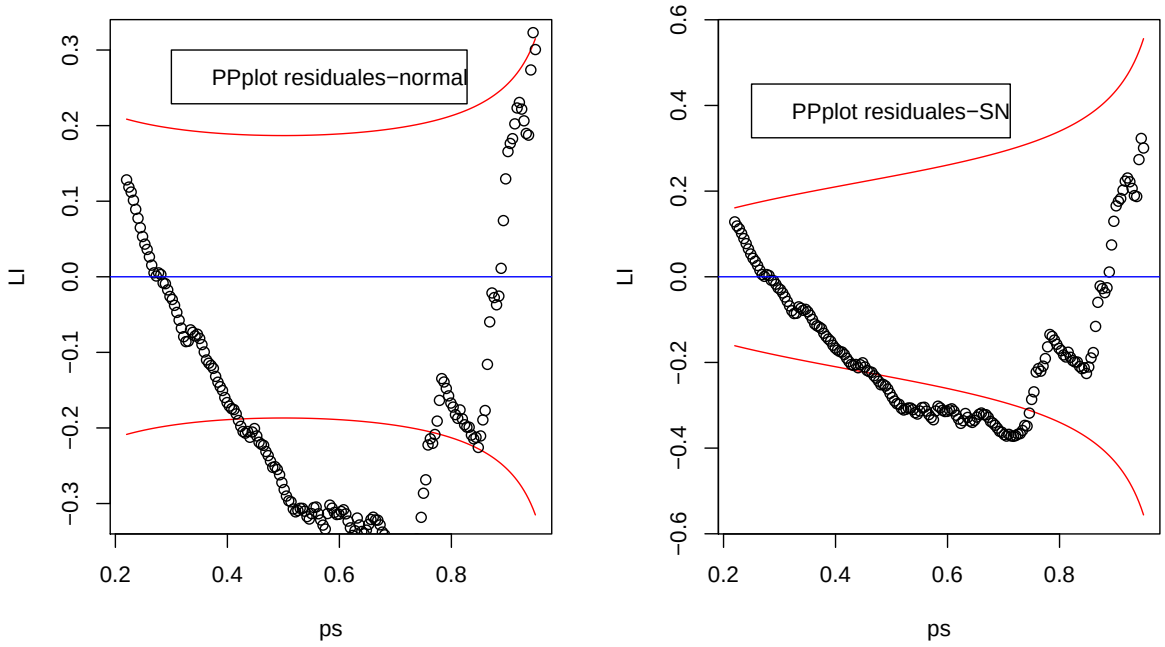


Figura 4: Gráfico de probabilidad normal y Normal sesgada para residuales.

Es claro el comportamiento sesgado aparente en el residual total, sin embargo, no es lo mismo para el efecto aleatorio, pues en éste todos los puntos quedan dentro de las bandas bajo la distribución normal sesgada.

Esto lleva a realizar una estimación del modelo usando una transformación de Box Cox para la respuesta de densidad celular, con un valor del respectivo parámetro lambda de 0.1515, que optimiza la función de verosimilitud. En el modelo cambian un poco los parámetros, pero se encuentran evidencias para no rechazar normalidad en los residuales, con la prueba de Shapiro Wilks, p=0.077.

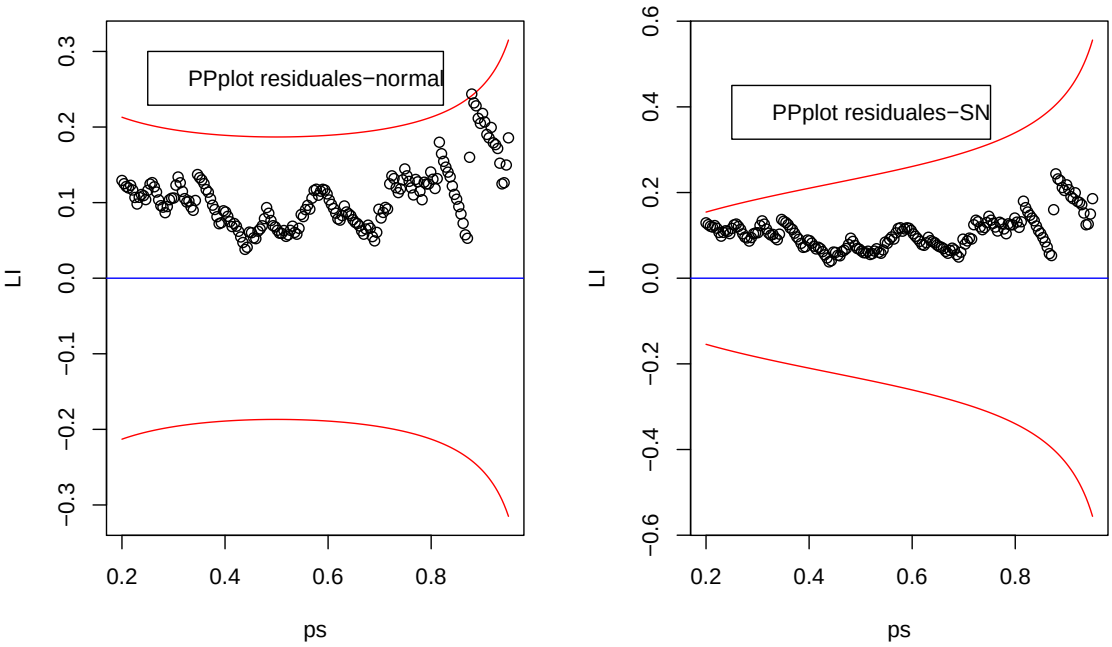


Figura 5: Gráfico de probabilidad normal y normal sesgada para residuales.

En la Figura 5 se observan los residuales para este nuevo modelo, encontrando más puntos dentro de las bandas normales. Al hacer la prueba de diferencia de proporciones para estos residuales, se encuentra igualdad con un nivel de confianza del 95 %: $(-0.02, 0.01)$ el IC incluye el 0, lo cual sugiere que el supuesto de normalidad puede ser viable con esta transformación, ya que las proporciones de puntos por fuera de las bandas son de 0 % para la distribución normal sesgada y 0.006 % para la distribución normal.

La prueba basada en el gráfico de probabilidad indica un problema de sesgo aparente en los residuales del modelo lineal mixto, además de mostrar una mejoría posterior a la transformación utilizada, aspectos que muestran un potencial de la prueba frente a la evaluación de la presencia de sesgos.

## 5. CONCLUSIONES

Se encontraron aciertos en la detección de normalidad sesgada para los residuales con la prueba propuesta, generando una alternativa para mostrar el comportamiento del residual total, pero también puede extenderse en cada uno de sus componentes. En los casos de normalidad sesgada encontrada

para los residuales, ocurre que los efectos aleatorios provenían previamente de una distribución t sesgada, mientras el residual crudo de una distribución normal sesgada, mostrando vía simulación el cambio en la forma de la distribución del residual total del modelo lineal mixto, posterior a su estimación.

Además, también detectó un problema de presencia de sesgo en los residuales del modelo lineal mixto estimado en la aplicación mostrada, mostrando aún más su potencial evaluador del sesgo.

De este trabajo pueden quedar algunos planteamientos sobre modificaciones u otros temas relativos. Una de éstas es la evaluación de una posible presencia de sesgo en los datos por individuo o subgrupo del modelo mixto, que aquí no fueron considerados, pero que quizás pueden representar efectos importantes de un subgrupo a otro. Otra posibilidad es la comparación entre la distribución teórica y de los datos, determinando así la mayor similitud de la variable aleatoria a una específica distribución y no tan ampliamente, como a una familia de distribuciones sesgadas.

**AGRADECIMIENTOS**

A la Facultad de Ciencias de la Universidad Nacional de Colombia, Sede Medellín. Trabajo patrocinado con recursos del proyecto "DIME 20101007954, Convocatoria Nacional 2009, Universidad Nacional de Colombia, Sede Medellín".

## Referencias

Arellano, R. B.; Bolfarine, H.; Lachos, V. (2005), Skew-normal Linear Mixed Models. *Journal of Data Science*, 3, 415–438.

Azzalini, A. (1985), A class of distribution which includes the normal ones. *Scand. J. Statist*, 12, 171–178.

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

Azzalini, A.; Capitanio, A. (1999), Statistical applications of the multivariate skew normal distribution. *J. Roy. Statist. Soc*, series B, 6, 579 – 602.

Gurka M.; Edwards, Ll.; Muller, K.; Kupper, L. (2006), Extending the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society*. Series A, 169(2), 273–288.

Lange, N.; Ryan, L. (1989), Assessing normality in random effects models. *Ann. Statist*, 17(2), 624–642.

Muñoz, H. (2004), Estudio *In Vitro* de los efectos e interacciones ambientales en el crecimiento y la producción de ácidos grasos poliinsaturados de las microalgas marinas. Tesis de Grado. Universidad Nacional de Colombia.

R Core Team (2014), R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Serfling, R. J. (1980), Approximation theorems of mathematical statistics. Wiley, New York. ISBN 0471024031

Valencia, M. (2010), Estimación en modelos lineales mixtos con datos continuos usando transformaciones y distribuciones no normales. Trabajo de Grado para optar al título de Magister en Ciencias-Estadística. Universidad Nacional de Colombia, Sede Medellín.

Verbeke, G.; Lesaffre, E. (1996), A linear mixed effects model with heterogeneity in the random effects population. *J. Amer. Statist. Assoc.*, 91, 217–221.

Verbeke, G.; Molenberghs, G. (2001), *Linear Mixed Models for Longitudinal Data*. Springer, NY.

Zhou, T.; He, X. (2007), Three-step estimation in linear mixed models with skew-t distributions. *Journal of Statistical Planning and Inference*, 138, 1542 – 1555.