

استخدام خوارزمية تعظيم التوقع في تقدير نموذج الانحدار الخطي العام معدما البيانات تتوزع توزيعاً طبيعياً ملتبساً مع التطبيق

امين حسين هادي
باحث

أ.م.د سهاد علي شهيد
الجامعة المستنصرية / كلية الإدارة والاقتصاد

P: ISSN : 1813-6729

<http://doi.org/10.31272/JAE.43.2020.124.13>

E : ISSN : 2707-1359

مقبول للنشر بتاريخ : 2020/4/27

تاريخ استلام البحث : 2020/1/13

المستخلص :

التوزيع الطبيعي الملتوي بمعلمة الالتواء (λ) هو احد التوزيعات الكاوسية لكن بإضافة معلمة الى الشكل تحدد مقدار الالتواء ، و يعد حالة خاصة من التوزيع الطبيعي . ان التباين في البيانات أي وجود بيانات غير متماثلة كان سبب في ظهور التوزيع الطبيعي الملتوي (SND) ، ويعد هذا التوزيع من التوزيعات المهمة نظراً لوجود معلمة الالتواء التي تتغير قيمتها باستمرار لذلك توفر هذه المعلمة نموذجاً جيداً لمحاكاة العديد من الظواهر التي تبرز لديها مثل تلك البيانات .

المحور الرئيس للدراسة هو تسليط الضوء على التوزيع الطبيعي الملتوي (SND) بوجود معلمة الالتواء (λ) لنموذج الانحدار الخطي العام (GLM) (Generalized Linear Regression Model) اذ تم تقدير معلمات الانموذج بشكل عام من خلال استعمال خوارزمية تعظيم التوقع (EMA) (Expectation Maximization Algorithm) التي أعطت مواصفات للمعلمات مرغوبة واكثر كفاءة . البيانات التي تم دراستها هي بيانات مالية تم الحصول عليها من سوق العراق للأوراق المالية والتي تمثل عدد الأسهم المتداولة لمجموعة من القطاعات وتم اخذ البيانات بشكل اسبوعي ولأربع سنوات للفترة من (2016-2019) وبلغ عدد المشاهدات (150) .



مجلة الإدارة والاقتصاد
العدد 124 / حزيران / 2020
الصفحات : 186-200

• مستل من رسالة ماجستير

1. المقدمة

أن مشكلة تقدير المعلمات مثل (σ^2, μ) لبيانات عشوائية كبيرة والتي تأخذ توزيعاً معيناً ، كانت أحد الموضوعات التي تدرس بشكل معمق . أن خوارزمية تعظيم التوقع (EMA) هي إحدى الطرق المعتمدة لإيجاد مقدرات المعلمات التوزيعات التي تتبع توزيعاً غير طبيعي . وخاصة عندما تكون البيانات تحتوي على قيم مفقودة (Missing Value) أو قيم شاذة (Outlier Value) خوارزمية تعظيم التوقع (EMA) قدمت من قبل الباحث (Dempster) وباحثون آخرون وقد تم استخدامها بشكل واسع لإيجاد مقدرات الإمكان الأعظم (MLE) ومن إيجابياته انه يتطلب عدم الحساسية للقيم الابتدائية (الأولية) . وأحد أسباب شيوع استخدام لخوارزمية (EMA) هو أن خطوة (M-step) تتضمن فقط بيانات كاملة لمقدرات (MLE) ، والتي غالباً تكون سهلة في الحساب . وعلى الرغم من أن طريقتي الإمكان الأعظم (MLE) وخوارزمية تعظيم التوقع (EMA) يبحثان عن المعلمات الأفضل في النموذج ، إلا أن عملية إيجاد النماذج مختلفة تماماً .

2- التوزيع الطبيعي الملتوي [2] ، [3] ، [5] :- Skew Normal Distribution

إن بعض الظواهر العملية تتطلب وضع نماذج لبيانات تمتلك توزيعاً محدداً لكنها ملتوية. ولاسيما البيانات المالية والاقتصادية والبيانات البيئية والتي لا تتبع على الأغلب التوزيع الطبيعي حيث يدخل الالتواء (Skewness) بالتوزيع المتماثل بطرق عديدة . منها وجود قيم شاذة (Outlier) في البيانات . إذ قام الباحثين Azzalini عام (1985) باقتراح فئة جديدة من التوزيعات ، والتي تشترك بخصائص مماثلة مع التوزيع الطبيعي وهو حالة خاصة من التوزيع الطبيعي واطلق عليه التوزيع الطبيعي الملتوي (SND) وعندما تكون الأخطاء العشوائية تمثل حد الخطأ الطبيعي الملتوي (Skew Normal Errors) . وفي حالة معلمة الالتواء تساوي صفر يرجع التوزيع الطبيعي الملتوي (SND) الى أساسه متمثل بالتوزيع الطبيعي (Normal Distribution).

1-2 دالة التوزيع الطبيعي الملتوي [12] :-

Function Of The Skew Normal Distribution

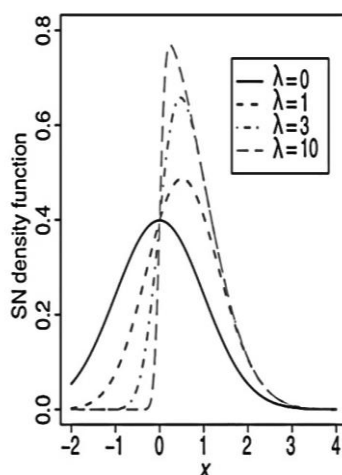
تكون دالة التوزيع الطبيعي الملتوي (في حالة الالتواء الموجب أو من الجهة اليمنى Right (Skew Normal Distribution) (RSND) ، وفي حالة الالتواء السالب أو من الجهة اليسرى Left (Skew Normal Distribution) (LSND) كالآتي :-

$$f(x, \mu, \sigma, \lambda) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2(1-\lambda)^2}} & , x \geq \mu \\ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2(1+\lambda)^2}} & , x < \mu \end{cases}$$

($-\infty < \lambda < +\infty$)

(1-1) ...

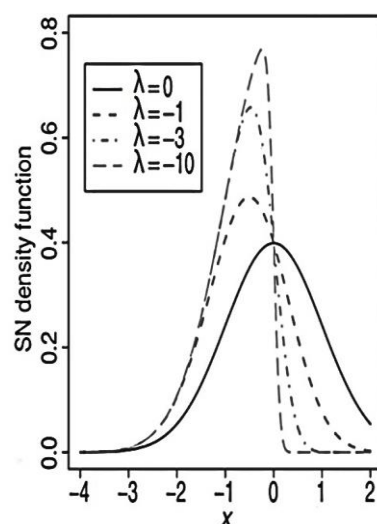
أن التوزيع الطبيعي الملتوي يكون ملتوياً باتجاه موجب أو من الجهة اليمنى عندما ($\lambda > 0$) ، ويكون ملتوياً باتجاه سالب أو من الجهة اليسرى عندما ($\lambda < 0$) . في حالة الالتواء الموجب يكون (الوسط الحسابي < الوسيط < المنوال) ، وفي حالة الالتواء السالب يكون (الوسط الحسابي > الوسيط > المنوال) .



التواء موجب

الشكل (1)

يوضح دالة الكثافة الاحتمالية (p.d.f) في حالة الالتواء الموجب



التواء سالب

الشكل (2)

يوضح دالة الكثافة الاحتمالية (p.d.f) في حالة الالتواء السالب

شكل (1) يوضح دالة الكثافة الاحتمالية (p.d.f) عندما $(\lambda = 0, 1, 3, 10)$ عندما يكون الالتواء موجب .
والشكل (2) يوضح دالة الكثافة الاحتمالية (p.d.f) عندما $(\lambda = 0, -1, -3, -10)$ عندما يكون الالتواء سالباً . وعندما يكون $\lambda = 0$ يرجع شكل التوزيع الى التوزيع الطبيعي الموضح في الشكلين (1) و (2) .

2-2 دالة الكثافة للتوزيع الطبيعي المعياري الملتوي [4]، [13] :-

Function For Standard Skew Normal Distribution

ان دالة الكثافة الاحتمالية (p.d.f) للتوزيع الطبيعي المعياري الملتوي يمكن صياغتها كالآتي :-

$$\phi(x; \lambda) = 2\phi(x)\Phi(\lambda x) \quad (-\infty < \lambda < \infty)$$

(2-1) ...

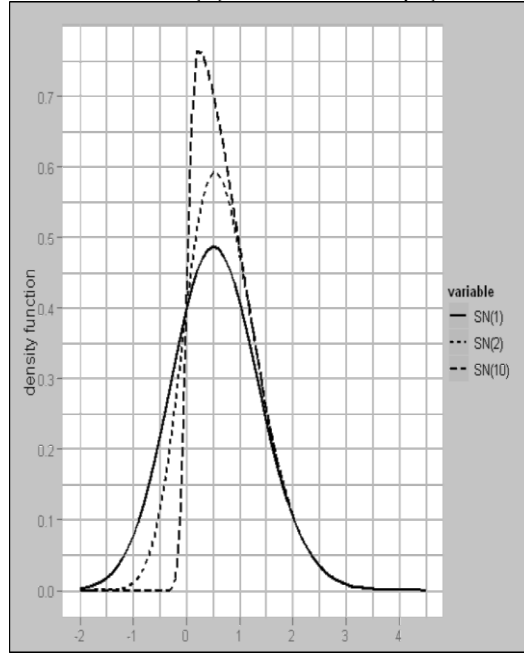
اذ إن :-

x : متغير عشوائي يمتلك توزيعاً طبيعياً ملتوياً (SND) بمعلمة λ وحدودها $(-\infty < \lambda < \infty)$.

$\phi(x)$: دالة التوزيع الاحتمالي المعياري (Standard Normal Distribution)

$\Phi(x)$: دالة التوزيع الاحتمالي التجميعي (c.d.f) للتوزيع المعياري .

λ : وتمثل معلمة الالتواء التي تحدد مقدار الالتواء ، وتنحصر قيمها بين $(-\infty, \infty)$.
والشكل (3) يوضح شكل دالة الكثافة الاحتمالية للتوزيع الطبيعي المعياري الملتوي (p.d.f) وعندما $X \sim SND(0,1,\lambda)$ و $(\lambda = 1,2,10)$ (اذ كلما قلت قيمة (λ) اقترب شكل التوزيع من التوزيع الطبيعي).



الشكل (3)

دالة الكثافة الاحتمالية للتوزيع الطبيعي المعياري الملتوي

يلاحظ في الشكل (3) ان دالة (p.d.f) عندما قيم $(\lambda = 1,2,10)$ والتي تمثل معلمة الالتواء اذ يصبح الشكل ملتوياً بشكل متزايد إلى اليمين بزيادة قيمة (λ) . وعندما تكون قيمة $(\lambda = 1)$ ، يصبح الشكل ملتوياً قليلاً إلى اليمين ، وعندما $(\lambda = 10)$ يصبح الشكل ملتوياً أكثر لليمين أي ان البيانات تكون ذات التواء اعلى على الجهة اليمنى من المنحني . عندها يكون الوسط الحسابي اكبر من الوسيط واكبر من المنوال .

3-2 خواص التوزيع الطبيعي الملتوي [14] ، [16] :

Proprieties Of Skew Normal Distribution

ليكن المتغير العشوائي $X \sim SND(\mu, \sigma^2, \lambda)$ فإن (X) يمتلك الخصائص التالية :-
1- دالة الكثافة الاحتمالية لـ (X) هي:-

$$f(x, \mu, \sigma^2, \lambda) = 2\phi(X; \mu, \sigma^2) \Phi\left(\frac{\lambda(x - \mu)}{\sigma}\right), \quad x \in R$$

2- يزداد التواء التوزيع عند زيادة القيمة المطلقة لـ (λ) سواء كان من جهة اليمين او اليسار.

3- إذا كان $(X \sim SND(\lambda))$ فإن $(-X \sim SND(-\lambda))$.

4- $1 - F(-x, \lambda) = F(x, -\lambda)$.

5- $F(x, 1) = \Phi^2(x)$.

6- الوسط الحسابي (Mean) لـ (X) :-

$$E(X) = \mu + c\sigma\zeta$$

اذ ان :-

$$c = \sqrt{\frac{2}{\pi}} \quad , \quad \zeta = \frac{\lambda}{\sqrt{1+\lambda^2}}$$

وعندما $X \sim SND(0,1,\lambda)$ يكون المتوسط :-

$$E(X) = \sqrt{\frac{2}{\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}}$$

7- التباين Variance لـ (X) :-

$$Var(X) = \sigma^2(1 - c^2\zeta^2)$$

وعندما $X \sim SND(0,1,\lambda)$ يكون التباين :-

$$Var(X) = 1 - \left(\sqrt{\frac{2}{\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}} \right)^2$$

8- يقاس معامل الالتواء (Skewness) للمتغير العشوائي (X) بواسطة $\gamma_1(X)$ وقيمتها تكون محصورة بين $(-\infty, \infty)$ ويحسب وفق الاتي :

$$\begin{aligned} \gamma_1(X) &= \frac{4 - \pi}{2} \frac{(E[X])^3}{(Var[X])^{3/2}} \\ &= \frac{4 - \pi}{2} \frac{\left(\sqrt{\frac{2}{\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}} \right)^3}{\left(1 - \left(\sqrt{\frac{2}{\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}} \right)^2 \right)^{\frac{3}{2}}} \end{aligned}$$

9- يقاس معامل التفرطح (Kurtosis) للمتغير العشوائي (X) بواسطة $\gamma_2(X)$ وقيمتها تكون محصورة بين $(0, +\infty)$ ويحسب وفق الاتي :-

$$\begin{aligned} \gamma_2(X) &= 2(\pi - 3) \frac{E[X]^4}{Var[X]^2} \\ &= 2(\pi - 3) \frac{\left(\sqrt{\frac{2}{\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}} \right)^4}{\left(1 - \left(\sqrt{\frac{2}{\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}} \right)^2 \right)^2} \end{aligned}$$

3- الانموذج الخطي تحت افتراض التوزيع الطبيعي الملتوي [7] ، [9] ، [10] :-

Linear Model With Skew Normal Distribution (LMSND)

عندما يتوزع المتغير المعتمد توزيعاً طبيعياً ملتوياً فإن حد الخطأ العشوائي يتوزع أيضاً التوزيع نفسه بسبب العلاقة الخطية بينهما . عندها يمكن صياغة العلاقة الخطية بأنموذج خطي وبأخذ الشكل الاتي :-

$$y_i = X_i' \beta + f(t_i) + \epsilon_i, \quad i = 1, 2, \dots, n \quad (3-1)$$

إذ إن :-

y_i : يمثل المتغير المعتمد (الاستجابة) لـ i^{th} من الوحدات التجريبية عندما $(i = 1, 2, \dots, n)$.

X_i : يمثل المتغيرات التوضيحية (المفسرة) ذات متجه ابعاده $(p \times 1)$.

β : متجه المعلمات وابعاده $(p \times 1)$ ، علماً بأن العنصر الاول منه يمثل الحد الثابت.

$f(t_i)$: اذ تمثل f دالة التمهيد (Smoothing Function) ، وهي تمثل دالة الجزاء (الغرامة) $f = (f(t_1^0), \dots, f(t_q^0))^T$ وتكون قيم مرتبة وشكلية من (t_i) $(t_1^0, t_2^0, \dots, t_q^0)$ و $(i = 1, 2, \dots, n)$.

ϵ_i : حد الخطأ العشوائي ويتوزع توزيعاً طبيعياً ملتوياً (SND)، $(\epsilon_i \sim SND(\mu, \sigma^2, \lambda))$ ، عندما $(i = 1, 2, \dots, n)$ ، وهو المركبة الأساسية لنموذج (LMSND) .

الانموذج من معادلة (3-1) يمكن إعادة صياغته وفق صيغة المصفوفات وكالاتي :-

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{Nf} + \underline{\epsilon} \quad (4-1)$$

اذ ان :-

N : مصفوفة الأثر (Incidence Matrix) ابعادها $(n \times p)$ بوجود الأزواج (i, j) وهي دالة المؤشر (Indicator Function) $(I(t_i = t_j^0))$ عندما $(j = 1, 2, \dots, q)$ و $(p = q)$.

كما اشرنا سابقاً في معادلة (3-1) و (4-1) ان الانموذج الخطي العام عندما تتوزع البيانات توزيعاً طبيعياً ملتوياً (SND) يتطلب ذلك إضافة دالة التمهيد $f(.)$ بهدف تحسين النتائج وجعل المقدرات أكثر كفاءة. سيتم اعتماد $f(.)$ على انها دالة الجزاء (Penalty Function) اذ انها دالة معلمية ، الهدف هو تحويل المشكلات المقيدة الى مشاكل غير مقيدة عن طريق ادخال جزاء مصطنع لانتهاك تلك القيود .

4- تقدير أنموذج الانحدار الخطي العام بوجود التوزيع الطبيعي الملتوي باستخدام خوارزمية تعظيم التوقع [6] ، [7] ، [8] ، [11] ، [15] :

Estimate The General Linear Regression Model In The Presence Of The Skew Normal Distribution By using The Expectation Maximization Algorithm (EMA)

ان خوارزمية تعظيم التوقع (EMA) تأخذ القيم التخمينية للمعلمات أولاً وبعدها تحتسب القيم المفقودة او الشاذة) ومن ثم يعدل الانموذج ليناسب التخمينات والبيانات المشاهدة .

ان تقدير متجه المعلمات (θ) في الانموذج لبيانات التوزيع الطبيعي الملتوي (SND) باستخدام طريقة الإمكان الأعظم (MLE) بشكل مباشر يؤدي الى تكلفة حسابية عالية ، فضلاً عن فقدان مواصفات تلك المقدرات المرغوب في تحقيقها . لذا يفضل استخدام خوارزمية تعظيم التوقع (EMA) والتي تساهم بأختزال الجهد الحسابي مع اكتساب الكفاءة للمقدرات .

أن تقدير المعلمات باستخدام خوارزمية تعظيم التوقع (EM) يمكن أن تلخص كالآتي :-
ومن معادلة (3-1)

$$y_i = X_i' \beta + f(t_i) + \epsilon_i , \quad i = 1, 2, \dots, n$$

ان أنموذج الانحدار الخطي العام لبيانات الالتواء الطبيعي (GLMSND) تشير الى ان $(y_i \sim \text{SND}(\mu_i, \sigma^2, \lambda))$ والذي ينتج عنه :-

$$\mu_i = x_i' \beta + n_i' f \quad \dots (5-1)$$

إذ إن :-

n_i' : يمثل الصف ith لـ N عندما $(i = 1, 2, \dots, n)$.

وعند تعظيم دالة الإمكان بدلالة معلمات الأنموذج $(\beta', \sigma^2, \lambda, f')$.

$$\begin{aligned} I(\theta) = \ell(\theta) = & -\frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2 \\ & + \sum_{i=1}^n \ln \left(\Phi \left(\frac{\lambda(y_i - \mu_i)}{\sigma} \right) \right) \quad \dots (6-1) \end{aligned}$$

الاجراء يستند على فكرة اللوغاريتم الإمكان الأعظم الجزائية (Log-likelihood penalization) والذي يتضمن دمج دالة الجزاء (Penalty function) بدالة لوغاريتم الإمكان الأعظم . ومعلمة الجزاء تحدد قيمتها من معيار معلومات بيز (BIC) والتي تلعب دور مهم في تحديد واختيار المتغير المعنوي وتحسب كالآتي :-

$$BIC(\alpha) = \ln[(n)[2 + p + df(\alpha)] - 2\ell_p(\theta, \alpha)]$$

$$\ell_p(\theta, \alpha) = \ell(\theta) - \frac{\alpha}{2} J(f) \quad \dots (7-1)$$

إذ إن :-

$J(f)$: دالة الجزاء والتي تمثل دالة التوزيع الطبيعي الملتوي (SND) . وتحسب كالآتي :

$$J(f) = \int_a^b [f'(t)]^2 dt$$

(α) : هي معلمة التمهيد (الجزاء) (Smoothing parameter) . والتي تراقب الأفضلية بين حسن المطابقة ودالة التمهيد المقدرة .

ليكن $(TN(\mu, \sigma^2; 0, +\infty))$ تشير الى التوزيع الطبيعي المشذب (أي مقطوع القيم السالبة) (Truncated Normal Distribution) بمتوسط (μ) وتباين (σ^2) ضمن الفترة $(0, +\infty)$.
وكما هو معلوم ان :-

$$\Phi\left(\frac{\lambda(y-\mu)}{\sigma}\right) = \int_0^{+\infty} \phi(z; \lambda(y-\mu), \sigma^2) \dots (8-1)$$

لذلك سيتم إعادة صياغة معادلة (8-1) وكالاتي :-

$$f(y; \mu, \sigma^2, \lambda) = 2\phi(y; \mu, \sigma^2) \int_0^{+\infty} \phi(z; \lambda(y-\mu), \sigma^2) dz \dots (9-1)$$

وكذلك فإن :-

$$f(y, z; \mu, \sigma^2, \lambda) = 2\phi(y; \mu, \sigma^2) \phi(z; \lambda(y-\mu), \sigma^2) \dots (10-1)$$

وبما أن :-

$$f(z|y) = \frac{f(y, z)}{f(y)}$$

فإن :-

$$f(z|y) = \frac{\phi(z; \lambda(y-\mu), \sigma^2)}{\int_0^{+\infty} \phi(z; \lambda(y-\mu), \sigma^2) dz}, z > 0 \dots (11-1)$$

اذ ان المتغير العشوائي (Z|Y) يتوزع

$$Z|Y = y \sim TN(\lambda(y-\mu), \sigma^2; 0, +\infty)$$

ومن خلال خصائص التوزيع الطبيعي المشذب (Truncated Normal Distribution) ان :-

$$E[Z|Y = y] = \lambda(y-\mu) + \sigma W_\Phi(\lambda(y-\mu)/\sigma) \dots (12-1)$$

$$E[Z^2|Y = y] = \lambda^2(y-\mu)^2 + \sigma^2 + \sigma(\lambda(y-\mu))W_\Phi(\lambda(y-\mu)/\sigma) \dots (13-1)$$

إذ ان :-

$$W_\Phi(u) = \phi(u)/\Phi(u) \dots (14-1)$$

ليكن $y = (y_1, y_2, \dots, y_n)'$ و $z = (z_1, z_2, \dots, z_n)'$ وعلى اعتبار فرض (z) بيانات مفقودة . بالتالي دالة الإمكان الأعظم الكاملة بوجود $y_c = (y', z')'$ تصاغ بالشكل الاتي :-

$$\ell_c(\theta|y_c) = C - n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n z_i^2 + \frac{\lambda}{\sigma^2} \sum_{i=1}^n z_i(y_i - \mu_i) - \frac{1 + \lambda^2}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2 \dots (15-1)$$

إذ ان C ثابتاً لا يعتمد على المعلمات غير المعلومة .

اذ ان خوارزمية تعظيم التوقع يمكن تلخيصها بأربع خطوات كالآتي :-

- المرحلة الأولى : التهيئة Initialization

الخوارزمية تبدأ بوجود بعض القيم الابتدائية للمعلمات $\theta_i^{(k)}$ حيث ان (k^{th}) هو تكرار عند (i^{th}) من التوزيعات المركبة الاسية (التوزيع الطبيعي الملتوي) على سبيل المثال $(\lambda_i^{(0)}, \mu_i^{(0)}, \sigma_i^{(0)})$ يتم اختيارها لـ (j^{th}) من العينات والمقاسة وفق مركبات التوزيع الطبيعي الملتوي (i^{th}) .

ويمكن الحصول على مجموعة قيم الابتدائية للمعاملات بحساب $\hat{\beta}^{(0)}$ و $\hat{\sigma}^{2(0)}$ من تقدير النموذج الخطي بأحد طرق التقدير المعروفة مثل المربعات الصغرى لـ Y على x وكذلك يمكن حساب :-

$$\hat{f}^{(0)} = (N'N - \alpha\sigma^2k)^{-1}N'(y - X\hat{\beta}^{(0)})$$

اما قيمة $\hat{\lambda}^{(0)}$ يمكن الحصول عليها من معامل التواء العينة (The sample skewness coeff) وكما يلي :-

$$\hat{\lambda}^{(0)} = y - x\hat{\beta}^{(0)} - N\hat{f}^{(0)}$$

دالة الإمكان الأعظم الابتدائية تحسب وفق المعادلة الآتية :-

$$\begin{aligned} \ell_c(\theta|y_c)^{(0)} &= C - n \log \sigma^{2(0)} \\ &- \frac{1}{2\sigma^{2(0)}} \sum_{i=1}^n z_i^2 + \frac{\lambda^{(0)}}{\sigma^{2(0)}} \sum_{i=1}^n z_i (y_i - \mu_i^{(0)}) \\ &- \frac{1 + \lambda^{2(0)}}{2\sigma^{2(0)}} \sum_{i=1}^n (y_i - \mu_i^{(0)})^2 \dots (16-1) \end{aligned}$$

- المرحلة الثانية: E-step

حساب القيمة التخمينية لتكرار (k^{th}) لاحتمال ان (j^{th}) من العينات تنتمي (i^{th}) من مركبات التوزيع الطبيعي الملتوي . وبوجود التقديرات الأولية.

$$\begin{aligned} \hat{\theta}^{(k)} &= (\hat{\beta}^{(k)}, \hat{\sigma}^{2(k)}, \hat{\lambda}^{(k)}, \hat{f}^{(k)}) \\ , k &= 0, 1, 2, \dots; k \in R \end{aligned}$$

بعدها سنجد الدالة :-

$$\begin{aligned} Q(\theta|\hat{\theta}) &= E[\ell_c(\theta|y_c)|y, \hat{\theta}^{(k)}] \\ &= -n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n \widehat{z}_i^{2(k)} \\ &+ \frac{\lambda}{\sigma^2} \sum_{i=1}^n \widehat{z}_i^{(k)} (y_i - \mu_i) \\ &- \frac{1 + \lambda^2}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2 \dots (17-1) \end{aligned}$$

اذ ان :-

$$\begin{aligned} \hat{z}_i^{(k)} &= E[Z_i|y_i, \hat{\theta}^{(k)}] \\ \widehat{z}_i^{2(k)} &= E[Z_i^2|y_i, \hat{\theta}^{(k)}] \end{aligned}$$

وان :-

$$Z_i|Y_i = y_i \sim TN(\lambda(y_i - \mu_i), \sigma^2; 0, +\infty)$$

و قيم $\hat{z}_i^{(k)}$ و $\widehat{z}_i^{2(k)}$ تحسب وفق الصيغ الآتية :-

$$\hat{z}_i^{(k)} = \hat{\lambda}^{(k)} \hat{\eta}_i^{(k)} + \hat{\sigma}^{(k)} W_{\Phi} \left(\frac{\hat{\lambda}^{(k)} \hat{\eta}_i^{(k)}}{\hat{\sigma}^{(k)}} \right)$$

... (18-1)

$$\begin{aligned} \widehat{z}_i^{2(k)} &= [\hat{\lambda}^{(k)} \hat{\eta}_i^{(k)}]^2 + \hat{\sigma}^{2(k)} \\ &+ \hat{\lambda}^{(k)} \hat{\eta}_i^{(k)} W_{\Phi} \left(\frac{\hat{\lambda}^{(k)} \hat{\eta}_i^{(k)}}{\hat{\sigma}^{(k)}} \right) \dots (19-1) \end{aligned}$$

اذ ان :-

$$\hat{\eta}_i^{(k)} = y_i - x_i' \hat{\beta}^{(k)} - n_i' \hat{f}^{(k)}$$

$$, i = 1, 2, \dots, n$$

اذ ان :

$\hat{\eta}$:- يمثل حد الخطأ العشوائي .

إن قيم θ التي تعظم دالة الإمكان تحسب كالآتي :-

$$Q_p(\theta|\hat{\theta}) = Q(\theta|\hat{\theta}) - \frac{\alpha}{2} J(f) \quad \dots (20-1)$$

ويمكن تعريف دالة الجزاء (Penalty function)

$$J(f) = \int_a^b [f''(t)]^2 dt \quad \dots (21-1)$$

إذ أن $f''(t)$ هي المشتقة الثانية بالنسبة لـ $f(t)$ بحدود $[a, b]$

حيث أن الخطوة E- تحتسب $\hat{z}_i^{(k)}$ و $\widehat{z}_i^{2(k)}$ عندما $(i = 1, 2, \dots, n)$ باستخدام المعادلتين (18-1) و (19-1).

بعدها الخطوة M لتعظيم $Q_p(\theta|\hat{\theta})$ بوجود (θ) ، وللحصول على تقديرات جديدة $\hat{\theta}^{(k+1)}$ يكون كالآتي

- المرحلة الثالثة: M-step

التقدير الجديد يتم احتسابه وفق الآتي :-

نحْدث $\hat{\beta}^{(k)}, \hat{f}^{(k)}, \widehat{\sigma}^2^{(k)}$ و $\hat{\lambda}^{(k)}$ وكالآتي :-

$$\hat{\beta}^{(k+1)} = (X'X)^{-1}X'[y - N\hat{f}^{(k)} - \hat{\lambda}^{(k)}$$

$$/ (1 + (\hat{\lambda}^{(k)})^2) \hat{z}_i^{(k)}] \quad \dots (22-1)$$

$$\hat{f}^{(k+1)} = \left(N'N + \frac{\alpha \widehat{\sigma}^2^{(k+1)}}{1 + (\hat{\lambda}^{(k)})^2} K \right)^{-1} N'$$

$$\left(y - X\hat{\beta}^{(k)} - \frac{\hat{\lambda}^{(k)}}{1 + (\hat{\lambda}^{(k)})^2} \hat{z}_i^{(k)} \right) \quad (23-1)$$

$$\widehat{\sigma}^2^{(k+1)} = \{1_n^T \hat{z}^{2(k)} - 2\hat{\lambda}^{(k)} \hat{z}^{(k)}\}$$

$$(y - \hat{\mu}^{(k)}) + (1 + (\hat{\lambda}^{(k)})^2) S\hat{\beta}^{(k)} \hat{f}^{(k)} \} / 2n \quad \dots (24-1)$$

$$\hat{\lambda}^{(k+1)} = \hat{z}^{(k)'} (y - \hat{\mu}^{(k)}) / S(\hat{\beta}^{(k)} \hat{f}^{(k)})$$

$$\dots (25-1)$$

إذ ان :-

1_n : يمثل متجه واحد $(n \times 1)$ 1_n

$$\hat{z}^{(k)} = (\hat{z}_1^{(k)}, \dots, \hat{z}_n^{(k)})'$$

$$\widehat{z}^{2(k)} = (\widehat{z}_1^{2(k)}, \dots, \widehat{z}_n^{2(k)})'$$

$$S(\beta, f) = (y - \mu)'(y - \mu)$$

$$\hat{\mu}^{(k)} = X\hat{\beta}^{(k)} + N\hat{f}^{(k)}$$

- المرحلة الرابعة : اختبار التقارب Convergence check

استناداً لمعادلة (16-1) فإن دالة الإمكان الأعظم الجزائية ستكون :-

$$\ell_c(\theta|y_c)^{(k+1)} = C - n \log \sigma^{2(k+1)} - \frac{1}{2\sigma^{2(k+1)}} \sum_{i=1}^n z_i^2 + \frac{\lambda^{(k+1)}}{\sigma^{2(k+1)}}$$

$$\sum_{i=1}^n z_i (y_i - \mu_i^{(k+1)}) - \frac{1 + \lambda^{2(k+1)}}{2\sigma^{2(k+1)}} \sum_{i=1}^n (y_i - \mu_i^{(k+1)})^2 \dots (26-1)$$

وعندها إذا كانت دالة الإمكان أقل من القيمة التخمينية (القادمة) $|\ell_c(\theta|y_c)^{(k+1)}| > |Q(\theta|\hat{\theta})|$ فإن الاجراء يرجع الى الخطوة الثانية ، وفي حالة العكس $|\ell_c(\theta|y_c)^{(k+1)}| < |Q(\theta|\hat{\theta})|$ فإن الخوارزمية تنتهي .

5- معيار أكاي للمعلومات [1] :-

(The Akaike information criterion) (AIC)

اقترح هذا المعيار من قبل الباحث Akaike عام (1974م) ويتم ايجاده وفق الصيغة الآتية :-
 $AIC = n(1 + \ln(2\pi)) + n \times \ln(\sigma^2) + 2m \dots (27-1)$

اذ ان :-

n : عدد المشاهدات في العينة .

σ^2 : تباين البواقي (Residual Variance) .

m : عدد المعلمات المقدرة في معادلة الانحدار باستثناء معلمة (β_0)

ويستفاد في هذا المعيار في بيان أي من نماذج الانحدار المقدرة للمتغير المعتمد من خلال استخدام طرائق التقدير المختلفة هو الأمثل لذلك المتغير المعتمد لغرض الاعتماد عليه في تمثيل البيانات والتنبؤ للمتغير المعتمد . علماً انه كلما كانت قيمة (AIC) للنموذج المقدر قليل كلما دل ذلك على ان هذا النموذج المقدر هو الأفضل والطريقة المستخدمة في تقديره هي الأفضل في التقدير للبيانات قيد الدراسة .

6- الجانب التطبيقي

نستعرض في الجانب التطبيقي نموذج التوزيع الطبيعي الملتوي و تقدير نموذج الانحدار الخطي العام للتوزيع الطبيعي الملتوي باستخدام خوارزمية تعظيم التوقع (EMA) . حيث تم التطبيق على بيانات تم الحصول عليها من سوق العراق للأوراق المالية .

1-6 البيانات

ان البيانات تم الحصول عليها من سوق العراق للأوراق المالية والتي تمثل عدد الأسهم المتداولة لمجموعة من القطاعات وتم اخذ البيانات بشكل اسبوعي ولأربع سنوات للفترة من (2016-2019) وبلغ عدد المشاهدات (150) .

إذ ان المتغير المعتمد (y) الذي يمثل المجموع الكلي (Total) لعدد الأسهم المتداولة لجميع القطاعات التي تم تداولها في السوق لكل أسبوع (اي هو عدد الأسهم المتداولة في سوق العراق) علماً ان مجموع الاسهم المتداولة يتجاوز (9) قطاعات مختلفة اضافة لشركات القطاع الخاص وقطاعات مختلفة . اما المتغيرات المفسرة ويبلغ عددها ستة متغيرات و هي :-

- 1- (X_1) ويمثل عدد الأسهم المتداولة للقطاع المصرفي (Bank) .
- 2- (X_2) ويمثل عدد الأسهم المتداولة لقطاع الاتصالات (Telecommunication) .
- 3- (X_3) ويمثل عدد الأسهم المتداولة لقطاع الخدمات (Services) .
- 4- (X_4) ويمثل عدد الأسهم المتداولة لقطاع الصناعة (Industry) .
- 5- (X_5) ويمثل عدد الأسهم المتداولة لقطاع الفنادق والسياحة (Tourism&Hotels) .
- 6- (X_6) ويمثل عدد الأسهم المتداولة لقطاع الزراعة (Agriculture) .

2-6 اختبار البيانات

تم توظيف اختبار (Anderson Darling Test) لمعرفة فيما اذا كانت متغيرات الدراسة تتبع التوزيع الطبيعي والذي يعتمد على الفرضية التالية :-

البيانات تتبع التوزيع الطبيعي: H_0

البيانات لا تتبع التوزيع الطبيعي: H_1

الجدول (1) ادناه يمثل قيم احصاءات الاختبار لكافة متغيرات الدراسة ، ويتبين من خلاله ان كافة قيم مستوى الدلالة (P-Value) كانت اقل من مستوى المعنوية (0.05) لذلك ترفض فرضية العدم ، وهذا يعني ان كافة متغيرات الدراسة لا تتبع التوزيع الطبيعي .

جدول (1)

اختبار Anderson darling لمتغيرات الدراسة

Variable	AD Statistic	P-value
Total (y)	17.817	< 2.2e-16
Bank (X_1)	19.376	< 2.2e-16
Telecommunication (X_2)	32.778	< 2.2e-16
Services (X_3)	32.479	< 2.2e-16
Industry (X_4)	13.36	< 2.2e-16
Tourism & Hotels (X_5)	19.535	< 2.2e-16
Agriculture (X_6)	17.187	< 2.2e-16

3-6 طريقة تقدير خوارزمية تعظيم التوقع (EMA) :-

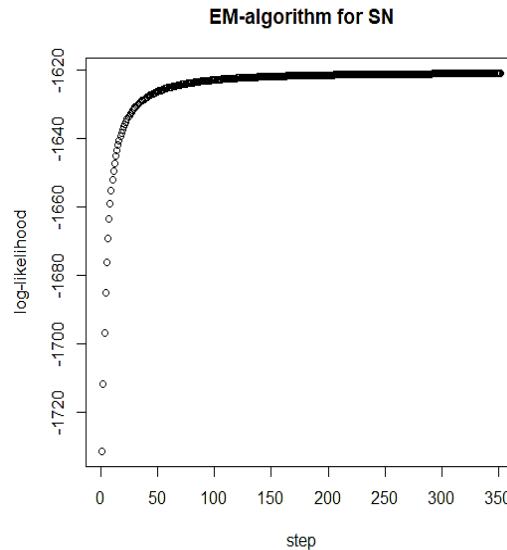
لتقدير الانموذج الخطي من خلال خوارزمية تعظيم التوقع (EM-GLM) تم اتباع الخطوات المشار اليها في الفقرة (4) ، كما هو مبين في الجدول (2) والذي يشمل تحديد قيم المعلمات الابتدائية المقدرة في الانموذج الانحدار الخطي العام بالطريقة التقليدية والمشار اليها في مرحلة التهيئة (Initialization) . وقد تم تقديرها اذ بلغت ($\mu_0 = 832, \sigma_0^2 = 0.9, \lambda_0 = 2.1$) اما فيما يخص الدالة ($f_0(t)$) فقد تم توظيف دالة التوزيع الطبيعي المنتظم للحصول عليها و بلغ عدد التكرارات العظمى بلغت (351) تكرار . وبعدها تطبيق خطوات تعظيم التوقع (E-step) وذلك من خلال تحديد التكرار الذي يعطي اعلى قيمة لدالة الإمكان .

جدول (2)

نتائج خوارزمية تعظيم التوقع EMA

μ_0	σ_0^2	λ_0	$f_0(t_i)$	Max Iteration	Best log-likelihood
832	0.9	2.1	$f_0(t_i) \sim UN(0,1)$	351	-1620.868

ويلاحظ في الجدول أعلاه ان اعلى قيمة بلغت (-1620.868) وتعطي اعظم قيمة لدالة الإمكان .



الشكل (4)

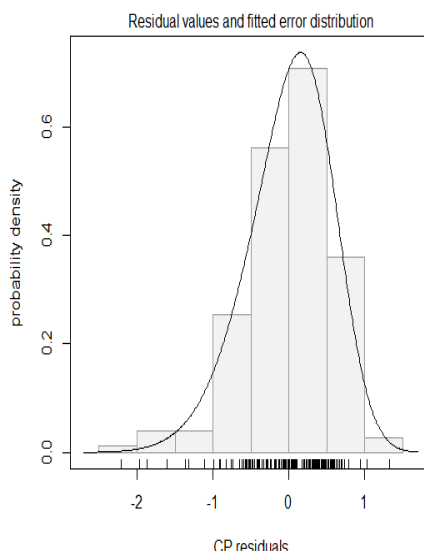
عدد التكرارات وافضل تكرار لتعظيم توقع دالة الامكان
يلاحظ في الشكل (4) التكرار الامثل لخوارزمية (EMA) بالنسبة لتقدير معلمات المتغير المعتمد

وبعد ان تم تحديد افضل تكرار يعطي اعظم قيمة في دالة الإمكان يتم تقدير نموذج الانحدار الخطي العام بطريقة (EMA) والجدول (3) يبين قيمة المعلمات المقدرة حسب معادلات (23-1) ، (24-1) ، (25-1) . وتم حساب المعلمات (β) عن طريق المعادلة (22-1) ، وتم استخراج معيار اكيائي وفق معادلة (27-1) جدول (3)

الانموذج المقدر بطريقة (EMA)

	Beta	s.d	Z	Sig
Intercept	8.404e+00	1.299e-01	6.468e+01	0.000 **
Bank	4.383e-05	8.548e-06	5.127e+00	0.000 **
Telecommunication	2.812e-04	3.103e-04	9.063e-01	0.365
Services	1.734e-04	1.889e-04	9.181e-01	0.359
Industry	1.845e-04	8.681e-05	2.125e+00	0.034
Tourism & Hotels	4.971e-04	1.856e-03	2.678e-01	0.789
Agriculture	-2.573e-03	1.889e-03	-1.36e+00	0.173
σ^2	0.9862	0.035		
λ_{res}	-0.4728	0.904		
AIC	236.890			
λ	5.9583			
rmse	0.3597			
rmae	0.4408			

ومن خلال قيم احصاءات اختبار (Z) ومستويات الدلالة المناظرة لها تبين ان المتغيرات التوضيحية التي تمثل عوائد الاسهم للقطاعات المختلفة والتي كانت ذات تأثير معنوي في عوائد الاسهم لسوق العراق للأوراق المالية هي فقط القطاع المصرفي ، فقد كانت قيم مستويات الدلالة لهذه التقديرات اقل من مستوى المعنوية ، في حين كانت قيم مستويات الدلالة لبقية المتغيرات التوضيحية أي عوائد اسهم قطاع الخدمات والقطاع الزراعي قطاع الاتصالات وقطاع الصناعة وقطاع السياحة اكبر من مستوى المعنوية وهذا يدل على عدم وجود تأثير لهذه المتغيرات في الأنموذج .



الشكل (5)

يبين توزيع الخطأ وقيم البواقي للنموذج المقدر بطريقة (EMA)

يبين الشكل (5) توزيع الخطأ (Error Distribution) وقيم البواقي للنموذج المقدر بطريقة (EMA) ، اذ ان توزيع الخطأ يكون ذي التواء سالب او اليسار . وهذا ما تم تأشيريه في قيمة معامل الالتواء المقدر للأخطاء العشوائية (λ_{res}) في جدول (3) اذ كانت قيمتها بالسالب .

الاستنتاجات

استنتج الباحثين ان نتائج تقدير خوارزمية تعظيم التوقع (EMA) لنموذج الانحدار الخطي العام وجود متغيرات معنوية في الانموذج وهي (قطاع المصارف) وتبين وجود متغيرات ظهرت معنوية ولكنها في الواقع غير معنوية .

اذ ان قيمة جذر متوسط مربعات الخطأ (rmse) فقد بلغت (0.3597) ومعامل التواء (5.9583) اما معيار اكاكي (AIC) فقد بلغ (236.890) وبلغ جذر متوسط مطلق الأخطاء (mae) مقداره (0.4408) اما التباين (σ^2) فقد بلغ (0.9862) كما بلغت قيمة معامل الالتواء المقدر للأخطاء العشوائية (-0.4728) اي ان الاخطاء ملتوية نحو اليسار .

ان الانموذج المقدر وفق طريقة (EMA) ترشح عنه وجود متغير توضيحي معنوي وهو (قطاع المصارف) .

التوصيات

اعتماد خوارزمية تعظيم التوقع (EMA) في تقدير معلمات نموذج الانحدار الخطي العام عندما البيانات لا تتبع التوزيع الطبيعي نتيجة لوجود قيم شاذة .

أظهرت نتائج التقدير بطريقة (EMA) ان سوق العراق للأوراق المالية يعتمد اساساً على قطاع المصارف مما يعطي مؤشر على عدم الاهتمام بالقطاعات الأخرى لذلك نوجه متخذي القرار بزيادة الاستثمار وتطوير القطاعات الأخرى والتي من شأنها انعاش القطاع الاقتصادي للدولة .

المراجع

- 1- Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csáki, F. (eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281.

- 2- Azzalini, A. & Capitanio (2013). The skew-normal and related families (Vol. 3). Cambridge University Press.
- 3- Azzalini, A. (1985). A class of distributions which includes the normal ones. Scandinavian journal of statistics, 171-178.
- 4- Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. Statistica, 46(2), 199-208.
- 5- Azzalini, A., & Valle, A. D. (1996). The multivariate skew-normal distribution. Biometrika, 83(4), 715-726.
- 6- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.
- 7- Ferreira, C. S., & Paula, G. A. (2017). Estimation and diagnostic for skew-normal partially linear models. Journal of Applied Statistics, 44(16), 3033-3053.
- 8- Ganjavi, A., Christopher, E., Johnson, C. M., & Clare, J. (2017, April). A study on probability of distribution loads based on expectation maximization algorithm. In 2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT) (pp. 1-5). IEEE.
- 9- Lin, T. I., Lee, J. C., & Yen, S. Y. (2007). Finite mixture modelling using the skew normal distribution. Statistica Sinica, 909-927.
- 10- Lyu, W., & Feng, Z. (2018, April). Estimation and Application of Skew-normal Data for Generalized Linear Regression. In 2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018). Atlantis Press.
- 11- McLachlan, G., & Krishnan, T. (2007). The EM algorithm and extensions (Vol. 382). John Wiley & Sons.
- 12- Mudholkar, G. S., & Hutson, A. D. (2000). The epsilon-skew-normal distribution for analyzing near-normal data. Journal of Statistical Planning and Inference, 83(2), 291-309.
- 13- Ngunkeng, G. (2013). Statistical analysis of skew normal distribution and its applications (Doctoral dissertation, Bowling Green State University).
- 14- Venegas, O., Sanhueza, A. I., & Gómez, H. W. (2011). An extension of the skew-generalized normal distribution and its derivation. Proyecciones (Antofagasta), 30(3), 401-413.
- 15- Xu, L., & Jordan, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. Neural computation, 8(1), 129-151.
- 16- Yalçinkaya, A., Şenoğlu, B., & Yolcu, U. (2018). Maximum likelihood estimation for the parameters of skew normal distribution using genetic algorithm. Swarm and Evolutionary

Using The Expectation Maximization Algorithm In Estimating The General Linear Regression Model When The Data Are Distributed Skew Normal With The Application

Amin Hussein Hadi

Prof. Dr. Suhad Ali Shahid

Abstract

A twisted normal distribution by the torsion parameter (λ) is one of the caustic distributions, but by adding a parameter to the shape that determines the amount of torsion, it is a special case of a normal distribution. The discrepancy in the data, i.e. the presence of asymmetric data, was the reason for the emergence of the twisted normal distribution (SND), and this distribution is one of the important distributions due to the presence of the torsion parameter whose value changes constantly, so this parameter provides a good model for simulating many of the phenomena that arise with such data.

The main focus of the study is to shed light on the twisted normal distribution (SND) with the presence of the torsion parameter (λ) of the Generalized Linear Regression Model, as the model parameters were generally estimated through the use of the Expectation Maximization algorithm (EMA). Maximization Algorithm), which gave desirable and more efficient parameter specifications. The data that was studied are financial data obtained from the Iraq Stock Exchange, which represent the number of shares traded for a group of sectors. The data were taken on a weekly basis for four years for the period from (2016-2019) and the number of views reached (150).