JRUCS
Journal of AL-Rafidain University College for Sciences

PISSN: (1681-6870); EISSN: (2790-2293)
مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

AL- Rafidain University College

تقدير دالة المخاطرة باستعمال أنموذج الانحدار log-logistic لبيانات التلوث البيئي

أ.د.انتصار عريبي فدعم	سماح صباح حسن
entsar_arebe@coadec.uobaghdad.edu.iq	samahsabah1993@gmail.com
قسم الاحصاء - كلية الإدارة والاقتصاد - جامعة بغداد، بغداد، العراق	
أ.م.د اقبال محمود علوان	
iqbal.alwan@coadec.uobaghdad.edu.iq	
قسم الاحصاء - كلية الإدارة والاقتصاد - جامعة بغداد، بغداد، العراق	

المستخلص

يحتوي التوزيع Log-Logistic على دالة خطر غير خطية (غير رتيبة) مما يجعله مناسباً لنمذجة بعض مجموعات بيانات البقاء الحية وحمايته من المخاطر الناجمة من التلوث البيئي في عصرنا الحديث مما ادى الى انتشار عدد من الامراض ومنها أمراض السرطان وفرط الحركة ومشاكل التنفس وغيرها من الامراض التي بدأت تنتشر في مجتمعنا اليوم .

تم وصف نموذج الانحدار Log-Logistic الذي تتقابل فيه دالة الخطر لعينات المنفصلة مع الوقت ويوفر ايضاً نموذجاً خطياً لاحتمالات البقاء على قيد الحياة في اي وقت يتم اختياره .

في هذا البحث تم تقدير دالة المخاطرة لنموذج الانحدار Log-Logistic باستعمال طريقة الامكان الاعظم وطريقة الخوارزمية الجينية في ظل وجود الاحصاءات المرتبة المعممة وتم تطبيق هذه الطرق على بيانات تلوث الهواء التي تم الحصول عليها من شركة مصافي الوسط في بغداد (مصفى الدورة) والتي تمثل القياسات اليومية لمركبات التلوث البيئي والتي تستند الى الوقت للمدة من 2018 الى 2020.

معلومات البحث

تواريخ البحث
تاريخ تقديم البحث: 2022/12/15
تاريخ قبول البحث: 2023/3/3
تاريخ رفع البحث على الموقع: 2023/12/31

الكلمات المفتاحية
توزيع log-logistic، أنموذج الاحصاءات المرتبة المعممة، بيانات تلوث بيئي، طريقة الامكان الاعظم، طريقة الخوارزمية الجينية، انموذج الانحدار log-logistic ، دالة المخاطرة، GLM

للمراسلة:
أ.د. انتصار عريبي فدعم
entsar_arebe@coadec.uobaghdad.edu.iq

https://doi.org/10.55562/jrucs.v54i1.583

1. المقدمة (Introduction)

ان مفهوم الاحصاءات المرتبة المعممة التي قدمها العالم الالماني Udo kamps[20] عام 1995 أوجد اهتماماً كبير في السنوات الاخيرة لكونه أنموذجاً موحداً يحتوي على عدة نماذج من المتغيرات العشوائية المرتبة التي تنتج من التجارب العملية المختلفة التي من اشهرها واكثرها استعمالاً: الاحصاءات المرتبة الاعتيادية والقيم المسجلة بنوعيها العليا والدنيا والقيم المسجلة من الرتبة k والعينات المراقبة بأنواعها والاحصاءات المرتبة التتابعية وقيم بفيفر المسجلة.

ويمكن الحكم على اي نموذج يحاول تمثيل تجربة المخاطرة لظاهرة تلوث الهواء بشكل واقعي على مدى ملاءمته من خلال قرب دالة الخطر لنظرية المستمدة من النموذج الخطر الحقيقي المشتق تجريبياً من البيانات باستعمال تقنيات جداول المخاطرة، قد تشمل النماذج المجهزة لبيانات المخاطرة على اشكال معلمية او لا معلمية لدالة الخطر. هذا يعتمد ما اذا كان هذا النموذج قد تم تعريفه كتوزيع معروف او اذا كان غير محدد تماماً، في هذا البحث سنهتم فقط بأشكال المعلمية التي يسهل ملاءمتها بشكل عام على الاشكال لامعلمية.

يتشابه التوزيع log-logistic الى حد كبير في الشكل مع توزيع log-normal ولكنه اكثر ملاءمة للاستعمال في تحليل بيانات تلوث الهواء وتأثيره على قيد الحياة، هذا بسبب النموذج الرياضي لتعامل مع المشاهدات الخاضعة للرقابة والتي تحدث بشكل متكرر في مثل هذه البيانات المساهمة التي قدمتها المشاهدة الخاضعة للرقابة الصحيحة في الاحتمالية تساوي قيمة دالة الناتجة في وقت الرقابة، لا يجب ان نضع نموذجاً لتوزيعات دالة المخاطرة داخل كل عينة بل يجب ايضاً اعطاء بعض المقاييس للاختلاف في المخاطرة بين العينات [4]. ويمكن القيام بذلك عن طريق تعيين معلمة واحدة للتوزيع تساوي دالة خطية للمتغيرات المشتركة التي تحدد العينات بما يؤدي الى نموذج الانحدار، في حالة log-logistic ستظهرقيمة هذا النموذج الخطي لها معنى ويمكن تفسيره بسهولة .

سيظهر ايضاً انه في نموذج الانحدار هذا فأن دالة الخطر العينات مختلفة لاتتناسب مع مرور الوقت كما في توزيع Weibull ولكن نسبتها تميل الى الوحدة هذه الخاصية مرغوبة عندما تميل التأثيرات او الاختلافات بين مراحل تلوث الهواء عند المراقبة الى تضاؤل بمرور الوقت وتصبح احتمالات المخاطرة على قيد الحياة لمجموعات مختلفة من التلوث اكثر تشابهاً .

عام 2021 م، قام الباحث (Abd-Elmonem Teamah)[3] وأخرون بدراسة التوزيعات ذات الذيل الثقيل التي لها دور كبير في دراسة مجموعات بيانات المخاطرة. قدموا في هذا البحث نموذجًا إحصائيًا جديدًا نسبيًا والذي يسمى (exponentiated log-logistic distribution)، تم اشتقاق خصائصه الإحصائية مثل العزوم، والدالة المولدة للعزوم، والدالة الكمية، والانتروبيا، وإلاحصاءات المرتبة. استعمل ست طرائق تقدير تم استنتاج بعض المقاييس الاكتوارية مثل القيمة المعرضة للخطر، و استعمال ثلاث مجموعات بيانات حقيقية من مجالات مختلفة لإظهار كيفية ملاءمة هذه النماذج المقترحة لمجموعات البيانات هذه ومقارنة بالعديد من النماذج الأخرى المعروفة والنماذج ذات الصلة. وفي عام (2022) قدم الباحثان (Entsar. D and Samah. H)[8] بدراسة الاحصاءات المرتبة لتوزيع log-logistic في تقدير معلماته باستعمال طريقة بيز القياسية المعتمدة على دالة الخسارة التربيعية (Bayslf) ولتحديد الطريقة الامثل من خلال مقارنتها مع طريقة الامكان الاعظم MLE وفق اسلوب محاكاة بأخذ نماذج مختلفة للمعالم الاولية وحجوم عينات مختلفة وبمعايير مقارنة MSE,IMSE وكذلك تم تطبيقها على بيانات حقيقية لمرضى سرطان الثدي وتم ايجاد دالة البقاء والمخاطرة لهذا التوزيع.

2. الاحصاءات المرتبة المعممة (Generalized order statistics)

تسمى المتغيرات العشوائية $U(j,n,\tilde{m},s)$, أذ ان $j=1,2,\ldots,n$ بالإحصاءات المرتبة المعممة المنتظمة اذا كانت دالة الكثافة الاحتمال المشتركة بالصيغة الاتية :

$$f^{U(1,n,\tilde{m},s)..U(n,n,\tilde{m},s)}(Y_1 \ldots . Y_n) = C_{n-1}[\prod_{i=1}^{n-1}(1-Y_i)^{m_i}](1-Y_n)^{s-1} \quad (1)$$

اذ ان :

$$n \in N; n \geq 2; S \geq 1$$
$$\tilde{m} = (m_1, \ldots, m_{n-1}) \in R^{n-1}$$
$$0 < Y_1 \leq \cdots \leq Y_n \leq 1$$
$$C_{n-1} = \prod_{i=1}^{n} Y_i = s\prod_{i=1}^{n} Y_i$$
$$Y_i = s + n - j + \sum_{i=j}^{n-1} m_i > 0$$

لجميع قيم $j0 \in \{1,2,..,n-1\}$

يمكن الحصول على الدالة الاحتمالية المشتركة للإحصاءات المرتبة المعممة لعدد n من المتغيرات العشوائية $(j,n,\tilde{m},s)$ T أذ ان $j\in\{1,2,\ldots,n\}$ التي تتبع أي توزيع احتمالي بدالة كثافة الاحتمالية هي $f(y)$ ودالة التوزيع التراكمية هي $F(y)$ وتأخذ الصيغة الاتية أنظر kamps(1995)

$$f^{y(1,n,\tilde{m},s)..y(n,n,\tilde{m},s)}(y_1 \ldots . y_n) = C_{n-1}\left[\prod_{i=1}^{n-1}(1-F(y_i))^{m_i} f(y_i)\right]\left[\left(1-F(y_n)\right)^{s-1} f(y_n)\right]) \quad (2)$$

اذ ان $F^{-1}(0) < y_1 \leq \cdots \leq y_n < F^{-1}(1)$

يحتوي أنموذج الاحصاءات المرتبة المعممة على العديد من نماذج المتغيرات العشوائية المرتبة ,وبأختيار مناسب للمعلمات $(\tilde{m}, s)$ نحصل على نماذج الاحصاءات المرتبة المعممة .

3. الاحصاءات المرتبة ذات القيم المسجلة ذات المرتبة العليا

Order Statistics with Higher-Order Recorded Values

اذا كانت $y_1 \dots y_n$ سلسلة لانهائية من متغيرات عشوائية مستقلة ومتطابقة التوزيع وكانت $f(y)$ ، $F(y)$ هما دالة كثافة الاحتمالية ودالة التوزيع التراكمية على التوالي للمتغير العشوائي y فأن المشاهدات y والمشاهدة y_j قيمة مسجلة عليا اذا كانت قيمتها اكبر من كل المشاهدات الاولية اي $y_j > y_{j-1}$ حيث $j>1$.

بفرض ان $y_{s(1)} \dots y_{s(n)}$ هي اول n قيمة مسجلة عليا من مجتمع دالة كثافة الاحتمالية $f(y)$ ، ودالة التوزيع التراكمية $F(y)$. يمكن الحصول على القيم المسجلة العليا كحالة خاصة من الاحصاءات المرتبة المعممة وذلك بوضع $m=-1$ =y_i لجميع $y_i = 1$, $k=1$,$i=1,2,,n-1$ في معادلة رقم (2) فنحصل على دالة كثافة الاحتمالية المشتركة لجميع قيم $y_{s(1)} \cdots y_{s(n)}$ بالشكل الاتي :

$$f(Y_1 \dots . Y_n) = \left[\prod_{i=1}^{n-1} (\bar{F}(Y_i)^{-1} f(Y_i)\right] \left[\bar{F}(Y_n) \ \right]^0 f(Y_n) \tag{3}$$

$$= \prod_{i=1}^{n-1} \frac{f(y_i)}{\bar{F}(y_j)} f(y_n) \tag{4}$$

$$= \prod_{i=1}^{n-1} \frac{f(y_i)}{\bar{F}(y_j)} f(y_n) \tag{5}$$

$$= f(y_n) \prod_{i=1}^{n-1} h(y_i) \tag{6}$$

اذ ان $-\infty < y_1 < y_2 < \cdots < \infty$

$$h(y_i) = \frac{f(y_i)}{\bar{F}(y_i)} \qquad ; \bar{F}(y_i) = 1 - F(y_i) \tag{7}$$

وبالتالي يمكن الحصول على المقدرات دالة المخاطرة لتوزيع log-logistic بالاعتماد على القيم المسجلة ذات المرتبة العليا [16] .

4. انموذج الانحدار Log-Logistic (Log-Logistic Regression Model)

يعتبر التوزيع log-logistic بديلاً ممتازاً لتوزيع Weibull لنمذجة المعلمية لبيانات المخاطرة وحيث يكون معدل الخطر غير رتيب و يكون من المرغوب ايضاً تقارب دالة الخطر لعينات المنفصلة ويمكن تركيب النموذج على اللبيانات الخاضعة للرقابة من الجهة اليمنى واليسرى. ويمكن كتابة دالة التوزيع log-logistic لعينة واحدة على النحو التالي [4]:

$$F(y_i, \varphi, \theta) = \frac{1}{1 + y_i^{-\varphi} e^{-\theta}} \qquad ; \ 0 \le y_i \le \infty \ ; \varphi > 0 \tag{8}$$

اذ ان :

y_i: يمثل المتغير الترتيبي (الوقت)

θ : يمثل معلمة القياس للموقع

φ: يمثل معلمة الشكل

بالنسبة لنموذج الانحدار الذي يشمل العديد من العينات سيكون φ هو نفسه لعينات بينما سيتم وضعθ بالشكل الاتي :

$$\theta_s = \beta_0 + \beta_1 y_{s1} + \cdots + \beta_P y_{sp} = \beta' y_s \tag{9}$$

لجميع بيانات مخاطر التلوث في العينة (S=1,2,3,..,s) و y_s تمثل المتجه للمتغيرات المشتركة التي تحدد العينة β هو متجه لمعلمات غير معروفة، احتمالية الوفاة حسب الوقت y لفرد العينة S_{th}. ودالة $F(y_i, \varphi, \theta)$ في معادلة رقم (3) يمكن كتابتها بدلالة $F_s(y)$

$$\left\{\frac{F_s(y)}{1-F_s(y)}\right\} = \varphi \log y + \beta' y_s \quad (10)$$

لتصبح نموذجاً خطياً لاحتمالات اللوغارتميات عند الوفاة بوجود المتغير الترتيبي y بالنسبة لعينتين مع متجهات من المتغيرات المشتركة y_1, y_2 والفرق بين دالة الاحتمالية يكون كالاتي :

$$log\left\{\frac{F_1(y)}{1-F_1(y)}\right\} - log\left\{\frac{F_2(y)}{1-F_2(y)}\right\} = \beta'(y_1 - y_1) \quad (11)$$

مستقل عن الوقت y_i وبالتالي فأن احدى دالة الاحتمالية هي مضاعف للأخرى بالنسبة لعينتين مع اختلاف وحدة في المتغير المشترك (جميع الاختلافات المتغيرة الاخرى تساوي صفر) فأن هذا الثابت يساوي e^{β_i} ويمكن التعرف على β_i على انه نسبة الاحتمالات اللوغاريتمات هذه حالة خاصة لنموذج الاحتمالات النسبية في نموذج الاحتمالات النسبية العامة يمكن استبدال $\varphi \log y$ بأي دالة [7] y. ويمكن اعطاء دالة المخاطرة للعينة بالشكل التالي :

$$h_s(y_i) = \frac{\varphi}{y_i\left(1 + y_i^{-\varphi} e^{-\theta_s}\right)} \quad (12)$$

5. طرائق التقدير (Estimation methods)

5.1. طريقة الامكان الاعظم بالاعتماد على القيم المسجلة ذات المرتبة العليا

(The Method of Maximum Likelihood Based on the Highest-Order Recorded Values)

تعتبر هذه الطريقة من اهم الطرائق التقدير المهمة لما تحتويه من خواص جيدة ومنها الثبات كفاءة عالية واتساق في بعض الاحيان، ويمكن اعادة كتابة معادلة رقم (6) :

$$f\left(Y_{(1)}, Y_{(2)} \dots . Y_{(n)}\right) = f(y_n) \prod_{i=1}^{n-1} h(y_i) \quad ; -\infty < y_{(1)} < y_{(2)} < \cdots < \infty \quad (13)$$

$$= \prod_{i=1}^{n-1} \frac{\varphi}{y_i\left(1 + y_i^{-\varphi} e^{-\theta}\right)} * \frac{\varphi e^{-\theta} y_i^{-\varphi-1}}{\left(1 + y_i^{-\varphi} e^{-\theta}\right)^2} \quad (14)$$

وبعد التبسيط نحصل على :

$$f\left(Y_{(i)}/\varphi, \theta\right) = \varphi^{2n} e^{-\theta_n} \prod_{i=1}^{n-1} y_i^{-\varphi-2}\left(1 + y_i^{-\varphi} e^{-\theta}\right)^{-3} \quad (15)$$

ويمكن كتابة دالة الامكان اللوغارتمية(log likelihood)

$$ln\, f\left(Y_{(i)}/\varphi, \theta\right) = 2n\, ln\varphi - \theta_n - (\varphi - 2) \sum_{i=1}^{n-1} \ln y_i - 3 \sum_{i=1}^{n-1} \ln(1 + y_i^{-\varphi} e^{-\theta}) \quad (16)$$

$$\frac{\partial ln\, f\left(Y_{(i)}/\varphi, \theta\right)}{\partial \theta} = -1 + 3e^{-\theta} \sum_{i=1}^{n-1} \frac{y_i^{-\varphi}}{1 + y_i^{-\varphi} e^{-\theta}} \quad (17)$$

$$\frac{\partial ln\, f\left(Y_{(i)}/\varphi, \theta\right)}{\partial \varphi} = \frac{2n}{\varphi} - \sum_{i=1}^{n-1} \ln y_i + 3e^{-\theta} \sum_{i=1}^{n-1} \frac{y_i^{-\varphi} \ln y_i}{1 + y_i^{-\varphi} e^{-\theta}} \quad (18)$$

ومن بعدها نساوي المعادلتين (17) و (18) بالصفر ومن ثم نحلها بإحدى الطرق العددية للحصول على مقدرات الامكان الاعظم لمعلمات φ و θ للتأكد من أن المقدرات هي مقدرات الأمكان الأعظم نحتاج الى حساب مشاهدات معلومات فيشر Observed (Fisher Information) سيمكننا ذلك من انشاء فترات ثبات لمعلمات بناء على اساس التوزيع s-normal واختبارالاحتمالية من خلال المحاكاة ,والان نستمد معلومات فيشر لمشاهدات الاحتمالية باستخدام المعادلتين (17) و(18) يصبح لدينا :

$$I_{(\hat{\theta},\hat{\varphi})} = \begin{bmatrix} \frac{\partial^2 ln\, f(Y_{(i)}/\varphi,\theta)}{\partial \boldsymbol{\varphi}^2} & \frac{\partial^2 ln\, f(Y_{(i)}/\varphi,\theta)}{\partial \boldsymbol{\varphi}\partial\theta} \\ \frac{\partial^2 ln\, f(Y_{(i)}/\varphi,\theta)}{\partial \boldsymbol{\varphi}\partial\theta} & \frac{\partial^2 ln\, f(Y_{(i)}/\varphi,\theta)}{\partial \boldsymbol{\theta}^2} \end{bmatrix}$$

$$\frac{\partial^2 ln\, f(Y_{(i)}/\varphi,\theta)}{\partial \boldsymbol{\theta}^2} = 3e^{-\theta}\sum_{i=1}^{n-1}\frac{-y_i^{-\varphi}}{\left(1+y_i^{-\varphi}e^{-\theta}\right)^2} \quad (19)$$

$$\frac{\partial^2 ln\, f(Y_{(i)}/\varphi,\theta)}{\partial \boldsymbol{\varphi}\partial\theta} = -3e^{-\theta}\sum_{i=1}^{n-1}\frac{-y_i^{-\varphi}\ln y_i}{\left(1+y_i^{-\varphi}e^{-\theta}\right)^2} \quad (20)$$

$$\frac{\partial^2 ln\, f(Y_{(i)}/\varphi,\theta)}{\partial \boldsymbol{\varphi}^2} = \frac{-2n}{\varphi^2} - 3e^{-\theta}\sum_{i=1}^{n-1}\frac{-y_i^{-\varphi}\ln^2 y_i}{\left(1+y_i^{-\varphi}e^{-\theta}\right)^2} \quad (21)$$

ومقدر الامكان لدالة المخاطرة $h(y_i)$ حيث$y_i > 0$ بالاعتماد على القيم المسجلة ذات المرتبة العليا يكون كالاتي:

$$h_s(y_i) = \frac{\varphi}{y_i\left(1+y_i^{-\varphi}e^{-\theta_s}\right)} \quad (22)$$

5.2. طريقة الخوارزمية الجينية بالأعتماد على القيم المسجلة ذات المرتبة العليا

(Genetic algorithm method based on the highest-ranked recorded values)

تعد الخوارزمية الجينية من طرائق االبحث العشوائي تعالج مشكلة ما من اجل الوصول الى افضل النتائج الممكنة تتمحور حول تقنيات التطور من منطلق نظرية دار للتطور وتنص على البقاء للأفضل بتقليده عمل الطبيعة من خلال الاحتفاظ بالصفات الجيدة الموجودة في جيل الاباء ونقلها الى جيل الابناء هدفها الحصول على ذرية فورية تتمتع افضل صفات الاباء على اقل تقدير[14] .

- **مراحل الخوارزمية الجينية**

تختلف الخوارزمية الجينية حسب اختلاف فروع التقنيات التطويرية الا أنها مشتركة بهذا المراحل التالية[1]:

1. البداية: تتمثل في تعداد سكاني عشوائي من الكروموسومات (فضاء البحث) وبعبارة اخرى هي مجموعة حلول المسألة
2. التهيئة: هي عملية انشاء الجيل الابتدائي تتضمن توليد الكروموسومات عشوائية بقدر حجم المجتمع وحسب طبيعة المشكلة[19]
3. دالة الهدف ودالة التقييم (المفاضلة): من خلالها يتم تقييم الكروموسومات بحيث يعطي كل كروموسومات قيمة معينة تمثل مدى كفاءة الى مدى اقترابه من الحل وتكون دالة الهدف اما عظمى (Maximum) او صغرى (Minimum) وعليها تعتمد دالة التقييم
4. الاستنتاج: عملية توليد جيل جديد من الافراد (الكروموسوم) التي يتم اختيارها من خلال عملية الاختيار وفق مبدأ البقاء للأصلح من ثم اجراء عليها عملية التهجين وعملية الطفرة لأنتاج ابناء الجيل اللاحق .
5. مجتمع جديد : يتضمن توليد جيل جديد بتكرار المراحل التالية الى ان يكتمل الجيل ويتمثل الجيل بمايأتي[19] :

أ. الاختيار: هي عملية اختيار كروموسومات مناسبة من الجيل القديم لتكوين الجيل الجديد يتم انشاؤها وفقاً لقيم دالة التقييم لغرض وجود الكروموسومات مع اعلى قيمة لدالة التقييم في الاجيال الجديدة [5]

ب. التهجين: يتم اختيار الكروموسومات الجيدة من الجيل الاول يتم التزاوج بين كرموسومين لتكوين الجيل الجديد (الابناء) اعتماداً على الكروموسومات (الام) بأخذ الصفات الجيدة منها من طرائقها التهجين بنقطة وبنقطتين وغيرها [14].

ت. الطفرة: بعد انشاء الجيل الجديد (الابناء) وباحتمال وجود الطفرة فتتم عملية الطفرة بأحداث بفترات عشوائية في صبغتها الكروموسومية وهذا يؤدي بنا الى الحفاظ على الصفات الجيدة بين الجينات في الكروموسوم الواحد والوصول الى الحل بشكل اسرع وفيها يحدث التبادل بين الكروموسوم وعندما لاتكون هنالك طفرة فتجري عملية استنساخ الكروموسومات (الاباء) مباشرة دون حدوث عملية التهجين [6].

- **تم الاشتقاق والتوظيف من قبل الباحث**

1. الاختبار: نختبر الحل بتوفر شرط التوقف من عدمه فعند توفره تتوقف الخوارزمية الجينية ويفيد الحل الجديد من تكون اخر جيل [10].
2. معيار التوقف: يستمر تكوين الاجيال بشكل متعاقب من اجل تحسين امثلية الحل حتى يتحقق شرط التوقف الذي يعتمد على مقياس توقف الخوارزمية الجينية (الحل الامثل) ويختلف هذا المقياس حسب المشكلة المراد معالجتها.
3. الانهاء : تنتهي الخوارزمية الجينية عند وجود احد العوامل التالية [11]:
   * ايجاد الحل الامثل
   * الوصول الى عدد الاجيال المطلوبة
   * توفر قيمة معينة مثل حساب كلفة الانتاج
   * الوقوع في قيمة الصغرى المحلية ولايمكن الخروج منها

- **تطبيق مراحل الخوارزمية الجينية في أنموذج الانحدار log-logistic الثنائي :**

نطبق مراحل الخوارزمية الجينية في معادلة دالة الهدف لكل طريقة لإيجاد تقديرات معلمات أنموذج الانحدار log-logistic الثنائي وفقاً لمايأتي [19]:

1. البداية: تكوين الكروموسوم من خلال قيم β_p التي تشكل جينات الكروموسوم وان (P=0,1,2,...,p) ضمن الاعداد الحقيقية.
2. التهيئة: انشاء الجيل الابتدائي عن طريق ايجاد قيمة اولية للجينات مع القيم العشوائية لمجموعة القيود الاخرى.
3. دالة الهدف: يتم تقييم الكروموسوم من حيث الكفاءة وصولاً الى الحل الامثل بتحديد قيمة β_p
4. اجراء عملية الاختبار للكروموسوم الذي يمتلك قيمة دالة الهدف صغيرة بأختيار الاحتمال الكبير لها وايجاد دالة التقييم له من خلال المعادلة التالية:

$$Fitness\ Function = \frac{1}{1 + objective\ function}$$

من خلال صيغة دالة التقييم نستطيع ايجاد الاحتمالية هذه الدالة (افضل القيم) حسب الصيغة الرياضية الاتية [18] :

$$C_{(i)} = \frac{f(i)}{\sum_{i=1}^{n} f(i)} \qquad (23)$$

اذ ان:

$C_{(i)}$: تمثل احتمالية الفرد i

f(i): دالة التقييم الفرد i

n : يمثل حجم المشاهدات

1. في هذه الخطوة يتم تهجين للكروموسومات الجيدة في صفاتها عن طريق التزاوج بين كل كروموسومين ويتطبق احد معاييره وهو التهجين المنظم بالأعتماد على احتمالية التهجين Pc وتقارن هذه القيمة مع قيمة الجينات للكرموسومين (الاباء) لتكوين الجيل الجديد الابناء ويحدث التبادل عندما تكون قيمة الجين الاكبر او تساوي القيمة الاحتمالية [7].
2. اخر خطوة يمكن ان تمر بها الكروموسومات هي عملية الطفرة وايضا تعتمد على مقدار احتمالي (Pm) للمعلمات استبدال جينات منقاة عشوائياً مع قيمة جديدة ايضاً حصلنا عليها بشكل عشوائي وفق الصيغة التالية:

مجموع الجينات = عدد الجينات في الكروموسوم * عدد السكان

6. الجانب العملي (The Practical side)

6.1. التلوث البيئي (Environmental pollution)

يعرف التلوث البيئي بأنه احداث تغير في النظام البيئي بشكل يؤدي الى عدم استطاعة النظام البيئي على اداء دوره الطبيعي في التخلص من الملوثات وخاصة العضوية منها بالعمليات الطبيعية الامر الذي يؤدي حدوث خلل في النظام البيئي [13].

وبمفهوم ادق يكون معنى التلوث كل ما يسهم في تغير اي عنصر من عناصر البيئة سواء كان هذا العنصر كائناً حياً كالأنسان والحيوان والنبات او مكوناً طبيعياً غير حي كالهواء والماء والتربة والاشعاع وغيرها وفيما يأتي عرض موجز لأهم مصادر تلوث الهواء والاضرار الصحية والبيئية الناجمة عن هذه الملوثات[9].

6.2. مصادر ملوثات الهواء (Sources of air pollutants)

بشكل عام تكون مصادر تلوث الهواء على نوعين مصادر صناعية وطبيعية , اما المصادر الصناعية فأنها تكون مصادر متحركة او ثابتة وتشمل كمحطات توليد الطاقة الكهربائية ومصانع الاسمنت والمعادن ومنشأت صناعة النفط والغاز الطبيعي التي تساهم هذه المصادر في زيادة خطورة التلوث بدرجات مختلفة فعلى سبيل المثال صناعة النفط تؤدي الى تلوث الهواء بأكاسيد الكبريت والنيتروجين والامونيا واوكسيد الكاربون وكبريتد الهيدروجين وغيرها من الغازات كما تنطلق بعض المواد المركبات العضوية الطيارة في الهواء والمحيط بمستويات البترول ومصانع الاصباغ ...الخ [9]

لذلك نسعى في هذا البحث الى اثبات صحة هذه الفكرة الشائعة بفحص مستويات التلوث بالنسبة الى حجم انتاج النفط على وجه التحديد من خلال البيانات المقدمة من شركة مصافي الوسط .

6.3. مراحل اجراء الجانب العملي (Stages of conducting the practical side):

* **المرحلة الاولى: اسلوب جمع البيانات (Data collection method)**

تضمن البحث الحصول على بيانات مناخية من موقع (weather history for KQTZ) على شبكة الانترنيت بأعتباره من افضل المواقع التي توفر سجلات الطقس التاريخية بسهولة لجميع البلدان حسب الرمز البريدي الخاص بكل منطقة (مدينة). اما بيانات ملوثات الهواء فقد تم الحصول عليها من شركة مصافي الوسط التي تمثل القياسات اليومية لمركبات التلوث البيئي التي تستند للمدة ايلول 2018 الى كانون الاول 2020 وقد تم استحصال موافقة شركة مصافي الوسط (مصفى الدورة). ومجموعة البيانات هذه تتعقب العناصر المسببة للتلوث الجوي من خلال القياسات المتكررة للمركبات مثل مركبات الكاربون (CO_X) ومركبات الكبريت (SO_X) ومركبات النتروجين (NO_X) فضلاً عن المتغيرات التفسيرية بما في ذلك سرعة الرياح , درجات الحرارة ,كميات النفط المنتجة مع الاشارة الى انه تم اخذ قياسات البيانات للعوامل المناخية المسجلة في نفس الساعة التي تم فيها قياس كل الملوثات في المصفى.

* **المرحلة الثانية: وصف البيانات (Description of the data)**

بالنظر لخلفية البيانات فقد تألفت مجموعة البيانات على ذلك (9) سمة مميزة اذ كانت اثنين فئوية (التاريخ، الوقت) وكانت متغيرين الاستجابة هما:

- ✓ ثنائي اوكسيد الكاربون (CO_2) :يعتبر احد اهم عناصر الهواء الطبيعية غير ملوثة ولكن زيادة الطبيعية في نسبته سيحدث كارثة لمايسببه هذا الغاز من ظاهرة الاحتباس الحراري .
- ✓ كبريتيد الهيدروجين (H_2S) : الذي ينتج اثناء عمليات تقطير البترول الذي يحتوي على الكبريت واثناء التحلل اللاهوائي للمواد العضوية في محطات التنقية.

وخمسة متغيرات تفسيرية هما متوسط درجات الحرارة لكل ساعة بالـ (C^0) ومتوسط نقطة الندى ومتوسط الرطوبة لكل ساعة بالدرجة المئوية ومتوسط سرعة الرياح بالـ (كم / ساعة) ومتوسط كمية النفط الخام المستعمل في عملية التصفية($م^3$/ ساعة) .

* **المرحلة الثالثة: ترميز البيانات (coding of the data)**

بما ان تم قياس مستوى بيانات التلوث لمدة اكثر من سنتين (28) فأن عدد القياسات يكون صغيرا نسبياً لتحديد مسارات معقدة في متغيرات الاستجابة وبناء على ذلك تم افتراض أنموذج logit للعينة الذي بموجبه تم تحويل البيانات الى استجابات ثنائية للتنبؤ بنتائج التلوث، وتمت معالجة البيانات استناداً الى الحدود القصوى المسموح بها لملوثات الهواء المنبعثة من مصادر الاحتراق . اذ ان تعد القيم الدنيا والمساوية الى الحدود القصوى غير ملوثة [0] وتعد القيم الاعلى من الحدود القصوى ملوثة [9] .

وان الحدود القصوى المسموح بها لتركيز كل ملوث من الملوثات التي يسمح بطرحها بموجب المعايير الوطنية ان غاز CO_2 قيمته $500\ ug/m^3$ و غاز H_2S قيمته $10\ ug/m^3$

* **المرحلة الرابعة : اختبار فرضية للأنموذج العام (Hypothesis testing of the general model)**

غالباً ما يتم اختبار معنوية المعلمات المناظرة لقبول او رفض فرضية العدم اقل 0.05 والتي تنص على ان

H_0 : الزيادة في معدل التلوث لاتعتمد على زيادة في الظروف المناخية وانتاج النفط

H_1: الزيادة في معدل التلوث تعتمد على زيادة في الظروف المناخية وانتاج النفط

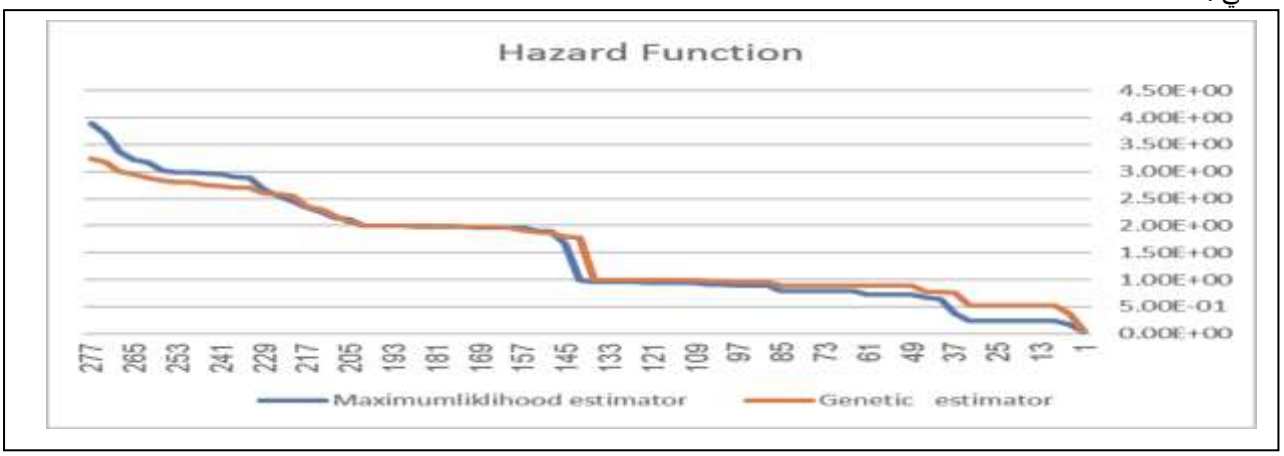
* **المرحلة الخامسة : تحليل البيانات (analysis of the data)**

سوف يتم عرض نتائج الجانب العملي وتحليلها للوصول الى مدى ملاءمة البيانات الحقيقية مع أنموذج الانحدار log-logistic المقدر من خلال اجراء الاختبارات الخاصة بمقدرات دالة المخاطرة ومقدرات الانموذج log-logistic وفيما يلي سوف يتم عرض النتائج في الجداول التي سيتم تحليلها وحسب الجداول وكالاتي :

جدول (1): تقدير معلمات أنموذج الانحدار log-logistic الثنائي بطريقة الامكان الاعظم والخوارزمية الجينية

Estimated Parameter	$\widehat{h}_{(y)_{MLE}}$	$\widehat{h}_{(y)_{G.E}}$
$\hat{\beta}_0$	-1.014	-3.941
$\hat{\beta}_1$	-0.073	2.041
$\hat{\beta}_2$	-0.091	-0.283
$\hat{\beta}_3$	-0.004	0.331
$\hat{\beta}_4$	0.015	0.815
$\hat{\beta}_5$	0.093	0.077

نلاحظ في جدول رقم (1) تنفيذ مقدرات أنموذج الانحدار log-logistic من خلال تحديد متجه المعلمات $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$ وتعويضها مع المتغيرات التوضيحية نحصل على مقدرات أنموذج الانحدارlog-logistic الثنائي .



شكل (1): مقدرات دالة المخاطرة لأنموذج الانحدار log-logistic الثنائي بطريقة الامكان الاعظم والخوارزمية الجينية

نلاحظ من شكل رقم (1) ان سلوك مقدرات دالة المخاطرة لطريقتين ظهرت بشكل متزايد لأنموذج الانحدار log-logistic الثنائي.

جدول (2): المعلمات المقدرة لدالة المخاطرة والخطأ المعياري لكافة المتغيرات التوضيحية لأنموذج الانحدار log-logistic الثنائي بطريقة الامكان

Estimated parameter	$\hat{h}_{(y)_{MLE}}$	Standard error $SE(\hat{\beta}_i)$	$\frac{\hat{\beta}}{SE(\hat{\beta}_i)}$	*significations*
$\hat{\beta}_0$	-1.014	2.034	*-0.499*	*N0N-Sig*
$\hat{\beta}_1$	-0.073	0.032	*-2.285*	*Sig*
$\hat{\beta}_2$	-0.091	0.066	-1.361	*N0N-Sig*
$\hat{\beta}_3$	-0.004	0.004	*-0.444*	*N0N-Sig*
$\hat{\beta}_4$	0.015	0.009	1.598	*N0N-Sig*
$\hat{\beta}_5$	0.093	0.035	2.423	*Sig*

نلاحظ في جدول رقم (2) تقديرات معلمات انموذج وقيم الخطأ المعياري لكل معلمة مقدرة والنتائج التي تم الحصول عليها وهي تتبع التوزيع الطبيعي القياسي بمستوى دلالة (α=0.05) وتقارن مع قيمة Z الجدولية $\mp 1.96 = Z_{\frac{1}{2}(1-0.05)}$ والعمود الاخير يمثل معنوية المتغيرات التوضيحية (اقل من 0.05) قيد الاختبار يكون معنوياً . وعليه تم ازالة $(\hat{\beta}_0, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ لانها لا تؤثر بشكل كبير في الانموذج المقدر اي ان المتغيرات (متوسط نقطة الندى، متوسط الرطوبة، متوسط سرعة الرياح) ليس لها اي تأثير، بينما$(\hat{\beta}_1, \hat{\beta}_5)$ (متوسط درجة الحرارة ومتوسط كمية النفط المنتج) لها تأثير معنوي على غاز (CO_2) و(H_2S) في انتشاره بكميات كبيرة في الهواء. وفيما اذا ذلك ان قوة العلاقة بين المتغيرات التوضيحية ومتغير الاستجابة من خلال احتساب احصاءة R^2 ظهرت قيمتها تساوي 0.32 وهذا يعني ان كمية النفط المنتج ومتوسط درجات الحرارة تفسر الاختلافات في الاستجابات المشاهدة لحالات (CO_2) و(H_2S) بنسبة 32 % .

جدول (3): المعلمات المقدرة لدالة المخاطرة والخطأ المعياري لكافة المتغيرات التوضيحية لأنموذج الانحدار log-logistic الثنائي بطريقة الخوارزمية الجينية

Estimated parameter	$\hat{h}_{(y)_{G.E}}$	Standard error $SE(\hat{\beta}_i)$	$\frac{\hat{\beta}}{SE(\hat{\beta}_i)}$	*significations*
$\hat{\beta}_0$	-3.941	22.400	*-0. 170*	*N0N-Sig*
$\hat{\beta}_1$	2.041	0.651	*3.138*	*Sig*
$\hat{\beta}_2$	-0.283	0.583	-0.486	*N0N-Sig*
$\hat{\beta}_3$	0.331	0.601	*-0.501*	*N0N-Sig*
$\hat{\beta}_4$	0.815	0.037	*21.284*	*Sig*
$\hat{\beta}_5$	0.077	0.020	*3.818*	*Sig*

نلاحظ في جدول رقم (3) تقديرات المعلمة وقيم الخطأ المعياري لكل معلمة مقدرة والنتائج التي تم الحصول عليها وهي تتبع التوزيع الطبيعي القياس بمستوى دلالة (α=0.05) وتقارن مع قيمة Z الجدولية $\mp 1.96 = Z_{\frac{1}{2}(1-0.05)}$ والعمود الاخير يمثل معنوية المتغيرات التوضيحية (اقل من 0.05) قيد الاختبار يكون معنوياً .

وعليه تم ازالة ($\hat{\beta}_0, \hat{\beta}_2, \hat{\beta}_3$) لانها لا تؤثر بشكل كبير في الانموذج المقدر اي ان المتغيرات (متوسط نقطة الندى , متوسط الرطوبة) ليس لها اي تأثير، بينما ($\hat{\beta}_1, \hat{\beta}_4, \hat{\beta}_5$) (متوسط درجة الحرارة ومتوسط سرعة الرياح ومتوسط كمية النفط المنتج) لها تأثير معنوي غاز (CO_2) و(H_2S) على انتشاره بكميات كبيرة في الهواء .وفيما اذا ذلك ان قوة العلاقة بين المتغيرات التوضيحية ومتغير الاستجابة من خلال احتساب احصاءة R^2 ظهرت قيمتها تساوي 0.48 وهذا يعني ان كمية النفط المنتج ومتوسط درجات الحرارة ومتوسط سرعة الرياح تفسر الاختلافات في الاستجابات المشاهدة لحالات (CO_2) و(H_2S) بنسبة 48 % .

7. الاستنتاجات (Conclusions)

1. بالنسبة لأنموذج الانحدار log-logisticالثنائي بطريقة الامكان الاعظم بأعتماد على الاحصاءات المرتبة ذات القيم المسجلة العليا تبين من خلال اختبار قيم (Z) ان العوامل التي كان لها تأثير معنوي واقل من 0.05 على متغير الاستجابة لغاز (CO_2) و(H_2S) هي متوسط درجات الحرارة وكمية النفط المنتج .
2. كان عامل متوسط درجة الحرارة ومتوسط سرعة الرياح ومتوسط كمية النفط المنتج هو الاكثر تأثيرا وبمعنوية اقل من 0.05 على متغير الاستجابة لغاز (CO_2) و(H_2S) باستعمال اختبار (Z) لأنموذج الانحدار log-logistic الثنائي بطريقة الخوارزمية الجينية .
3. اثبتت كلتا الطريقتين (MLE,G.E) ان المتغيرين كمية النفط المنتج ومتوسط درجات الحرارة يعتبران من العناصر الرئيسية على انتشار غاز (CO_2) و(H_2S) وهذا ماتم تاكيده من قبل المهندسين الكيميائيين بهذا المجال .
4. اظهرت مقدرات دالة المخاطرة بشكل متزايد بازدياد انتشار نسبة غاز(CO_2) و(H_2S) في الهواء

8. التوصيات والدراسات المستقبلية(Recommendations and future studies)

1. استعمال الطرائق التي تم عرضها لتقدير دالة المخاطرة ومعلمات أنموذج الانحدارlog-logistic المتعدد والترتيبي والذي يمكن استعماله لنمذجة البيانات في مجالات التلوث البيئي مثل تلوث الماء والتربة والتلوث الاشعاعي .
2. اجراء دراسات مستقبلية تتناول الاعتماد على طرائق التقدير الاعتيادية والمحسنة بالخوارزمية الجينية المذكورة في البحث والمقارنة بينهما في حالة وجود قيم شاذة في البيانات.
3. استعمال متغيرات اخرى غير التي ذكرت في البحث مثل متغيرات الاستجابة التي تشمل غاز اكاسيد نتروجين والامونيا واوكسيد الكاربون والاوزون.
4. اضافة بيانات من اجل تسهيل تطوير أنموذج تنبؤ مثالي واقتراح ان البيانات اليومية للمدة من 10 الى 20 سنة تسفر عن نموذج اكثر دقة.
5. ينبغي ان تنظر شركة مصافي الوسط ايضاً في مواصلة وتحسين عمليات جمع البيانات عن الملوثات المحتملة في مختلف الظروف المناخية مثل النفايات السائلة والصلبة التي يتم الحصول عليها من خلال عمليات التكرير وذلك لتسهيل تقييم المخاطر على مسار الهواء في الدراسات المستقبلية المماثلة .
6. واحدة من الدراسات الرئيسية هي لفحص المستويات المستمرة لانبعاثات الكاربون الناجمة عن عمليات التكريرفي مصافي النفط بالنسبة لكميات النفط المنتجة.
7. ان تناول هذا الموضوع من شأنه تسهيل التعرف على امكانية تلوث الكاربون عن طريق مصافي النفط بأعتبار ان المتغيرات البيئية ايضاً تديم التلوث وتحفز التفاعلات بين الملوثات.

المصادر (References)

- **المصادر الأجنبية (Foreign References)**

[1] A. Misra, "Portfolio optimization of commercial Banks- An Application of Genetic Algorithm", EJBM, Vol. (5), pp. 120-129 ,(2013).

[2] A. Soliman, A. Abd – Ellah , A. H. and, K. S. Sultan ,"Comparison of estimates using record statistics from Weibull model: Bayesian and non-Bayesian approaches". Computational Statistics and Data Analysis, Vol. 51:2065-2077 ,(2006).

[3] A.M. Abd-Elmonem, Teamah ,et.al "Heavy-Tailed Log-Logistic Distribution: Properties, Risk Measures and Applications", Master Thesis, Department of Mathematics, Faculty of Science, Tanta University, Tanta 31527, Egypt , (2021).

[4] B. Steve ,"Log-logistic Regression Models for Survival Data", Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. (32), No. 2,pp. 165-171 , (1983).

[5] D. Raghupathikumar, K. Raja, "Genetic algorithm based scheduling of an input queued switch", International Journal of Soft Computing, Vol. (6), No. 5, 2011.

[6] E. Demir , Ö. Akkus, " An Introductory Study on How the Genetic Algorithm Works in the Parameter Estimation of Binary Logit Model", IJS:BAR, pp.162-180 , (2015).

[7] H. Liu, C. Ong, "Variable selection in clustering for marketing segmentation using genetic algorithms", Expert Systems with Applications, Vol. (34), No. 1, (2008).

[8] H., Samah Sabah and AL Doori, "Estimating the survival and risk functions of a log-logistic distribution by using order statistics with practical application", Int. J. Nonlinear Anal. Appl., Vol. (13), No. 1, (2022).

[9] I. Youssef Khalil, "Some methods for estimating the conditional logistic regression model in the case of longitudinal data and their application in environmental pollution." A master's thesis in statistics submitted to the College of Administration and Economics at the University of Baghdad (2017).

[10] J. Pasia, A. Hermosilla, et al., "A useful tool for statistical estimation: genetic algorithms", JSCS, Vol. (75), pp. 237 – 251, (2005).

[11] K.H. Lee, K.W. Kim, "Performance comparisons of particle swarm optimization and genetic algorithm for inverse surface radiation problem", Vol. (88), pp. 330-337, (2015).

[12] M. Burkschat, E. Cramer, and U. Kamps, "Dual generalized order statistics " , International Journal of Statistics, Vol. LXI, No. 1, pp. 13-26 (2003) ,

[13] M. S. Ahmed, "Forecasting of Environmental Pollution Using Box-Genghis Models for Al-Waziriya Station", Diploma Research in Statistics, College of Administration and Economics, University of Baghdad, (2015).

[14] Ö. Akkus, E. Demir, "Comparison Some Classical And Meta-Heuristic Optimization Techniques in The Estimation Of The Logit Model Parameters", IJAR, pp.1026-1042, (2016).

[15] P. McCullagh, "Regression models for ordinal data" J. R. Statist. Soc. B, Vol. 42, No. 2, 109-142,(1980).

[16] Q. Abdul, G. Enas,"Estimating Lomax Distribution Parameters Using Generalized Ordered Statistics", A master's thesis in statistics, College of Computer Science and Mathematics, University of Mosul, (2018)

[17] R. Sarah Adel ,"The use of the genetic algorithm in estimating the parameters of the binary logistic regression model with practical application." A master's thesis in statistics, College of Administration and Economics, University of Baghdad, (2019) .

[18] S. Hadji, and et al., "Theoretical and experimental analysis of genetic algorithms based MPPT for PV systems", International Conference on Technologies and Materials for renewable energy, Environment and Sastainability, pp. 772-787 (2015).

[19] S. Sefian , M. Benbouziane,"Portfolio Selection Using Genetic Algorithm" , MPRA , pp. 1-12 ,(2012) .

[20] U. Kamps "A Concept of Generalized Order Statistics". Teubner, Stuttgart,Germany. Journal of Statistical Planning and Inference, Vol. 48: 1-23(1995).

Journal of AL-Rafidain University College for Sciences (2023); Issue 54; 123 – 133

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

Estimating the Risk Function Using the Log-Logistic Regression Model for Environmental Pollution Data

Prof. Dr. Entsar A. Fadam AL-Doori	Samah S. Hassan
entsar_arebe@coadec.uobaghdad.edu.iq	samahsabah1993@gmail.com
Department of Statistics - College of Administration and Economics - Baghdad University, Baghdad, Iraq	
Assist. Prof. Dr. Iqbal M. Alwan	
iqbal.alwan@coadec.uobaghdad.edu.iq	
Department of Statistics - College of Administration and Economics - Baghdad University, Baghdad, Iraq	

Article Information

Article History:
Received: December, 15, 2022
Accepted: March, 3, 2023
Available Online: December, 31, 2023

Keywords:
Log-logistic distribution, generalized ordered statistics model, environmental pollution data, maximum likelihood method, genetic algorithm method, log-logistic regression model, risk function, GLM

Correspondence:
Prof. Dr. Entsar A. Fadam AL-Doori
entsar_arebe@coadec.uobaghdad.edu.iq

https://doi.org/10.55562/jrucs.v54i1.583

Abstract

Log-Logistic distribution contains a non-linear risk function (non-monotonous), it can be used to model some survival data sets and shield them from the risks associated with environmental pollution in the modern era, which led to the spread of various diseases, such as cancer, hyperactivity, breathing issues, and other diseases, that have become prevalent in our society. A Log-Logistic regression model has been described in which the risk function of each sample that relates to time and also provides a linear model of survival probability at any specified time. In this research, the risk function of the Log-Logistic regression model was estimated using the maximum likelihood approach and the genetic algorithm method in the presence of generalized order statistics. These techniques were applied to air pollution data obtained from the Central Refineries Company in Baghdad (Al-Dora Refinery) daily environmental pollution compounds, which are based on time for the years 2018 to 2020.