المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

اختيار المتغيرات في نموذج الانحدار اللوجستي باستخدام خوارزمية اليراعات المضيئة المعدلة

هبة سليمان داؤد

قسم الاحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات ، جامعة الموصل ، الموصل ، العراق

معلومات النشر

تاريخ المقالة:
تم استلامه في 25 كانون الاول 2023
تم القبول في 12 شباط 2024
متاح على الانترنيت 1 حزيران 2024

الكلمات الدالة:
اختيار المتغيرات، خوارزمية اليراع المضيئة، المحاكاة، انموذج الانحدار الاسي.

المراسلة:
هبة سليمان داؤد
heba.sulaiman82@uomosul.edu.iq

الخلاصة

يعتبر نموذج الانحدار اللوجستي هو الأكثر استخدامًا في العديد من التطبيقات ، وهو من النماذج الرئيسية في عائلة النماذج الخطية المعممة.. وكغيره من سائر نماذج الانحدار، قد يحتوي النموذج على متغيرات مستقلة كثيرة ما يؤثر سلباً على دقة النموذج وبساطته في تفسير النتائج. تهدف هذه الدراسة إلى استخدام خوارزمية اليراع المضيئة المعدلة ومقارنتها مع طرائق اخرى في اختيار المتغيرات في نموذج الانحدار الاسي باستخدام المحاكاة والبيانات الحقيقة . وأظهرت النتائج أنه بالمقارنة مع الطرائق الأخرى المستخدمة سابقا، فإن الاسلوب المقترح يؤدي أداء أفضل ويساعد على خفض متوسط مربع الخطأ للنموذج.

DOI 10.33899/IQJOSS.2024.183255

مقدمة

تعالج التقنيات الجديدة النمو الهائل للبيانات، اذ تساعد هذه التقنيات الباحثين على نقل كميات هائلة من البيانات إلى المعلومات. قد تحتوي البيانات الضخمة على متغيرات غير ذات صلة أو زائدة عن الحاجة. ولذلك يفضل الباحثون اختيار المهم من هذه المتغيرات عن طريق اختيار مجموعة فرعية صغيرة من المتغيرات المهمة المتوفرة من مجموعات البيانات.

تعتبر دراسة أي مشكلة أو ظاهرة من المجالات الاقتصادية، الاجتماعية، الطبية أو غيرها، من أهم أسس البحث العلمي. فالغاية الرئيسية من دراستها هي تحديد المعادلة الرئيسية التي تمثل تلك الظاهرة بدقة، وذلك عن طريق جمع البيانات المتعلقة بها من مختلف المصادر المتاحة. ومن ثم يتم تحليل تلك البيانات باستخدام تقنيات الإحصاء والتحليل الرياضي لتحديد العلاقات بين المتغيرات المختلفة وتصميم نماذج إحصائية تصف تلك العلاقات. وهذا يشكل المدخل الأساسي لفهمها بشكل أعمق وتحديد معالمها الرئيسية. ويشار إلى أن هذه العملية في علم الإحصاء بنمذجة الظواهر(Månsson, 2013). ومن بين جميع

نماذج الانحدار الخطي المعممة، يمكن القول أن نموذج الانحدار اللوجستي هو أحد أشهر هذه النماذج ، حيث يتم استخدامه بشكل واسع في العديد من التطبيقات.

الانحدار اللوجستي هو احد نماذج المعتمد على التصنيف الثنائي من خلال التنبؤ باحتمالية حدوث نتيجة أو حدث. يقدم النموذج نتيجة ثنائية أو ثنائية التفرع تقتصر على نتيجتين محتملتين: نعم/لا، 1/0، أو صحيح/خطأ. يقوم الانحدار اللوجستي بتحليل العلاقة بين واحد أو أكثر من المتغيرات المستقلة ويصنف البيانات إلى فئات منفصلة. يتم استخدامه على نطاق واسع في النمذجة التنبؤية، حيث يقوم النموذج بتقدير الاحتمال الرياضي لما إذا كان المثيل ينتمي إلى فئة معينة أم لا (Alharthi, Lee, & Algamal, 2021).

غالبية البيانات في الواقع التطبيقي الحقيقي تحتوي على مشاكل مثل مشكلة العدد الكبير من المتغيرات المستقلة المدروسة، وهي من المشاكل المعروفة لدى الباحثين الإحصائيين، وتؤثر سلباً على عملية التقدير. في بعض الحالات، يمكن أن تؤدي هذه المشكلة إلى تجاهل بعض المتغيرات التوضيحية المهمة. حيث اصبحت الاساليب التقليدية لاختيار المجموعات الجزئية غير جيدة في اداء وظيفتها حيث اصبحت اكثر تكلفة في حسابها ، اضافة الى ذلك فان معايير المعلومات لاختيار المتغيرات مثل معيار أكاكي للمعلومات (Akaike information(AIC)) ومعيار بيز للمعلومات (Bayesian information criterion (BIC)) اصبحت غير عملية في اختيار المتغيرات التوضيحية وذلك بسبب تعقيدها الحسابي الذي ينمو بشكل طردي مع ازدياد عدد المتغيرات التوضيحية (Özkale & Arıcan, 2018).

يهدف هذا البحث إلى توظيف خوارزمية اليراعات المضيئة المعدلة ومقارنتها مع طرائق إختيار المتغيرات التوضيحية في أنموذج الانحدار اللوجستي الاخرى باستخدام المحاكاة والبيانات الحقيقة، من خلال تسليط الضوء على عدد من العوامل التي قد تؤثر على جودة هذه الطرائق ووجوب استخدامها ضمن شروط معينة دون غيرها من الطرائق.

يعتبر العالم Yang اول من استخدم خوارزمية اليراعات المضيئة عام 2007 وطورها في عام 2009 لاجراء التصنيفات للبيانات الهندسية وبعدها تم استخدامها في الكثير من المجالات الاحصائية مثل تنقيب البيانات و في جال تعلم الالة وغيرها من المجالات.

2. نموذج الانحدار اللوجستي (LRM)

يعرف الانحدار اللوجستي على انه نوع من انواع نماذج الانحدار اللاخطية الذي تكون فيه العلاقة بين المتغير التابع (الاستجابة) (y_i) ومجموعة من المتغيرات التوضيحية (Özkale & Arıcan, 2016) علاقة غير خطية $(x_1, x_1, x_2 \dots x_k)$، اذ يكون فيها المتغير التابع (الاستجابة) متغير نوعي (Varathan & Wijekoon, 2018). قد يأخذ المتغير التابع في نموذج الانحدار اللوجستي صفتين فقط ويرمز لهاتين الصفتين بـ (0) او(1)، وهو ما يطلق عليه بالانحدار اللوجستي الثنائي (Binary logistic regression). اما فيما يخص المتغيرات التوضيحية، فيمكن ان تكون هذه المتغيرات مستمرة او متقطعة سواءا وصفية كانت او عددية.

يبنى نموذج الانحدار اللوجستي على فرض أساسي هو أن المتغير التابع الذي نهتم بدراسته هو متغير ثنائي الصفة ويتبع توزيع برنولي وفق الدالة الاحتمالية المعرفة بالصيغة الاتية (*Özkale & Arıcan, 2018*)

$$p\,(Y = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{1}$$

اذ ان: π_i : تمثل احتمال حدوث الاستجابة عندما $y_i = 1$ و $1 - \pi_i$: تمثل احتمال عدم حدوث الاستجابة عندما $y_i = 0$.

يمكن تعريف الاحتمال (π_i) رياضيا بدلالة المتغيرات التوضيحية والدالة اللوجستية وكما في الصيغة الاتية :

$$\pi_i = \frac{e^{X_i\beta}}{1+e^{X_i\beta}} \tag{2}$$

اذ ان: $\underline{\beta}$:متجه من المعلمات أبعاده (p×1) و $X_i = \{x_{i1}, x_{i2}, \dots x_{ip}\}$: متجه صفي من المتغيرات التوضيحية أبعاده (1×p).

يكون الهدف الرئيس من الانحدار اللوجستي الثنائي هو تفسير التغير في قيم المتغير التابع من خلال تفسير حدوث الاستجابة باحتمال (π_i) او عدم حدوث الاستجابة باحتمال $(1 - \pi_i)$. بناءاً على ذلك، وكما هو معروف عند بناء نموذج الانحدار، فان من خلال المعادلة الاخيرة (2) يتضح ان العلاقة بين المتغير التابع والمتغيرات التوضيحية علاقة غير خطية وغالبا ما تأخذ الدالة اللوجستية شكلاً منحنياً. ويلجأ الاحصائيون غالبا إلى التحويل الخطي لهذه النماذج لإزالة انحناءات معلماتها وذلك لتأثير هذه الانحناءات السلبي في حالة وجودها على خصائص المقدرات اذ بالإمكان افتراض علاقة معينة تربط بين المتغير التابع والمتغيرات التوضيحية الاخرى، لذلك تم اقترح تحويل دالة اللوجت (Logit Function) التي تقوم بتحويل علاقة الانحدار اللاخطية بين المتغيرات التوضيحية ودالة احتمال الاستجابة (π_i) في نموذج الانحدار اللوجستي الى علاقة انحدار خطي، وذلك من خلال اخذ اللوغاريتم الطبيعي للمقدار $(\frac{\pi_i}{1-\pi_i})$ كما مبين في المعادلات (Steyerberg, Borsboom, van Houwelingen, Eijkemans, & Habbema, 2004; Varathan & Wijekoon, 2018) الاتية:

$$\text{logit}(\pi_i) = \ln\frac{\pi_i}{1-\pi_i} \tag{3}$$

$$\text{logit}(\pi_i) = \ln\frac{\frac{e^{X_i\beta}}{1+e^{X_i\beta}}}{\frac{1}{1+e^{X_i\beta}}} \tag{4}$$

$$\text{logit}(\pi_i) = \ln\left(e^{X_i\beta}\right) = X_i\beta \; = \left(Bo + \sum_{j=1}^{k} Bj\, xij\right) \tag{5}$$

اذ ان: $\beta_0, \beta_1, \dots, \beta_p$: معالم مجهولة يتم تقديرها.

دالة الإمكان الأعظم لنموذج الانحدار اللوجستي الذي يتبع توزيع برنولي تكون بالصيغة الاتية

$$yi \sim \text{Bernoulli}(\pi i) \tag{6}$$

اذا:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\Pi_I)^{1-Y_I} \tag{7}$$

$$L(\beta) = \prod_{i=1}^{n}(1-\Pi_I)\exp\left[\sum_{i=1}^{n} y_i \ln\left(\frac{\Pi_I}{1-\Pi_I}\right)\right] \tag{8}$$

وحسب خاصية التحويل اللوجستي (دالة اللوجستك) فان

$$\ln\frac{\pi_i}{1-\pi_i} = X_i\beta \tag{9}$$

فان $L(\beta)$ تساوي:

$$L(\beta) = \prod_{i}^{n}(1-\pi_i)exp\left[\sum_{i=1}^{n} y_i(X_i\beta)\right] \tag{10}$$

وبأخذ اللوغاريتم إلى دالة الإمكان.

$$\ln L(\beta) = \sum_{i=1}^{n}\ln(1-\pi_i) + \left[\sum_{i=1}^{n} y_i(X_i\beta)\right] \tag{11}$$

اذا:

$$\ln L(\beta) = \left[\sum_{i=1}^{n} y_i(X_i\beta)\right] - \sum_{i=1}^{n}\ln(1+e^{X_i\beta}) \tag{12}$$

وبأخذ المشتقة الأولى الى لوغاريتم دالة الإمكان الأعظم ثم مساواة المشتقة بالصفر.

$$\frac{\partial \log L(\beta, X)}{\partial \underline{\beta}} = 0 \tag{13}$$

فان المعادلات الناتجة من المشتقة الأولى هي معادلات غير خطية والتي ليس لها حل واضح، لذلك يتم حل هذه المعادلات عن طريق الطرق العددية والتي منها الطريقة العددية الأكثر شيوعا هي خوارزمية نيوتن رافسون .

3. خوارزمية اليراعات المضيئة Firefly Algorithm (FFA)

في السنوات الأخيرة أصبح الاهتمام متزايد بتصميم خوارزميات التحسين المستوحاة من الطبيعة وتطويرها، حاول الباحثون إيجاد الالهام من مصادر مختلفة في الطبيعة مثل النحل والنمل واليراعات والاسماك والطيور والنباتات وأنظمة الامواج والأنهار. يعد ذكاء السرب أداة مهمة لحل العديد من المشكلات المعقدة في البحث العلمي، إذ تمّت دراسة خوارزميات ذكاء السرب على نطاق واسع حيث تم تطبيقها بنجاح على مجموعة متنوعة من مشكلات التحسين المعقدة نظراً لتمتعها بالبساطة والمرونة والكفاءة العالي(Yang, 2010) .

تعتمد معظم خوارزميات التحسين المستوحاة من الطبيعة على ذكاء السرب، وتشكل الخوارزميات القائمة على ذكاء السرب جزءاً كبيرا من الخوارزميات المعاصرة، وأصبحت هذه الخوارزميات مستخدمة على نطاق واسع في التحسين وتحليل البيانات وكذلك في التعلم الآلي والذكاء الإصطناعي. وتعُد خوارزمية اليراعات المضيئة(Firefly Algorithm FA) واحدة من أحدث أساليب ذكاء السرب الجديدة واقوى خوارزميات التحسين التي تم تطويرها لأول مرة من قبل الباحث Yangفي بداية عام 2008.

أثبتت الخوارزمية أنها فعالة وذات أداء جيد في حل مشكلات التحسين المختلفة. تم إيجاد خوارزمية اليراعات من محاكاة السلوك الاجتماعي لليراعات المضيئة على أساس جاذبية الفلاش (الأضواء الساطعة) من خلال تمثيل ميزة بعض الخصائص الوامضة لليراعات وكيفية التفاعل معها، إذ أن وميض اليراعة هو نظام إشارة يستخدم لجذب يراعة أخرى (Long Zhang, Shan, & Wang, 2016; Li Zhang, Srisukkham, Neoh, Lim, & Pandit, 2018).

حيث يمكن حساب المسافة بين اثنين من اليراعات في المواقع بالمسافة الديكارتية والتي يمكن حسابها بإستخدام المعادلة الآتية :

$$r_{ij} = \|xi - xj\| = \sqrt{\sum_{d=1}^{D}\left(x_{id} - x_{jd}\right)^2} \tag{14}$$

إذ أن x_i : هو موقع اليراعة i إذ $x_i = \{x_{1i}, x_{2i}, ..., x_{id}\}$ ، وأن x_j : موقع اليراعة j إذ $x_i = \{x_{1i}, x_{2i}, ..., x_{id}\}$ ، D عدد الابعاد وأن $d \in D$

يمكننا تلخيص آلية عمل خوارزمية اليراعات (FA) بالخطوات الآتية(Yang, 2010) :

1- جميع اليراعات للجنسين، إذ يمكن أن تنجذب كل يراعةٍ إلى كل اليراعات الأخرى. إذ أنّ اليراعات الأقل جاذبية (أشراقا) تنجذب إليها اليراعات الأكثر جاذبية(إشراقاً).

2- تتناسب جاذبية اليراعة مع شدة الضوء الذي يتناقص كلما زادت المسافة عن اليراعات الأخرى.

3- يتم تحديد جاذبية اليراعة من خلال موقعها داخل مساحة البحث.

4- تؤدي القيمة الأفضل لوظيفة اللياقة في موقع معينٍ إلى زيادة جاذبية اليراعة.

لكل فراشة شدة ضوء أو سطوع يتم إستخدام قيمتهُ لتقييم جودتها. إن سطوع اليراعة i في موقع معين x نستطيع أن نشير إليه بالآتي:

$$I(x_i) = f(x_i) \tag{15}$$

حيث أن شدة ضوء اليراعة تتناسب طردياً مع سطوعها وترتبط بالقيم الموضوعية. عند المقارنة بين اليراعات، تنجذب اليراعة التي لها شدة ضوء منخفضة نحو اليراعة الأخرى ذات الضوء الأعلى، شدة ضوء اليراعة تعتمد على I_o من الضوء المنبعث من اليراعة والمسافة r_{ij} بين زوج من اليراعات. يمكن وصف شدة الضوء $I(r)$ من خلال دالة متناقصة بشكل رتيب لـ r_{ij} والتي يمكن صياغتها كالآتي:

$$I(r) = I_o e^{-(\gamma rij)2} \tag{16}$$

γ: هو عامل امتصاص تأثير الضوء.

ونظراً لأن الجاذبية لكل فراشة تتناسب مع شدة الضوء التي تراها اليراعات المجاورة، لذلك يجب السماح للجاذبية بالتنوع بإختلاف درجة الإمتصاص، حيث يمكن تحديد الشكل الرئيسي لتباين الجاذبية Z بالمعادلة التالية : (Xu, Yu, Chen, & Zuo, 2018)

$$Z(r) = Z_o e^{-(\gamma rij)2} \tag{17}$$

إذ أن $Z(r)$ تمثل دالة جاذبية اليراعة عند المسافة r وZ_0 هي الجاذبية الأولية لليراعة عند مسافة (r = 0) ويمكن أن تكون ثابتة. عند التنفيذ Z_0 تساوي الواحد ولمعظم المشاكل. حيث يتم تحديث الحركة للفراشات حسب المعادلة الآتية:

$$x_i^{(t+1)} = x_i^{(t)} + Z_o e^{-\gamma r^2{}_{ij}}(x_j^{(t)} - x_i^{(t)}) + \alpha_t \boldsymbol{\varepsilon}_i^t \tag{18}$$

إذ أن α_t : هو معامل التوزيع العشوائي. $\boldsymbol{\varepsilon}_i^t$: متجه لأرقام عشوائية مأخوذة من توزيعUniform.

يعتمد تأثير هذه الحركة العشوائية في المعلمة α_t فيما إذا تم إختياره ليكون كبيراً فإن الحل x_i سيتحرك بشكل عشوائي مبتعداً عن الموقع، بخلاف إذا كان α_t صغيرة جداً، فستتحرك في الموقع وقد تصبح ضئيلة مقارنة بالحركة نحو اليراعات الأكثر إشراقاً.

في BFFA ، تُستخدم وظيفة النقل لتعيين مساحة بحث مستمرة إلى مساحة ثنائية ، وتم تصميم عملية التحديث لتبديل مواقع النجوم بين 0 و 1 في مساحات البحث الثنائية. من أجل بناء هذا المتجه الثنائي ، وظيفة النقل في المعادلة. (18) يمكن استخدامها ، حيث يكون الحل الجديد مقيدًا بالقيم الثنائية فقط

$$x_i^t = \begin{cases} 1 & \text{if} \quad T(x) > \alpha \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

اذ ان $\alpha \in [0,1]$ هي عبارة عن رقم عشوائي وان $T(x)$ هي دالة تحويل. ان دالة التحويل تعرف بالشكل الاتي:

$$T_{BCSA}(x_i^t) = \frac{1}{1+e^{10(x_i^t - 0.5)}}, \tag{20}$$

في هذا البحث تم اقتراح استخدام دالة تحويل متغيرة خلال الزمن. اي ان دالة التحويل هذه سوف تتغير خلال تكرار الحل. تم هذا الاقتراح من خلال اضافة معلمة تحكم وهي θ ، اذا تحتاج هذه المعلمة الى قيمة عليا وقيمة دنيا لها من خلال المعادلة الخاصة بها وهي:

$$\theta = \theta_{\min} + (\theta_{\max} - \theta_{\min})\,\mathrm{e}^{-t}, \tag{21}$$

وعليه سوف تصبح دالة التحويل المقترحة بالشكل التالي:

$$T_{\mathrm{TV}}\left(x_i^t\right) = \frac{1}{1+e^{10(x_i^t-0.5)/\theta}}, \tag{22}$$

من أجل إتمام هدف البحث وتحقيقهُ، وبالاعتماد على هذه التقنية، فإن كل عنصر (يراعة) في المجموعة سيكون لديه d من المواقع التي تمثل عدد المتغيرات التوضيحية في انموذج الانحدار اللوجستي. بناءً على ذلك، فإن توظيف خوارزمية اليراعات المضيئة تكون وفق الخطوات التالية:

الخطوة الأولى: تحديد حجم المجموعة (عدد اليراعات) وهو 30 فراشة، حيث إن كل فراشة سيكون له متجه من عدد المتغيرات المستقلة فضلاً عن ذلك تحديد عدد التكرارات داخل خوارزمية اليراعات المضيئة حيث استقرت النتائج عند التكرار 500.

الخطوة الثانية: توليد القيم الأولية التي تحتاجها الخوارزمية، التي ستمثل القيم الأولية الافتراضية ، فإن توليدها سيكون من التوزيع المنتظم المستمر وفق الفترة [0,1].

الخطوة الثالثة: لغرض اختيار القيم المُثلى، تم الاعتماد على Fitness Function وفق الصيغة الآتية:

$$\text{Fitness Function} = \min\left[\frac{\sum_{i=1}^{n}(\mathbf{y_i} - \hat{m}(\mathbf{X}))^2}{n}\right] \tag{23}$$

الخطوة الرابعة: بالاعتماد على أقل قيمة تحصل عليها أي فراشة وفق المعادلة (22) يتم تحديث مواقع باقي اليراعات.

الخطوة الخامسة: نستمر بالحل لحين الوصول الى أعلى تكرار للخوارزمية، الذي تم تحديده بالخطوة الأولى والذي سيمثل الحل الأمثل.

x_1	x_2	....	x_{p-1}	x_p
1	0	....	1	0

الشكل 1: الية اختيار المتغيرات حسب خوارزمية اليراعات المضيئة

3- معايير تقييم طرائق اختيار المتغيرات

3-1 معايير تقييم دقة التنبؤ

اولا: خطأ التنبؤ (PE) (Prediction Error)

ويعرف بانه مربع الفرق بين القيمة الحقيقية لمتغير الاستجابة والقيمة التنبؤية المرافقة له، ويعرف رياضيا بالمعادلة التالية :

$$PE = (y - \hat{y})^T (y - \hat{y}) \tag{24}$$

وبالاعتماد على هذا المعيار يتم تحديد الطريقة الافضل التي تعطي اقل قيمة مقارنة بالطرائق الاخرى.

ثانيا: معايير تقييم دقة اختيار المتغيرات

بما ان الطرق المقترحة بصورة عامة تعمل على اختيار المتغيرات، لذلك من المهم تقييم وقياس قدرة هذه الطرائق وجودتها في كيفية اختيار المتغيرات المهمة. ولذلك، تم الاعتماد على معيارين في دراستنا لهذا الغرض وبالشكل التالي:

(1) معيار التقييم"C"

هو معيار التقييم الذي يرمز له بـ**(C)** والذي يعرف بانه عدد المعاملات الحقيقية ذات القيم الصفرية والتي تم تقديرها بشكل صحيح على انها ذات قيم صفرية.

(2) معيار التقييم "I"

معيار التقييم الذي يرمز له بـ(**I**) وهو يعرف على انه عدد المعاملات الحقيقية ذات القيم غير الصفرية والذي تم تقديرها بشكل غير صحيح على انها ذات قيم صفرية. تعتمد جودة طرائق الجزاء من ناحية معايير تقييم دقة اختيار المتغيرات على من يعطي اعلى قيمة لـ (**C**) واقل قيمة لـ (**I**) .

3- نتائج المحاكاة

لقد تم تصميم تجربة ومحاكاتها باستعمال لغة البرمجة (R) حيث تم توليد المتغير (y_i) في انموذج انحدار كاوس المعكوس، حيث تم استخدام اسلوب مونت كارلو (Mont Carlo) في المحاكاة حيث تم تعيين قيم حجم العينات (n) حيث تم استخدام ثلاث احجام من العينات وهي (30,100,150,250) وذلك لأجل دراسة المقارنة وفق العينات باختلاف أنواعها. سوف تتم المقارنة مع كل من طريقة معيار بيز ومعيار اكاكي.

اولا : تم توليد بيانات المتغير y التي تتبع انموذج الانحدار اللوجستي وكالاتي :

$$\hat{Y} = \frac{e^{x_i\hat{\beta}}}{1 + e^{x_i\hat{\beta}}}$$

ثانياً : تم توليد مصفوفة المتغيرات التوضيحية **X** ذات ابعاد $(n \times p)$ التي تتبع التوزيع الطبيعي المتعدد **(Multivariate Normal Distribution)** كالاتي :

$$\boldsymbol{X} \sim \mathbf{MN}(\boldsymbol{\mu}, \mathbf{M})$$

حيث ان **M** هي مصفوفة التباين المشترك، حيث ان $\boldsymbol{M_{ij}} = \boldsymbol{r}^{|\boldsymbol{i}-\boldsymbol{j}|}$، عندما $(i, j = 1,2, \dots, p)$ حيث ان المتغيرات التوضيحية تكون مرتبطة.

ثالثاً : تم تكرار التجربة (100) مرة وذلك لغرض تقليل التحيز في تجارب مونت كارلو (Mont Carlo).

رابعاً : تم توليد بيانات نموذج انحدار بواسون تبعا لقيم متجه معلمات الانحدار $\boldsymbol{\beta}$ الذي ابعاده $(1 \times p)$ وكانت قيم متجه معلمات الانحدار $\boldsymbol{\beta}$ كالاتي $\boldsymbol{\beta} = (\mathbf{1.8}, \mathbf{2.5}, \mathbf{1}, -\mathbf{4}, -\mathbf{7}, \mathbf{0}, \dots, \mathbf{0})^{\mathrm{T}}$ ، حيث ان المعلمات غير الصفرية عددها $q = 5$ ، وان المعلمات الصفرية تساوي $p - q$. اذ تم اعتبار p=10, 50, 100

الجداول الآتية توضح النتائج العملية:

جدول (1) : معدل معايير تقييم طرائق الاختيار عندما n=30

p	Method	PE	C	I
10	**AIC**	24.503	1	0
	BIC	22.955	2	0
	FFA	**17.722**	**5**	**0**
50	**AIC**	22.879	3	0
	BIC	21.331	3	0
	FFA	**16.098**	**5**	**0**
100	**AIC**	22.112	2	1
	BIC	20.564	3	0
	FFA	**15.331**	**5**	**0**

جدول (2) : معدل معايير تقييم طرائق الاختيار عندما n=100

p	Method	PE	C	I
10	**AIC**	23.465	1	0
	BIC	21.917	3	0
	FFA	**16.684**	**5**	**0**
50	**AIC**	21.841	3	0
	BIC	20.293	4	0
	FFA	**15.06**	**5**	**0**
100	**AIC**	21.074	2	1
	BIC	19.526	3	0
	FFA	**14.293**	**5**	**0**

جدول (3) : معدل معايير تقييم طرائق الاختيار عندما n=150

p	Method	PE	C	I
10	AIC	21.687	2	0
	BIC	20.139	3	0
	FFA	14.906	**5**	**0**
50	AIC	20.063	4	0
	BIC	18.515	4	0
	FFA	13.282	**5**	**0**
100	AIC	19.296	3	1
	BIC	17.748	3	0
	FFA	12.515	**5**	**0**

جدول (4) : معدل معايير تقييم طرائق الاختيار عندما n=250

p	Method	PE	C	I
10	AIC	20.649	1	0
	BIC	19.101	2	0
	FFA	13.868	**5**	**0**
50	AIC	19.025	3	0
	BIC	17.477	4	0
	FFA	12.244	**5**	**0**
100	AIC	18.258	3	1
	BIC	16.71	3	0
	FFA	11.477	**5**	**0**

سيتم تحليل وتفسير نتائج تجربة المحاكاة تبعا لمعايير دقة التنبؤ ومعيار دقة اختيار المتغيرات. من خلال ملاحظة الجدول (1) و (2) و (3) و (4) الذي يوضح قيم معايير كل من (PE, C, I) للطرائق BIC و AIC والطريقة المقترحة FFA يمكن استخلاص ما يلي:

1- عندما تتغير قيمة معلمة التشتت وبغض النظر عن قيمة حجم العينة، يتبين ان طريقة (FFA) اعطت اقل قيم (PE) حيث بلغ مقدار التحسن بالتنبؤ بالاعتماد على المعيار (PE) بمقدار 35.14% و 30.86% عند (n=50) و $\tau = 1.5$ مقارنة بـ(AIC وBIC) على الترتيب.
2- عندما يتغير حجم العينة وبغض النظر عن قيمة معلمة التشتت، اعطت طريقة (FFA)افضل النتائج مقارنة بالطرائق الاخرى حيث تحسن التنبؤ بالاعتماد على المعيار (PE).
3- بالاعتماد على معايير اختيار المتغيرات، فقد امتلكت طريقة (FFA) اعلى قيم (C) الذي هو عدد المعاملات الحقيقية ذات القيم الصفرية والتي تم تقديرها بشكل صحيح على انها ذات قيم صفرية، واعطت اقل قيم (I) الذي يعرف انه عدد المعاملات الحقيقية ذات القيم غير الصفرية والذي تم تقديرها بشكل غير صحيح على انها ذات قيم صفرية.
4- ظهرت طريقة AIC كأسوأ طريقة في اختيار المتغيرات لأنها تعطي أعلى قيم لـ (PE) وكذلك كأسوأ طريقة في اختيار المتغيرات كونها تميل الى اختيار متغيرات توضيحية غير مهمة.

5- الجانب التطبيقي

في هذا الجانب، يتم إجراء مقارنة بين أداء الطريقة المقترحة ومقدرات أخرى عن طريق استخدام البيانات الحقيقة. لغرض اتمام الفائدة المرجوة من البحث والطريقة المقترحة، تم التطبيق على بيانات تحتوي على تعدد خطي بين المتغيرات التوضيحية والتي أخذت من بيانات استخدمت من قبل (النعيمي، اسوان محمد طيب،2005) حول مرض الثلاسيميا الذي يصاب به الأطفال وبحجم 150 مريض. وقد تم اختيار عشرة متغيرات توضيحية وهي: العمر الحقيقي للطفل (بالشهر) (x1)، عمر المريض عند المرض مقاساً (بالشهر) (x2)، تضخم الكبد مقاساً (بالسنتمتر) (x3)، هيموكلوبين الدم (x4)، مكداس الدم (خلايا الدم المضغوط) (x5)، الخلايا الشبكية (x6)، ارومة حمراء (x7)، الهيموكلوبين الجيني (x8)، عدد وحدات الدم (x9)، بداية نقل الدم حسب العمر مقاساً (بالشهر) (x10). في حين يمثل متغير الاستجابة وهو متغير ثنائي الصفة: العمر من العظم مقاسا بالشهر اكبر من او يساوي 60 و العمر من العظم مقاسا بالشهر اقل من 60.

تم إجراء تقييم لنموذج الانحدار اللوجستي باستخدام طرائق اختيار المتغيرات المشار اليها من خلال حساب قيم متوسط مربعات الخطأ وكذلك عدد المتغيرات المستقلة التي تم اختيارها. توضح النتائج الملخصة في الجدول رقم 5 أن الاسلوب المقترح FFA تفوقت في الأداء على الطرائق الأخرى، حيث حققت أدنى قيمة لـ MSE واقل عدد من المتغيرات المستقلة التي تم اختيارها.

جدول 5: نتائج الجانب التطبيقي

Method	MSE	Variables
AIC	39.561	7
BIC	37.248	6
CSA	**27.931**	**4**

6- الاستنتاجات

تشير النتائج التي تم الحصول عليها من خلال المحاكاة والبيانات الحقيقية في انموذج الانحدار اللوجستي إلى أن استخدام اسلوب FFA يؤدي إلى نتائج ممتازة عند استخدام معيار MSE و PE، مما يجعله موثوقاً للمستخدمين في التنبؤ بالنتائج وتقييم النماذج الإحصائية. وبالإضافة إلى ذلك، يبدو أن حجم العينة n له تأثير كبير على قيم PE ، حيث تنخفض قيمه PE عند زيادة حجم العينة، مما يعني زيادة الدقة. وعلى الجانب الآخر، عند زيادة قيمة عدد المتغيرات المستقلة نلاحظ أيضاً انخفاض في قيمة PE . ويجدر بالذكر أن استخدام معيار MSE مع يؤدي إلى نتائج أفضل في التنبؤ بالنتائج وتقييم النماذج الإحصائية. علاوة على ذلك ان الاسلوب المقترح ابدى قوته باختيار اقل عدد من المتغيرات المستقلة.

Reference

1. Al-Naimi, Aswan Muhammad Tayyab Rashid, 2005, “Testing Variables in Letter Regression,” unpublished master’s thesis, College of Computer Science and Mathematics, University of Mosul, Iraq.
2. Alharthi, A. M., Lee, M. H., & Algamal, Z. Y. (2021). Gene selection and classification of microarray gene expression data based on a new adaptive L1-norm elastic net penalty. *Informatics in Medicine Unlocked, 24*. doi:10.1016/j.imu.2021.100622
3. Månsson, K. (2013). Developing a Liu estimator for the negative binomial regression model: method and application. *Journal of Statistical Computation and Simulation, 83*(9), 1773-1780.
4. Özkale, M. R., & Arıcan, E. (2016). A new biased estimator in logistic regression model. *Statistics*, 1-21. doi:10.1080/02331888.2015.1123711
5. Özkale, M. R., & Arıcan, E. (2018). A first-order approximated jackknifed ridge estimator in binary logistic regression. *Computational Statistics*, *34*(2), 683-712. doi:10.1007/s00180-018-0851-6
6. Steyerberg, E. W., Borsboom, G. J., van Houwelingen, H. C., Eijkemans, M. J., & Habbema, J. D. (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*, *23*(16), 2567-2586. doi:10.1002/sim.1844
7. Varathan, N., & Wijekoon, P. (2018). Liu-Type logistic estimator under Stochastic Linear Restrictions. *Ceylon Journal of Science*, *47*(1). doi:10.4038/cjs.v47i1.7483
8. Xu, H., Yu, S., Chen, J., & Zuo, X. (2018). An Improved Firefly Algorithm for Feature Selection in Classification. *Wireless Personal Communications*, *102*(4), 2823-2834. doi:10.1007/s11277-018-5309-1
9. Yang, X.-S. (2010). *Nature-inspired metaheuristic algorithms*: Luniver press.
10. Zhang, L., Shan, L., & Wang, J. (2016). Optimal feature selection using distance-based discrete firefly algorithm with mutual information criterion. *Neural Computing and Applications*, *28*(9), 2795-2808. doi:10.1007/s00521-016-2204-0
11. Zhang, L., Srisukkham, W., Neoh, S. C., Lim, C. P., & Pandit, D. (2018). Classifier ensemble reduction using a modified firefly algorithm: An empirical evaluation. *Expert Systems with Applications*, *93*, 395-422. doi:10.1016/j.eswa.2017.10.001

Variable Selection In Logistic Regression Model Using Modified Firefly Algorithms

Heba Suleiman Dawood

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Abstract: The logistic regression model is considered the most widely used in many applications, and it is one of the main models in the family of generalized linear models. Like other regression models, the model may contain many independent variables, which negatively affects the accuracy of the model and its simplicity in interpreting the results. This study aims to use the modified firefly algorithm and compare it with other methods for selecting variables in an exponential regression model using simulation and real data. The results showed that compared to other previously used methods, the proposed method performs better and helps reduce the mean square error of the model.

Keyword: Selection of variables, firefly algorithm , simulation , exponential regression model.