المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

اختيار المتغيرات في نموذج انحدار كاوس المعكوس باستخدام خوارزمية الغراب المعدلة

رفل أديب عثمان

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

معلومات النشر

تاريخ المقالة:
تم استلامه في 4 تموز 2023
تم القبول في 18 ايلول 2023
متاح على الإنترنت في 1 كانون الاول 2023

الكلمات الدالة:
اختيار المتغيرات
خوارزمية الغراب
المحاكاة
انموذج انحدار كاوس المعكوس

المراسلة:
رفل اديب عثمان
Rafal.ad81@uomosul.edu.iq

الخلاصة

يعد نموذج انحدار كاوس المعكوس احد النماذج المهمة ، حيث يتم استخدامه بشكل واسع في العديد من التطبيقات. يتم وضع نموذج كاوس المعكوس في جداول عائلات النماذج الخطية المعممة كونه من النماذج الأساسية. 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 الغراب المعدلة ومقارنتها مع طرائق اخرى في اختيار المتغيرات في نموذج انحدار كاوس المعكوس باستخدام المحاكاة والبيانات الحقيقة . حيثُ بيَّنت النتائج أنَّ الاسلوب المقترح يُساهم في تقليل متوسط مربعات الخطأ للنموذج ويحقق أداءً أفضل مُقارنةً بالطرائق الأخرى المستخدمة سابقاً.

DOI: https://doi.org/10.33899/iqjoss.2023.0181218 , ©Authors, 2023College of Computer and Mathematical Science, University of Mosul Iraq. This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

مقدمة

تعتبر دراسة أي مشكلة أو ظاهرة في 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 التحليل الاحصائي المناسب 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[1].

ومن بين جميع نماذج الانحدار الخطي المعممة، يمكن القول أن نموذج انحدار كاوس المعكوس هو من النماذج الواسعة الاستخدام ، اذ يتم استخدامه بشكل واسع في العديد من التطبيقات. يتم وضع نموذج كاوس المعكوس في جداول عائلات النماذج الخطية المعممة كونه من النماذج الأساسية [3] .

كما يتم استخدام طريقة انحدار كاوس المعكوس على نطاق واسع في العديد من مجالات الهندسة الصناعية واختبار الحياة والموثوقية والتسويق والعلوم الاجتماعية. وتكون هذه الطريقة أكثر فائدة في الحالات التي يكون المتغير المستجيب ملتو التواء موجب [4].

غالبية البيانات في الواقع التطبيقي الحقيقي تحتوي على مشاكل مثل مشكلة العدد الكبير من المتغيرات التوضيحية المدروسة، وهي من المشاكل المعروفة لدى الباحثين ، 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 مثل طريقة الاختيار الامامية (Forward selection) و طريقة الحذف العكسي(Backward elimination) و طريقة الاختيار التدريجية (Stepwise selection) 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 [5].
يهدف هذا البحث إلى توظيف خوارزمية الغراب المعدلة ومقارنتها مع طرائق اختيار المتغيرات التوضيحية في أنموذج انحدار كاوس المعكوس 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.

نموذج انحدار كاوس المعكوس (IGRM)

يتم استخدام توزيع كاوس المعكوس ، الذي يحتوي على معلمتين موجبتين، وهما معلمة الموقع (μ) ومعلمة التشتت (τ) ، كتوزيع مستمر عندما يتبع متغير الاستجابة y_i نمطًا منحرفًا بشكل إيجابي. يشار إلى هذا التوزيع بالرمز $IG \sim (\mu,\tau)$ ، ويتم تعريف دالة كثافة الاحتمال لهذا التوزيع على النحو التالي:

$$f(y|\mu,\tau)=\left[\frac{\tau}{2\pi y^3}\right]^{1/2}\exp\left\{-\frac{\tau}{2\mu^2 y}(y-\mu)^2\right\}, y>0 \qquad (1)$$

ينتمي نموذج انحدار كاوس المعكوس (IGRM) إلى عائلة النماذج الخطية المعممة (GLM). يمكن تحويل المعادلة رقم (1) إلى شكل صيغة العائلة الأسية عن طريق اعادة كتابتها كالتالي [6، 7]:

$$f(y,\mu,\tau)=\left\{\frac{y\theta-b(\theta)}{\phi}+C(y,\phi)\right\} \qquad (2)$$

حيث ان :

θ: تسمى معلمة الربط أو دالة الارتباط The canonical parameter or link function

$b(\theta)$: تسمى الدالة التراكمية ,The cumulate function

ϕ : هي معلمة التشتت The dispersion parameter

C(y, ϕ): الحد الطبيعي The normalization term: هي دالة تطبيع تضمن أن المعادلة (2) دالة احتمالية. أي ان $c\left(y,\phi\right)$ ، هي دالة بدلالة ϕ و y تضمن أن $\int f_y\left(y;\theta,\phi\right)dy=1$ إذا كان المتغير y مستمر أو $\sum_y f_y\left(y;\theta,\phi\right)=1$ اذا كان الـ y متقطع [3، 8]

$$f\left(y,\mu,\tau\right)=exp\left\{\frac{-\tau\left(y^2+\mu^2-2\mu y\right)}{2\mu^2 y}-\frac{1}{2}\ln\left(2\pi y^3\right)+\frac{1}{2}\ln\left(\tau\right)\right\} \tag{3}$$

يمكن كتابة المعادلة (3) بشكل أبسط كالتالي:

$$f\left(y,\mu,\tau\right)=exp\left(\tau\left(\frac{-y}{2\mu^2}+\frac{1}{\mu}\right)+(-\frac{1}{2})(ln\left(\frac{2\pi y^3}{\tau}\right)+\frac{\tau}{y})\right) \tag{4}$$

وعليه فإن [6] :

$$x_i^T\beta=\frac{1}{\mu},\sqrt{\eta}=\frac{1}{\mu},\mu=\frac{1}{\sqrt{\eta}}$$

ومن خلال مقارنة المعادلة رقم (4) مع المعادلة رقم (1)، يتم الحصول على:

$$\theta=\frac{1}{2\mu^2},\ b\left(\theta\right)=\frac{1}{\mu}=\frac{1}{\sqrt{-2\theta}}=-\sqrt{-2\theta},\phi=\frac{1}{\tau}$$

ويمكن استخدام دالة الربط للحصول على المتوسط والتباين لمعادلة (4) كالاتي:

$$E\left[y\right]=b'\left(\theta\right)=\frac{\partial b}{\partial\mu}\frac{\partial\mu}{\partial\theta}=\frac{-1}{\mu^2}\left(-\mu^3\right)=\mu=\frac{1}{\sqrt{\eta}} \tag{5}$$

$$\mathrm{V}[\mathrm{y}]=\phi b''(\theta)=\phi \mathrm{V}[\mu]=\frac{1}{\tau}\frac{-1}{\mu^{-3}}=\frac{\mu^3}{\tau} \tag{6}$$

حيث ان $\phi=\frac{1}{\tau}$.

يمكن تعريف دالة الإمكان اللوغاريتمية لنموذج انحدار كاوس المعكوس باستخدام طريقة الإمكان الأعظم لتقدير معلماته. وتأخذ هذه الدالة الشكل التالي[7]:

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{n}\left(\tau\left(\frac{-y_i}{2\mu_i^2}+\frac{1}{\mu_i}\right)+(-\frac{1}{2})(ln\left(\frac{2\pi y_i^3}{\tau}\right)+\frac{\tau}{y_i})\right) \tag{7}$$

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{n}\left(\tau\left(\frac{-y_i\boldsymbol{x}_i^T\boldsymbol{\beta}}{2}-\sqrt{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)+(-\frac{1}{2})(ln\left(\frac{2\pi y_i^3}{\tau}\right)+\frac{\tau}{y_i})\right) \tag{8}$$

يتم حساب المشتقة الجزئية الأولى للمعلمات β لمعادلة (8) ومساواتها بالصفر. وبهذا الإجراء يتم الحصول على مقدر الإمكان الأعظم لـ (IGRM)وفقاً للصيغة المذكورة.

$$\frac{\partial\ell(\beta)}{\partial\beta}=\sum_{i=1}^{n}\frac{1}{2\phi}\left[y_i-\frac{1}{\sqrt{\mathrm{x}_i^T\beta}}\right]\mathrm{x}_i^T=0 \tag{9}$$

يتضح أن المشتقة الأولى لا يمكن حسابها بسبب عدم خطية المعادلة (9) بالنسبة للمعلمة β .وللتغلب على هذه المشكلة، يمكن استخدام التقنيات العددية كما ذكر في الدراسة التي أجراها[9]، مثل طريقة Newton Raphson iterative method [10] أو خوارزمية المربعات الصغرى الموزونة التكرارية (IRLS) ، للحصول على معلمات انحدار كاوس المعكوس (IGRM) ، حيث يتم تحديث المعلمات في كل تكرار باستخدام الصيغة التالية[7]:

$$\hat{\beta}^{(r+1)}=\hat{\beta}^{(r)}+\left(\mathrm{X}^T\hat{\mathrm{W}}^r\mathrm{X}\right)^{-1}\mathrm{X}^T\left(y-\hat{\mu}^r\right) \tag{10}$$

يتم الحصول على تقدير الامكان الاعظم MLE باستخدام خوارزمية IRLS ادناه، والتي تستند إلى عدد التكرارات r .

$$\hat{\beta}_{IGRM} = D^{-1}\mathrm{X}^{T}\hat{\mathrm{W}}\hat{z} \tag{11}$$

حيث ان

$$\mathrm{D} = \left(\mathrm{X}^{T}\hat{\mathrm{W}}\mathrm{X}\right),$$

$$\hat{\mathrm{W}} = \mathrm{diag}\left(\hat{\mu}_i^3\right)$$

$\hat{z}$ يمثل المتغير المعدل للاستجابة ويتم حساب قيمته على النحو التالي:

$$\hat{z}_i = \left(\frac{1}{\hat{\mu}_i^2}\right) + \left(\frac{\left(y_i - \hat{\mu}_i\right)}{\hat{\mu}_i^3}\right)$$

$$\hat{\mu} = \frac{1}{\sqrt{\mathrm{x}_i^{T}\hat{\beta}}}$$

خوارزمية الغراب Crow Search Algorithm (CSA)

تعد خوارزمية البحث عن الغراب واحدة من أحدث الخوارزميات التطورية المستوحاة من السلوك الاجتماعي للغراب. تم تقديم هذه الخوارزمية في عام 2016 من قبل Askarzadeh [11] . في CSA ، يتم تحفيز الفكرة من عملية تخزين الطعام الزائد في أماكن الاختباء ثم استعادتها في الوقت اللازم. من المعروف أن الغراب طائر ذكي للغاية يراقب الآخرين وهم يخفون طعامهم ويسرقه بمجرد مغادرتهم. بعد ارتكاب السرقة ، تختبئ لتتجنب الوقوع ضحية لها في المستقبل.

ليكن لدينا قطيع n_c من الغربان وكل غراب i له موقع عند التكرار tهو x_i^t . يتم حفظ مكان اختباء الطعام الذي يتبعه الغراب.

يتحرك في مستوى البحث ويحاول العثور على أفضل مصدر للطعام والذي يعرف

بـ M_i^t . ان نهج البحث في CSA له سيناريوهان محتملان ؛ الأول هو أن الغراب مالك مصدر الغذاء M_j^t لا يعرف أن الغراب السارق j يتبعه لذلك يصل الغراب اللص إلى مكان اختباء الغراب مالك مصدر الغذاء .حيث تتم عملية تحديث موضع الغراب اللص بواسطة

$$x_i^{t+1} = x_i^t + \tau \times fl \times (M_j^t - x_i^t), \quad i = 1,2,...n_c, \tag{12}$$

حيث ان fl تمثل مسافة الطيران وان τ هي رقم عشوائي ضمن الفترة 0 و 1.

أما السيناريو الثاني هو أن مالك الغراب يعرف أن غراب اللص يتبعه ، لذلك فإن الغراب المالك سوف يخدع الغراب بالذهاب إلى أي موقع آخر في مساحة البحث. يتم تحديث موضع الغراب بواسطة موضع عشوائي. في CSA ، يتم تحديد السيناريو من خلال التعبير التالي:

$$x_i^{t+1} = \begin{cases} x_i^t + \tau \times fl \times (M_j^t - x_i^t), & \text{if } \theta \geq \text{AP} \\ \text{random position}, & \text{otherwise} \end{cases} \tag{13}$$

حيث ان θ هي رقم عشوائي ضمن الفترة 0 و 1. اما AP هي احتمالية الادراك (الاحتمال الملحوظ للغراب j في التكرار).

لإجراء اختيار المتغير ، تم اقتراح خوارزمية ثنائية للبحث عن الغراب. على عكس CSA القياسي ، حيث يتم تحديث الحلول في مساحة البحث نحو المواضع ذات القيمة المستمرة فيBCSA (خوارزمية البحث عن الغراب الثنائية)، تم تصميم مساحة البحث على شكل شبكة منطقية ذات أبعاد n ويتم تحديث الحلول تدريجيا . بالإضافة إلى ذلك ، نظرًا لأن المشكلة تكمن في اختيار أو عدم تحديد متغير معين ، يتم استخدام متجه ثنائي للحل ، حيث يتوافق 1 مع ما إذا كان سيتم تحديد متغير لتكوين مجموعة البيانات الجديدة ، و 0 بخلاف ذلك. في أي خوارزمية ثنائية ، حيث يستخدم المرء متجه الخطوة لحساب احتمالية تغيير المواضع ، تؤثر دوال التحويل بشكل كبير على التوازن بين الاستكشاف والاستغلال [13, 14].

في BCSA ، تُستخدم دالة 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 ، دالة النقل في المعادلة. (14) يمكن استخدامها ، حيث يكون الحل الجديد مقيدًا بالقيم الثنائية فقط

$$x_i^t = \begin{cases} 1 & \text{if} \quad T(x) > \alpha \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

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

$$T_{BCSA}\left(x_i^t\right)=\frac{1}{1+e^{10(x_i^t-0.5)}}, \tag{15}$$

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 ، اذ تحتاج هذه المعلمة الى قيمة عليا وقيمة دنيا لها من خلال المعادلة الخاصة بها وهي:

$$\theta=\theta_{\min}+(\theta_{\max}-\theta_{\min})\,\mathrm{e}^{-t}, \tag{16}$$

وعليه سوف تصبح دالة التحويل المقترحة بالشكل التالي:

$$T_{\mathrm{TV}}\left(x_i^t\right)=\frac{1}{1+e^{10(x_i^t-0.5)/\theta}}, \tag{17}$$

من أجل إتمام هدف البحث وتحقيقهُ، وبالاعتماد على هذه التقنية، فإن كل عنصر (غراب) في المجموعة سيكون لديه d من المواقع التي تمثل عدد المتغيرات التوضيحية في انموذج انحدار كاوس المعكوس. بناءً على ذلك، فإن توظيف خوارزمية الغراب تكون وفق الخطوات التالية:

الخطوة الأولى: تحديد حجم المجموعة (عدد الغربان) وهو 25 غراب، حيث إن كل غراب سيكون له متجه من عدد المتغيرات التوضيحية فضلاً عن ذلك تحديد عدد التكرارات داخل خوارزمية الغراب حيث استقرت النتائج عند التكرار 300.

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:

$$\text{Fitness Function}=\min\left[\frac{\sum_{i=1}^{n}(\mathbf{y_i}-\hat{m}(\mathbf{X}))^2}{n}\right] \tag{18}$$

الخطوة الرابعة: بالاعتماد على أقل قيمة يحصل عليها أي غراب وفق المعادلة (18) يتم تحديث مواقع باقي الغربان.

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 الغراب

معايير تقييم طرائق اختيار المتغيرات

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 :

$$PE = (y - \hat{y})^T (y - \hat{y}) \quad \dots (19)$$

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

ثانيا: معايير تقييم دقة اختيار المتغيرات

بما ان الطرائق 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) .

نتائج المحاكاة

لقد تم تصميم تجربة ومحاكاتها باستعمال لغة البرمجة (R) وتم توليد المتغير (y_i) في انموذج انحدار كاوس المعكوس، و ذلك باستخدام اسلوب مونت كارلو (Mont Carlo) في المحاكاة اذ تم استخدام اربعة احجام من العينات وهي (30,50,100,150) وذلك لأجل دراسة المقارنة وفق العينات باختلاف أنواعها (صغيرة، متوسطة، كبيرة). سوف تتم المقارنة مع كل من طريقة معيار بيز ومعيار اكاكي.

اولا : تم توليد بيانات المتغير y التي تتبع انموذج انحدار كاوس المعكوس ولقيم معلمة التشتت المساوية الى $\tau = 0.5,1.5,3$ وكالاتي :

$$y_i \sim \text{inverse Gaussain}(\mu_i, \tau)$$

$$\mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

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

$$X \sim MN(\mu, M)$$

حيث ان M هي مصفوفة التباين المشترك، و ان $M_{ij} = r^{|i-j|}$، 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 β الذي ابعاده $(p \times 1)$ وكانت قيم متجه معلمات الانحدار β كالاتي $\beta = (1.5,2,0.8,-3.5,5,0,\dots,0)^T$ ، حيث ان المعلمات غير الصفرية عددها $q = 5$ ، وان المعلمات الصفرية تساوي $p - q$.

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

1- عندما تتغير قيمة معلمة التشتت وبغض النظر عن قيمة حجم العينة، يتبين ان طريقة (CSA) اعطت اقل قيم (PE) حيث بلغ مقدار التحسن بالتنبؤ بالاعتماد على المعيار (PE) بمقدار 33.03% و27.57% عند (n=50) و $\tau = 1.5$ مقارنة بـ(AIC و BIC)على الترتيب.

2- عندما يتغير حجم العينة وبغض النظر عن قيمة معلمة التشتت، اعطت طريقة (CSA)افضل النتائج مقارنة بالطرائق الاخرى حيث تحسن التنبؤ بالاعتماد على المعيار (PE).

3- بالاعتماد على معايير اختيار المتغيرات، فقد امتلكت طريقة (CSA) 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.

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

τ	Method	PE	C	I
0.5	AIC	22.152	1	0
	BIC	20.604	3	0
	CSA	15.371	5	0
1.5	AIC	20.528	3	0
	BIC	18.98	4	0
	CSA	13.747	5	0

3	AIC	19.761	4	1
	BIC	18.213	4	0
	CSA	12.98	5	0

جدول (2) : معدل معايير تقييم طرائق الاختيار عندما n=50

τ	Method	PE	C	I
0.5	AIC	21.114	1	0
	BIC	19.566	2	0
	CSA	14.333	5	0
1.5	AIC	19.49	3	0
	BIC	17.942	4	0
	CSA	12.709	5	0
3	AIC	18.723	3	1
	BIC	17.175	4	0
	CSA	11.942	5	0

جدول (3) : معدل معايير تقييم طرائق الاختيار عندما n=100

τ	Method	PE	C	I
0.5	AIC	19.336	1	0
	BIC	17.788	3	0
	CSA	12.555	5	0
1.5	AIC	17.712	3	0
	BIC	16.164	4	0
	CSA	10.931	5	0
3	AIC	16.945	3	1
	BIC	15.397	4	0
	CSA	10.164	5	0

جدول (4) : معدل معايير تقييم طرائق الاختيار عندما n=150

τ	Method	PE	C	I
0.5	AIC	18.298	1	0
	BIC	16.75	3	0
	CSA	11.517	5	0
1.5	AIC	16.674	4	0
	BIC	15.126	4	0
	CSA	9.893	5	0

3	AIC	15.907	3	1
	BIC	14.359	4	0
	CSA	9.126	5	0

الجانب التطبيقي

في هذا الجانب، يتم إجراء مقارنة بين أداء المقدر المقترح IGDK ومقدرات أخرى عن طريق استخدام البيانات الفعلية. ويتم تقييم أداء المقدرات باستخدام معيار MSE. وللتحقق من أداء الطريقة المقترحة IGDK باستخدام البيانات الفعلية، تم استخدام بيانات كيميائية محددة $(n, p) = (65,15)$، تمثل n عدد مشتقاتimidazole[4,5−b] pyridine وهي مركبات مضادة للسرطان ، في حين يمثل الرمز p المتغيرات التوضيحية والتي ترمز الخصائص الجزيئية (Yahya Algamal, 2019;10) . تتناول هذه الفقرة دور متغير (IC50) كمتغير الاستجابة في تقييم الأنشطة البيولوجية للمركبات المضادة للسرطان، وتسلط الضوء على أهمية دراسة العلاقة الكمية بين التراكيب الكيميائية والفاعلية البيولوجية باستخدام نمذجة QSAR. ويعرف QSAR بوصفه نموذجاً للأنشطة البيولوجية على أساس الخصائص الهيكلية لمجموعة من المركبات الكيميائية[15].

تم استخدام اختبار مربع كاي للمطابقة لتحديد مدى ملاءمة التوزيع الكاوسي المعكوس لمتغير الاستجابة(IC50) ، حيث أظهرت النتائج قيمة تساوي 5.2762 وقيمة $p\text{-value}$ تساوي 0.2601. وبناءً على هذه النتائج، يمكن اعتبار أن التوزيع كاوس المعكوس مناسب لمتغير الاستجابة المعتمد[7]. فيما يخص الدراسة، فقد تم تضمين 15 واصفاً جزيئياً يمثلون المتغيرات التوضيحية (مستقلة).

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

جدول 4: نتائج الجانب التطبيقي

Method	MSE	# variables
AIC	43.681	11
BIC	41.368	9
CSA	32.051	6

الاستنتاجات

تشير النتائج التي تم الحصول عليها من خلال المحاكاة والبيانات الحقيقية إلى أن استخدام اسلوب CSA يؤدي إلى نتائج ممتازة عند استخدام معيار MSE و PE، مما يجعله موثوقاً للمستخدمين في التنبؤ بالنتائج وتقييم النماذج الإحصائية. وعلى الجانب الآخر، عند زيادة قيمة معلمة التشتت نلاحظ أيضاً انخفاض في قيمةPE . 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 التوضيحية.

Reference

1. Ross, S.M., Introduction to probability and statistics for engineers and scientists. 2020: Academic press.
2. McCullagh, P. and J. Nelder, Generalized Linear Models. 1989, London: Chapman and Hall.
3. Peter K. D., G.K.S., Generalized Linear Models With Examples in R. 2018, New York: Springer.

4. yonis, F.a. and R.A. Othma, Shrinkage estimators in inverse Gaussian regression model: Subject review. IRAQI JOURNAL OF STATISTICAL SCIENCES, 2022. 19(35): p. 72-82.
5. Alkhateeb, A.N. and Z.Y.J.E.J.o.A.S.A. Algamal, Variable selection in gamma regression model using chaotic firefly algorithm with application in chemometrics. 2021. 14(1): p. 266-276.
6. Akram, M.N., et al., A new Liu-type estimator for the inverse Gaussian regression model. 2020. 90(7): p. 1153-1172.
7. Yahya Algamal, Z., Performance of ridge estimator in inverse Gaussian regression model. Communications in Statistics-Theory
Methods, 2019. 48(15): p. 3836-3849.
8. Olsson, U., Generalized linear models. An applied approach. Studentlitteratur, Lund, 2002. 18.
9. Salh, A.P.D.S.M., et al., Using Multinomial Logistic Regression model to study factors that affect chest pain. 2021. 17(53 part 2).
10. Mawlood, K.I., Estimating Hazard Function and Survival Analysis of Tuberculosis Patients in Erbil city. Tikrit Journal of Administration
Economics Sciences, 2021. 17(54 part 3).
11. Askarzadeh, A., A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm. Computers & Structures, 2016. 169: p. 1-12.
12. Sayed, G.I., A.E. Hassanien, and A.T. Azar, Feature selection via a novel chaotic crow search algorithm. Neural Computing and Applications, 2017.
13. Islam, M.J., X. Li, and Y. Mei, A time-varying transfer function for balancing the exploration and exploitation ability of a binary PSO. Applied Soft Computing, 2017. 59: p. 182-196.
14. Mafarja, M., et al., Binary dragonfly optimization for feature selection using time-varying transfer functions. Knowledge-Based Systems, 2018. 161: p. 185-204.
15. Algamal, Z.Y., & Lee, M. H. , A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression. Journal of Chemometrics, 2017. 31(10).

Variable selection in Inverse Gaussian regression model using modified crow search algorithm

Rafal Adeeb Othman

Rafal.ad81@uomosul.edu.iq

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

Abstract

The inverse Gaussian regression model is one of the important models, which is widely used in many applications. The inverse Gaussian model is placed in tables of families of generalized linear models as it is one of the basic 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 crow search algorithm and compare it with other methods in selecting the variables in the inverse Gaussian regression model using simulation and real data. The results showed that the proposed method contributes to reducing the average square error of the model and achieves better performance compared to other previously used methods.

Keyword: Choice of variables, Raven algorithm, Simulation ,Inverse Gauss regression model.