Tikrit Journal of Administrative and Economic Sciences, Vol. 20, No. 68, Part (1): 403-414
Doi: www.doi.org/10.25130/tjaes.20.68.1.23

Tikrit Journal of Administrative and Economics Sciences
مجلة تكريت للعلوم الإدارية والاقتصادية
EISSN: 3006-9149
PISSN: 1813-1719

Variable selection in Poisson regression model using Golden jackal optimization algorithm

Diaa Nazem Ahmed Khalaf*, Zakaria Yahya Nouri Al-Jamal, Munther Khalil Abdullah
College of Computer Science and Mathematics/Tikrit University

Keywords:
Ibn Awi Al-Dhahabi's algorithm, Poisson's regression, variable selection, simulation experiment description

A R T I C L E I N F O

Article history:
Received 15 Jul. 2024
Accepted 17 Aug. 2024
Available online 31 Dec. 2024

©2023 THIS IS AN OPEN ACCESS ARTICLE UNDER THE CC BY LICENSE
http://creativecommons.org/licenses/by/4.0/

*Corresponding author:
Diaa Nazem Ahmed Khalaf
College of Computer Science and Mathematics/Tikrit University

Abstract: Variable selection is a very helpful procedure for improving prediction accuracy by finding the most important variables that related to the response variable. Poisson regression model has received much attention in several science fields for modeling count data. Golden jackal optimization algorithm (GJO) is one of the recently efficient proposed nature-inspired algorithms that can efficiently be employed for variable selection. In this work, GJO algorithm is proposed to perform variable selection for Poisson regression model. Extensive simulation studies and real data application are conducted to evaluate the performance of the proposed method in terms of prediction accuracy and variable selection criteria. The results proved the efficiency of our proposed methods and it outperforms other popular methods.

Tikrit Journal of Administrative and Economic Sciences, Vol. 20, No. 68, Part (1): 403-414
Doi: www.doi.org/10.25130/tjaes.20.68.1.23

اختيار المتغيرات في نموذج انحدار بواسون باستخدام خوارزمية ابن اوى الذهبي

ضياء ناظم احمد خلف زكريا يحيى نوري الجّمال منذر خليل عبدالله
كلية علوم الحاسوب والرياضيات/جامعة تكريت

المستخلص

يعد نموذج انحدار بواسون أحد أهم نماذج الانحدار اللوغاريتمية الخطية، وهو الأداة التي يتم من خلالها نمذجة المتغير المعتمد عندما تكون قيم ذلك المتغير على شكل قيم قابلة للعد. 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 ابن اوى الذهبي ومقارنتها مع طرائق اخرى في اختيار المتغيرات في نموذج انحدار بواسون باستخدام المحاكاة والبيانات الحقيقة وتم استخدام أسلوب مونت – كارلو في المحاكاة لتوليد بيانات تتبع نموذج انحدار بواسون تبعا لعوامل مختلفة كحجم العينة وعدد المتغيرات المستقلة وتم الاعتماد على جانبين من جوانب تقييم أداء الطرائق المستخدمة: الأول هو تقييم دقة التنبؤ والثاني هو تقييم اختيار المتغيرات كمعيار للمقارنة، فقد أظهرت نتائج المحاكاة تفوق خوارزمية ابن اوى الذهبي مقارنةً بطرائق اختيار المتغيرات الأخرى. فضلا من ذلك، تم التطبيق على بيانات حقيقية تم جمعها من مصابين بمرض العجز الكلوي المزمن والذين يتعالجون بالغسيل الكلوي المستمر، وقد شخصت حالة المرضى من قبل أطباء مختصين بالتعاون مع مستشفى ابن سينا التعليمي – وحدة الكلية الاصطناعية.

الكلمات المفتاحية: خوارزمية ابن اوى الذهبي، انحدار بواسون، اختيار المتغيرات، وصف تجربة المحاكاة

1. المقدمة

تحليل الانحدار هو أداة احصائية تقوم ببناء أُنموذج احصائي وذلك لتقدير العلاقة بين متغير واحد يدعى المتغير التابع ومتغير آخر أو متغيرات عدة أخرى تدعى المتغيرات التوضيحية (التفسيرية)، بحيث ينتج معادلة احصائية توضح العلاقة بين المتغيرات. لقد احتل تحليل الانحدار بنماذجه المختلفة مكانة متميزة في توجهات العديد من علماء الإحصاء، ونالت نصيبها الوافر عبر المؤلفات الاحصائية المختلفة، وأصبح دورها مهم جدا في تطبيقات علوم الحياة المتنوعة خصوصاً في المجال الاقتصادي الذي أخذ على عاتقه اعتماد نماذج الانحدار بالدرجة الأساس لتكون أبرز وسائل الدعم العملي للنظريات الاقتصادية، فضلاً عن العلوم الأخرى كالصحية والحياتية والاجتماعية وغيرها (الراوي، 1978).

يفترض أنموذج الانحدار الخطي الكلاسيكي أن متغير الاستجابة يعتمد على مجموعة من المتغيرات التوضيحية، حيث يمكن أن تكون هذه المتغيرات عبارة عن متغيرات مستمرة أو متغيرات قابلة للعد، ومع ذلك، عندما يكون متغير الاستجابة بشكل متغيرات قابلة للعد مثل عدد المرضى، فأنه سوف لن تتحقق افتراضات الانحدار الخطي. لذلك تم اقتراح أنموذج انحدار بواسون كأحد نماذج الانحدار التي تتوافق مع هكذا حالات.

اختيار المتغيرات في بيانات العد باستخدام أنموذج انحدار بواسون هي واحدة من التحديات في تطبيق أنموذج انحدار بواسون عندما يكون عدد المتغيرات التوضيحية كبير، حيث اصبحت الاساليب التقليدية لاختيار المجموعات الجزئية مثل طريقة الأختيار الأمامية

(Forward selection) وطريقة الاختيار إلى الخلف (Backward elimination) وطريقة الاختيار التدريجية (Stepwise selection) غير جيدة في أداء وظيفتها حيث أصبحت أكثر تكلفة في حسابها، فضلا عن 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 (Algamal,2015).

لقد تناولت الدراسة الحالية أنموذج انحدار بواسون (Poisson Regression Model (PRM)) الذي يعد أحد النماذج الأكثر شعبية بين النماذج التي لديها متغير استجابة قابل للعد، وقد تم وصفه أولا عن طريق الباحثَين Nelder and Wedderburn (1972)، كحالة خاصة من النماذج الخطية المعممة Generalized linear models (GLMs)، وللوقوف على مدى أهمية المنهجية مقارنة بالطرائق التقليدية الأخرى سيتم اخضاع الأنموذج المستخدم ومن ثم توظيف معايير تقييم المعنوية لنتائج كل طريقة.

يهدف هذا البحث إلى توظيف خوارزمية ابن اوى الذهبي ومقارنتها مع طرائق اختيار المتغيرات التوضيحية في أنموذج انحدار بواسون أخرى 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.

2. **نموذج انحدار بواسون:** يعد نموذج انحدار بواسون أحد أهم نماذج الانحدار اللوغاريتمية الخطية، وهو الأداة التي يتم من خلالها نمذجة المتغير المعتمد عندما تكون قيم ذلك المتغير على شكل قيم قابلة للعد. وكغيره من سائر نماذج الانحدار، قد يحتوي النموذج على متغيرات مستقلة كثيرة ما يؤثر سلباً على دقة النموذج وبساطته في تفسير النتائج. يفترض هذا الأنموذج أن المتغير المعتمد y_i هو متغير استجابة يتبع توزيع بواسون وبمعلمة قدرها (μ)، كما تتبع الأخطاء العشوائية في الأنموذج توزيع بواسون بمعلمة قدرها (μ) Mansson and Kubria (Hossain And Ahmed,(2012)) . (2012) ويعرف وفق الدالة الاحتمالية المعرفة بالصيغة الآتية.

$$y_i = e^{XB+U} \qquad \dots (1)$$

ويمكن التعبير عنه ايضا بصيغة المصفوفات وكالآتي:

$$\mathbf{y} = \mathrm{Exp}(\mathbf{X\beta} + \mathbf{U}) \qquad \dots (2)$$

إذ إن:

$\mathbf{y}$: موجه المتغير التابع ذي درجة، $(n\times1)$ $\mathbf{X}$: مصفوفة المتغيرات المستقلة (التوضيحية) ذات الدرجة $(n\times(p+1))$ $\boldsymbol{\beta}$: موجه المعلمات ذو الدرجة $((p+1)\times1)$ $\mathbf{U}$: موجه الأخطاء العشوائية ذي الدرجة $(n\times1)$ n : حجم العينة P : عدد المتغيرات المستقلة (التوضيحية).

لأجل تقدير معلمات أنموذج انحدار بواسون باستخدام طرائق الإمكان الجزائية سيتم اللجوء الى تعظيم المشاهدات لتوزيع المتغير المعتمد (y_i)، إذا كان المتغير المعتمد (y_i) يتبع توزيع بواسون بمعلمة قدرها (μ_i) فتكون دالة التوزيع كما في الصيغة (1)والمعرفة سلفا بالشكل الآتي:

Tikrit Journal of Administrative and Economic Sciences, Vol. 20, No. 68, Part (1): 403-414
Doi: www.doi.org/10.25130/tjaes.20.68.1.23

$$f(y_i/\mu_i) = \frac{e^{-\mu i}\,\mu_i^{y_i}}{y_i!}$$

ومن خلال تعظيم المشاهدات لتوزيع المتغير المعتمد (y_i) الوارد في الصيغة أعلاه تكون دالة الإمكان الأعظم بالشكل الآتي:

$$L(y_1, y_2, \dots, y_n; \mu_i) = \frac{Exp\{-\sum_{i=1}^{n}\mu_i\}\;\mu_i^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i\,!} \qquad \dots (3)$$

وبأخذ اللوغاريتم الطبيعي لدالة الإمكان الأعظم للمشاهدات أعلاه نحصل على:

$$LogL(\mathbf{y_i}|x_i, \boldsymbol{\beta}) = -\sum_{i=1}^{n}\mu_i + \sum_{i=1}^{n} y_i\,(Log\{\mu_i\}) - Log\left\{\prod_{i=1}^{n} y_i\,!\right\} \quad \dots (4)$$

وبالاعتماد على الافتراض الثاني من الفروض الأساسية لأنموذج انحدار بواسون $\mu_i = Exp\{x_i^T\boldsymbol{\beta}\}$، يتم تعويض هذا الافتراض بالدالة (4) أعلاه وكما يأتي:

$$\begin{aligned} LogL(\mathbf{y_i}|x_i, \boldsymbol{\beta}) &= -\sum_{i=1}^{n}\left(Exp\{x_i^T\boldsymbol{\beta}\}\right) + \sum_{i=1}^{n} y_i\left(Log\left\{Exp\{x_i^T\boldsymbol{\beta}\}\right\}\right) \\ &\quad - Log\left\{\prod_{i=1}^{n} y_i\,!\right\} \\ &= \sum_{i=1}^{n}\left(y_i x_i^T\boldsymbol{\beta} - EXP\left(x_i^T\boldsymbol{\beta}\right) - \log y_i!\right) \qquad \dots (5) \end{aligned}$$

3. **خوارزمية ابن اوى الذهبي Golden Jackal Optimization Algorithm:** تعتمد معظم خوارزميات التحسين المستوحاة من الطبيعة على ذكاء السرب، وتشكل الخوارزميات القائمة على ذكاء السرب جزءاً كبيراً من الخوارزميات المعاصرة، وأصبحت هذه الخوارزميات مستخدمة على نطاق واسع في التصنيفات والتحسين ومعالجة الصور وذكاء الأعمال وكذلك في التعلم الآلي والذكاء الاصطناعي. وتعدُّ خوارزمية ابن آوى الذهبي (Golden Jackal Optimization Algorithm) واحدة من أحدث أساليب ذكاء السرب الجديدة وأقوى خوارزميات التحسين التي تم تطويرها لأول مرة في عام 2022 بواسطة Chopra and Ansari (Chopra & Mohsin Ansari, 2022).

أثبتت خوارزمية ابن أوى الذهبي أنّها فعالة وذات أداء جيد في حل مشكلات التحسين المختلفة. تم إيجاد خوارزمية ابن أوى الذهبي من محاكاة سلوك الصيد لديه. يعد ابن اوى الذهبي حيوان مفترس أرضي متوسط الحجم يعيش في شمال وشرق أفريقيا والشرق الأوسط وأوروبا وجنوب شرق آسيا وآسيا الوسطى. يبلغ طول جسم ابن اوى الذهبي حوالي 70 إلى 85 سم ويبلغ ارتفاعه عند الوقوف حوالي 40 سم وطول الذيل حوالي 25 سم ويكون الفراء عاده ما يكون خشنا ذو أطراف بنية ذهبية شاحبة إلى صفراء ويختلف باختلاف المنطقة والموسم. يسمح الجسم الصغير والأرجل الطويلة لابن آوى الذهبي بالركض لمسافات طويلة لصيد الفريسة. وتكون المراحل الأولى لصيد زوج ابن اوى الذهبي كالآتي:

أ. البحث والتقدم نحو الفريسة.
ب.تطويق الفريسة وازعاجها حتى تتوقف عن الحركة.
ج. الانقضاض على الفريسة.

تم تصميم استراتيجية الصيد لزوج ابن آوى الذهبي رياضيا وتحسين النموذج الرياضي. والخوارزمية توضح القسم الفرعي لعملية تطوير خوارزمية ابن آوى الذهبي ولتحسين الخوارزمية كطريقة بسيطة ومتجانسة يتم صياغة مساحة البحث حالها حال العديد من الخوارزميات الأخرى.

تبدأ خوارزمية ابن آوى الذهبي بتوليد الحل الأولي بشكل موحد على نطاق فضاء البحث وبالشكل الآتي:

$$\mathbf{Y_0 = Y_{min} + rand(Y_{max} - Y_{min})} \quad \ldots\ldots\ldots\ldots (.6)$$

حيث إن:

$\mathbf{y_{max}}$: الحد الأعلى للمتغيرات.

$\mathbf{y_{min}}$: الحد الأدنى للمتغيرات وهو متغير عشوائي يتبع التوضيع المنتظم ضمن الفترة 0 و1. وبالتالي نحصل على المصفوفة الأولية لقيم فضاء البحث والتي تكون كالآتي:

$$\mathbf{Prey} = \begin{bmatrix} \mathbf{Y_{1,1}} & \mathbf{Y_{1,2}} & \cdots & \mathbf{Y_{1,d}} \\ \mathbf{Y_{2,1}} & \mathbf{Y_{2,1}} & \cdots & \mathbf{Y_{2,d}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{Y_{n,1}} & \mathbf{Y_{n,2}} & \cdots & \mathbf{Y_{n,d}} \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

حيث إن:

$\mathbf{y_{ij}}$: يشير إلى البعدj للفريسة i.

n: اجمالي من الفرائس والمتغيرات.

d: يشير إلى موضع الفريسة للمعلمات.

بعد ذلك يتم تطبيق دالة اللياقة (دالة الهدف) لكل صف من صفوف المصفوفة الذي يمثل الفريسة الموضحة في المعادلة (7) وبالشكل الآتي:

$$\mathbf{F_{OA}} = \begin{bmatrix} \mathbf{f(Y_{1,1}; Y_{1,2}; \cdots; Y_{1,d})} \\ \mathbf{f(Y_{2,1}; Y_{2,1}; \cdots; Y_{2,d})} \\ \vdots \\ \mathbf{f(Y_{n,1}; Y_{n,2}; \cdots; Y_{n,d})} \end{bmatrix} \ldots\ldots\ldots\ldots (8)$$

حيث $\mathbf{F_{OA}}$ هي المصفوفة قيم دالة اللياقة لكل فريسة ويظهر $\mathbf{Y_{ij}}$ قيمة البعد j للفريسة i.

n: عدد الفرائس.

يعتمد أسلوب الخوارزمية على مرحلتين: مرحلة الاستكشاف (Exploration) ومرحلة الاستغلال (Exploitation).

أولاً. مرحلة الاستكشاف أو البحث عن الفريسة: تم اقتراح استراتيجية لاستكشاف خوارزمية ابن اوى الذهبي. وطبيعة ابن اوى هي معرفة كيفية إدراك الفريسة ومتابعتها بشكل متقطع لا يمكن

الامساك بالفريسة بسهولة ومن ثم ابن اوى ينتظر ويبحث عن فريسة أخرى. علما إن الصيد يقوده ذكر ابن اوى وتتبعه انثى ابن اوى.

$$\mathbf{Y_1(t) = Y_M(t) - E \cdot |Y_M(t) - rl \cdot Prey(t)|} \quad \dots\dots\dots (9)$$

$$\mathbf{Y_2(t) = Y_{FM}(t) - E \cdot |Y_{FM}(t) - rl \cdot Prey(t)|} \quad \dots\dots\dots\dots (10)$$

حيث إن:

(t): التكرار الحالي.

Prey(t): هو متجه الموضع ل الفريسة.

$\mathbf{Y_{FM}(t)}$و$\mathbf{Y_M(t)}$ **:** يشير إلى موضع ذكر وانثى ابن اوى.

$\mathbf{Y_1(t)}$ و $\mathbf{Y_2(t)}$ عبارة عن مواضع محدثة لذكروانثى ابن اوى تتوافق مع الفريسة.

E: هو التهرب من طاقة الفريسة ويتم حسابه على النحو الآتي:

$$\mathbf{E = E_1 * E_0} \quad \dots\dots\dots\dots (11)$$

حيث إن:

E_1 : يشير إلى الطاقة المتناقصة للفريسة.

E_0: يشير إلى الحالة الأولية لطاقتها.

$$\mathbf{E_0 = 2 * r - 1} \quad \dots\dots\dots (12)$$

حيث إن r هو رقم عشوائي بين (0,1).

$$\mathbf{E_1 = c_1 * \left(1 - \left(\frac{t}{T}\right)\right)} \quad \dots\dots\dots\dots (13)$$

تشير (T) إلى الحد الأقصى لعدد التكرارات وتشير C1 إلى قيمة ثابتة تساوي 1.5 يتم تقليل E_1 خطيا من 1.5 إلى 0.

$\mathbf{|y(t) - r|.Prey(t)}$في المعادلة (11) و(12) يحسب المسافة بين ابن اوى والفريسة. يتم طرح هذه المسافة بين ابن اوى والفريسة أو اضافتها. الوضع الحالي لابن اوى يعتمد على التهرب من طاقة الفريسة. r في المعادلة (11) مكافئ وفي المعادلة (12)هو متجه للأرقام العشوائية على اساس توزيع (ليقي) يمثل حركة ليفي مضاعفة r و**Prey** حركة الفريسة. ويتم حسابها باستخدام

$$\mathbf{rl = 0.05 * LF(y)} \quad \dots\dots\dots\dots (14)$$

هي دالة طيران الضريبة والتي يتم حسابها باستخدام

$$\mathbf{LF(y) = 0.01 \times \frac{\mu \times \sigma}{\left(\left|v^{\left(\frac{1}{\beta}\right)}\right|\right)}; \sigma = \left(\frac{\Gamma(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times \left(2^{\frac{1-1}{2}}\right)}\right)^{\frac{1}{\beta}}} \quad \dots\dots (15)$$

حيث إن:

μ وv: قيم عشوائية محصورة بين (0 و1).

β: هي مجموعة ثابتة افتراضية الى 1.5.

ويتم تحديث خوارزمية ابن اوى الذهبي عن طريق اخذ متوسط المعادلة (4) ومكافئ (5).

$$\mathbf{Y(t+1) = \frac{Y_1(t) + Y_2(t)}{2}} \qquad \dots\dots\dots\dots (16)$$

مرحلة الاستغلال أو التطويق والانقضاض على الفريسة عندما تتعرض الفريسة لمضايقة ابن اوى

$$\mathbf{Y_1(t) = Y_M(t) - E.|rl.Y_M(t) - Prey(t)|} \qquad \dots\dots.. (17)$$

$$\mathbf{Y_2(t) = Y_{FM}(t) - E.|rl.Y_{FM}(t) - Prey(t)|} \qquad \dots\dots\dots. (18)$$

يتم حساب طاقة التهرب من الفريسة **E** وفقا المعادلة (4) وأخيرا تم تحديث موافق ابن اوى وفقا المعادلة (16) ودالة **rl** في المعادلة (17) ومكافئ (18) هو تقدم تعسفي السلوك في مرحلة الاستغلال وتفضيل الاستكشاف والمحلية تجنب الأمثل ويتم حساب "**rl**" وفقا المعادلة (14) يساعد هذا العنصر في تجنب التباطؤ المحلي خاصة في الختام.

ثانياً. التحول من الاستكشاف إلى الاستغلال: في خوارزمية ابن اوى الذهبي يتم استخدام الطاقة الهاربة من الفريسة التحول من الاستكشاف إلى الاستغلال تنخفض طاقة الفريسة بشكل ملحوظ خلال سلوك التهرب. بالنظر إلى هذه الفريسة تم تصميم نموذج التهرب من الطاقة وفق المعادلة الطاقة الأولية $\mathbf{E_0}$ بشكل تعسفي ينحرف من (-1 الى 1) في كل تكرار عندما تنخفض فيه $\mathbf{E_0}$من (0 -1) تتضال الفريسة جسديا على الرغم من ارتفاع قيمة $\mathbf{E_0}$ من (0 إلى 1) فأنه يشير إلى تحسين في قوة الفريسة تنخفض قوة التهرب المتغير $\mathbf{E_0}$اثناء عملية التكرار عندما $\mathbf{E > 1}$ تبحث ازواج ابن اوى في أقسام مختلفة لاستكشاف الفريسة ومتى $\mathbf{E < 1}$ يهاجم ابن اوى الذهبي الفريسة وينقذها الاستغلال في الختام. تبدء عملية البحث في خوارزمية ابن اوى الذهبي بأنشاء مجموعة عشوائية من الفريسية أثناء التكرارات يتم تقدير الموقع المحتمل للفريسة من خلال صيد ابن اوى من الذكور والاناث زوج يقوم كل مرشح في السكان بتحديث المسافة التي تفصله عن زوج ابن اوى يتم تقليل المعلمة $\mathbf{E_1}$من (1.5 الى 0)لتأكيد الاستكشاف والاستغلال على التوالي زوج صيد ابن اوى الذهبي تنحرف عن الفريسة عندما يكون $\mathbf{E > 1}$ وتتجمع مع الفريسة عندما يكون $\mathbf{E < 1}$ تم الانتهاء من خوارزمية ابن اوى الذهبي من خلال تحقيق معيار النهاية برد الكود الزائف لخوارزمية ابن اوى الذهبي. الشكل (1) يوضح آلية عمل الخوارزمية في اختيار المتغيرات.

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

شكل (1): آلية اختيار المتغيرات حسب خوارزمية ابن اوى الذهبي

ثالثاً. معايير تقييم طرائق الجزاء: إن أسلوب تقييم أداء الطرائق الجزائية ومقارنة هذه الطرائق فيما بينها واختيار الطريقة الأفضل هو جانب مهم من جوانب تحليل البيانات. بشكل عام هناك جانبين من جوانب تقييم أداء الطرائق الجزائية: الاول هو تقييم دقة التنبؤ والثاني هو تقييم اختيار المتغيرات.

رابعاً. معايير تقييم دقة التنبؤ:

❖ **خطأ التقدير (Estimation Error) (EE):** ويعرف بأنه مربع الفرق بين قيمة المعلمات الحقيقية وقيمة المعلمات المقدرة ويعرف بالشكل الرياضي الآتي:

$$EE = (\hat{\beta} - \beta)^T(\hat{\beta} - \beta) \quad \dots (9)$$

حيث إن$\hat{\beta}$:هو متجه المعلمات المقدرة وفق الطرائق المستخدمة وβهو متجه المعلمات الحقيقية.

❖ **خطأ التنبؤ (Prediction Error) (PE):** ويعرف بأنه مربع الفرق بين القيمة الحقيقية لمتغير الاستجابة والقيمة التنبؤية المرافقة له، ويعرف رياضيا بالمعادلة الآتية:

$$PE = (y - \hat{y})^T.(y - \hat{y}) \quad \dots (10)$$

حيث إن $\hat{y} = Exp\{X^T\boldsymbol{\beta}\}$ وبالاعتماد على هذين المعيارين يتم تحديد الطريقة الأفضل التي تعطي أقل قيمة مقارنة بالطرائق الأخرى.

خامساً. معايير تقييم دقة اختيار المتغيرات: بما إن الطرق المقترحة بصورة عامة (I,C) 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 الآتي:

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

ب. **معيار التقييم "I":** معيار التقييم الذي يرمز له بـ (**I**) وهو يعرف على أنه عدد المعاملات الحقيقية ذات القيم غير الصفرية والذي تم تقديرها بشكل غير صحيح على أنها ذات قيم صفرية. تعتمد جودة طرائق الجزاء من ناحية معايير تقييم دقة اختيار المتغيرات على من يعطي أعلى قيمة لـ (**C**) وأقل قيمة لـ (**I**).

4. **وصف تجربة المحاكاة Description Simulation Experiment:** لقد تم تصميم تجربة ومحاكاتها باستعمال لغة البرمجة (R) حيث تم توليد المتغير (y_i) في انموذج انحدار بواسون الذي يتبع توزيع بواسون بمعدل مقداره (μ_i)، حيث تم استخدام أسلوب مونت كارلو (Mont Carlo) في المحاكاة حيث تم تعيين قيم حجم العينات (n) حيث تم استخدام ثلاث احجام من العينات وهي (50,100,150) وذلك لأجل دراسة المقارنة وفق العينات باختلاف أنواعها. سوف تتم المقارنة مع كل من طريقة LASSO والتي تمثل مختصر Least Absolute Shrinkage and Selection Operator وكذلك طريقة SCAD والتي تمثل مختصر Smoothly Clipped Absolute Deviation.

5. **دراسات المحاكاة Simulation Studies**

أولا: تم توليد بيانات المتغير y التي تتبع انموذج انحدار بواسون وكالآتي:

$$y \sim P(\exp(\mathbf{X\beta}))$$

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

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

حيث إن **M** هي مصفوفة التباين المشترك، حيث إن $\mathbf{M_{ij}} = \mathbf{r}^{|\mathbf{i}-\mathbf{j}|}$, عندما $(i, j = 1,2, \dots, p)$ حيث إن 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 أبعاده ($p \times 1$) وكانت قيم متجه معلمات الانحدار $\boldsymbol{\beta}$ كالآتي:

$$\boldsymbol{\beta} = (\mathbf{1.2}, -\mathbf{0.6}, \mathbf{0.8}, -\mathbf{0.4}, \mathbf{1.5}, \mathbf{0}, \dots, \mathbf{0})^T$$

حيث ان المعلمات غير الصفرية عددها $q = 5$، وإن المعلمات الصفرية تساوي $p - q$.

6. تفسير نتائج المحاكاة: سيتم تحليل وتفسير نتائج تجربة المحاكاة تبعا لمعايير دقة التنبؤ ومعيار دقة اختيار المتغيرات. من خلال ملاحظة الجدول رقم (1) و(2) و(3) الذي يوضح قيم معايير كل من (EE, PE, C, I) للطرائق الجزائية (LASSO, SCAD) المقترحة من قبل الباحثين Zou (2006), Zou and Hastie (2005), Tibshirani (1996), Fan and Li (2001), El-Anbari and Mkhadri (2013) والطريقة المقترحة GJO يمكن استخلاص ما يأتي:

أ. عندما يتغير معامل الارتباط بين المتغيرات من (0.5) إلى (0.7)، يتبين أن طريقة (GJO) أعطت أقل قيم (EE, PE) حيث بلغ مقدار التحسن بالتنبؤ بالاعتماد على المعيار (PE) بمقدار 69.36% و2.3% عند (r=0.5) و63.63% و2.83% عند (r=0.7) مقارنة بـ LASSO) وSCAD) على الترتيب، كما وبلغ التحسن بخطأ التقدير بالاعتماد على المعيار (EE) بمقدار 99.01% و45.22% عند (r=0.5) و97.66% و39.34% عند (r=0.7) مقارنة بـ(LASSO وSCAD) على الترتيب.

ب. عندما يكون معامل الارتباط مساوي إلى (0.9) أعطت طريقة (GJO)أفضل النتائج مقارنة بالطرائق الاخرى حيث تحسن التنبؤ بالاعتماد على المعيار (PE) بمقدار 52.59% و7.28% مقارنة بـ (LASSO وSCAD) على الترتيب، كما وبلغ التحسن في خطأ التقدير بالاعتماد على المعيار (EE) بمقدار 94.90% و78.77% مقارنة بـ (LASSO وSCAD) على الترتيب.

ج. بالاعتماد على معايير اختيار المتغيرات، فقد امتلكت طريقة (GJO) اعلى قيم (C) الذي هو عدد المعاملات الحقيقية ذات القيم الصفرية والتي تم تقديرها بشكل صحيح على أنها ذات قيم صفرية، وأعطت أقل قيم (I) الذي يعرف أنه عدد المعاملات الحقيقية ذات القيم غير الصفرية والذي تم تقديرها بشكل غير صحيح على انها ذات قيم صفرية عند قيم معامل الارتباط (0.5) و(0.7). في حين أظهرت طريقة (GJO) تباينا في معايير اختيار المتغيرات عند قيمة معامل الارتباط (0.9).

د. ظهرت طريقة (LASSO) كاسوأ طريقة في التقدير لأنها تعطي أعلى قيم لـ (PE وEE) وكذلك كأسوء طريقة في اختيار المتغيرات كونها تميل إلى اختيار متغيرات توضيحية غير مهمة.

جدول (1): معدل معايير تقييم طرائق الجزاء عندما n=50 وP=10

r	Method	PE	EE	C	I
0.5	LASSO	32.3507	2.1488	1	0
	SCAD	10.1541	0.0387	4	0
	GJO	9.8014	0.0281	5	0
0.7	LASSO	29.9037	2.0302	3.5	0
	SCAD	10.7742	0.0783	4	0
	GJO	10.1074	0.0463	5	0
0.9	LASSO	24.1644	1.9384	4	1
	SCAD	12.3546	0.4654	4.5	0
	GJO	11.328	0.2416	5	0

جدول (2): معدل معايير تقييم طرائق الجزاء عندما n=100 , p=10

r	Method	PE	EE	C	I
0.5	LASSO	19.3353	2.0341	2	0
	SCAD	6.9986	0.0699	4	0
	GJO	6.6514	0.0432	5	0
0.7	LASSO	18.6220	1.8992	3	0
	SCAD	7.5870	0.1520	4	0
	GJO	6.8547	0.0890	5	0
0.9	LASSO	14.8017	1.7818	5	1
	SCAD	8.5148	0.6986	5	1
	GJO	7.9514	0.5511	6	1

جدول (3): معدل معايير تقييم طرائق الجزاء عندما n=150 , p=10

r	Method	PE	EE	C	I
0.5	LASSO	8.4941	1.7540	4	0
	SCAD	3.8819	0.1872	4	0
	GJO	3.1247	0.1288	4	0
0.7	LASSO	7.7433	1.6771	4	0
	SCAD	4.2797	0.4467	4	0
	GJO	3.9825	0.3064	5	0
0.9	LASSO	6.2328	1.8516	5	2
	SCAD	4.7625	1.1938	5	2
	GJO	4.2961	1.3040	5	1

7. **الجانب التطبيقي:** لغرض اتمام الفائدة المرجوة من البحث، تم التطبيق على بيانات تتبع توزيع بواسون والتي أخذت من بيانات استخدمت من قبل (لقاء سعيد واخرون، 2011) حول مرض الفشل الكلوي المزمن حيث تم جمع (73) نموذج دم لأشخاص مصابين بمرض العجز الكلوي المزمن والذين يتعالجون بالغسيل الكلوي المستمر، وتم سحب نماذج الدم لمجموعة المرضى من قبل اجراء عملية الغسيل الكلوي التي تستغرق (3-4) ساعات، وقد شخصت حالة المرضى من قبل أطباء مختصين بالتعاون مع مستشفى ابن سينا التعليمي – وحدة الكلية الاصطناعية، تراوحت اعمارهم بين (20-80) سنة، وتتضمن (38) نموذجا للذكور و(35) نموذجا للإناث، ودونت المعلومات الخاصة بالمرضى على وفق استمارة استبيان خاصة لكل مريض أعدت لهذا الغرض لسنة 2013، حيث سجلت الدراسة ثمانية متغيرات توضيحية والتي يعتقد بان لها تأثير في متغير الاستجابة الذي يمثل عدد مرات الغسيل الكلوي بالشهر. يوضح الجدول رقم (4) وصف المتغيرات التوضيحية المستخدمة في الدراسة.

Tikrit Journal of Administrative and Economic Sciences, Vol. 20, No. 68, Part (1): 403-414
Doi: www.doi.org/10.25130/tjaes.20.68.1.23

جدول (4): وصف المتغيرات المستقلة المستخدمة في الدراسة

رمز المتغير التوضيحي	وصف المتغير التوضيحي	وحدة القياس
X1	الجنس	(ذكر =1، انثى =2)
X2	العمر	سنوات
X3	مدة المرض	الايام
X4	الوراثة	(نعم=1، كلا=2)
X5	نسبة اليوريا	(ملي مول/لتر)
X6	نسبة البروتين الكلي	غرام/100 ميليلتر
X7	نسبة الالبومين	غرام/100 ميليلتر
X8	نسبة الكلوبيولين	غرام/100 ميليلتر

يتم تقدير معلمات نموذج انحدار بواسون بواسطة مقدر الإمكان الأعظم (MLE) بغض النظر عن تقدير (B_0) ثم يتم ايجاد قيم$(\widehat{Y})$ لحساب متوسط مربعات الخطأ (MSE) للنموذج، ومن خلال ملاحظة الجدول رقم (5) الذي يوضح نتائج متوسط مربعات الخطأ للنموذج المقدر الذي تم الحصول عليها نلاحظ تفوق طريقة (GJO) على باق طرائق التقدير المستخدمة الأخرى، حيث انها اعطت اقل قيمة لمتوسط مربعات الخطأ مما يجعلها افضل طريقة للتقدير، ثم تأتي بعده طريقة (SCAD) بالمرتبة الثانية من حيث قيمة متوسط مربعات الخطأ، وكذلك كانت طريقتي (MLE , LASSO) أسوأ طريقتين كونها أعطت أعلى قيم لمتوسط مربعات الخطأ.

جدول (5): نتائج الطرق المستخدمة بالاعتماد على معيار MSE في بيانات مرضى العجز الكلوي

Methods	MSE
MLE	9.358487
LASSO	7.877187
GJO	5.1036
SCAD	6.9741

الاستنتاجات والتوصيات

اولاً. الاستنتاجات: لقد أظهرت نتائج المحاكاة والتطبيق العملي أن طريقة GJO هي أفضل من طرائق اختيار المتغيرات الاخرى وتغلبت عليها عندما يكون الارتباط بين المتغيرات (0.5) و(0.7)، حيث امتلكت طريقة GJO أقل قيم معايير (EE, PE, I) وأعلى قيم (C) لجميع نماذج المحاكاة عندما كان معامل الارتباط بين المتغيرات (0.5) و(0.7). كما أظهرت نتائج المحاكاة والتطبيق العملي ان طريقة LASSO هي أسوأ الطرائق، حيث إن طريقة LASSO أعطت أعلى قيم للمعايير (EE, PE, I) وأقل قيم (C) لجميع النماذج عندما كان معامل الارتباط بين المتغيرات (0.5) و(0.7) و(0.9).

المصادر

اولاً. المصادر العربية:

1. الراوي، خاشع محمود، (1978)، مدخل الى تحليل الانحدار، جامعة الموصل.
2. صبري، حسام موفق (2013)، مقارنة طرائق تقدير معلمات انموذج انحدار بواسون في ظل وجود مشكلة التعدد الخطي مع تطبيق عملي، أطروحة دكتوراه، غير منشورة، كلية الإدارة والاقتصاد، جامعة بغداد.
3. عبدالله، لقاء سعيد وعلوش، ذكرى علي والجراح، إسراء عبد الحق، (2011)، دراسة إنزيم ميتالوإندوببتايديز وعلاقته بمرض العجز الكلوي المزمن، مجلة علوم الرافدين، المجلد (22)، العدد (4)، ص (71-87).

ثانياً. المصادر الأجنبية:

1. Mehrabian, A. R., & Lucas, C., (2006), A novel numerical optimization algorithm inspired from weeds colonization. Ecological informatics, 1(4), 355-366.
2. Algamal, Z. Y. and Lee, M. H., (2015), Penalized Poisson Regression Model using adaptive modified Elastic Net Penalty. Electronic Journal of Applied Statistical Analysis,Vol. 08, Issue 02, 236-245.
3. El Anbari, M. and Mkhadri, A., (2013), The adaptive gril estimator with a diverging number of parameters. Communications in Statistics - Theory and Methods. 42(14), 2634–2660.
4. Fan, J., & Li, R., (2001), Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456), 1348-1360.
5. Hossain, S. and Ahmed, E., (2012), Shrinkage and penalty estimators of a Poisson regression model. Australian & New Zealand Journal of Statistics. 54(3), 359–373.
6. James, G., Witten, D., Hastie, T., and Tibshirani, R., (2024), An introduction to statistical learning. Springer, New York.
7. Månsson, K., Kibria, B. G., Sjolander, P & Shukur, G., (2012), Improved Liu Estimators for the Poisson Regression Model, International Journal of Statistics and Probability, Vol. 1, No. 1, pp. 1-6.
8. Tibshirani, R., (1996), Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 58(1), 267–288.
9. Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 101(476), 1418–1429.
10. Zou, H. and Hastie, T., (2005), Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B (Methodological). 67(2), 301–320.