



**Tikrit Journal of Administrative
and Economics Sciences**
مجلة تكريت للعلوم الإدارية والاقتصادية

ISSN: 1813-1719 (Print)



**Variable selection in Poisson regression model using Black Hole
optimization algorithm**

Abed Ali Hamad*^A, Zakariya Yahya Algamal^B

^A College of Administration and Economics, University of Anbar, Anbar, Iraq

^B College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Keywords:

Black hole algorithm, Poisson regression,
selection of variables.

ARTICLE INFO

Article history:

Received 06 Apr. 2023

Accepted 26 Apr. 2023

Available online 30 Aug. 2023

©2023 College of Administration and Economy, Tikrit
University. THIS IS AN OPEN ACCESS ARTICLE
UNDER THE CC BY LICENSE

<http://creativecommons.org/licenses/by/4.0/>



***Corresponding author:**

Abed Ali Hamad

College of Administration and
Economics, University of Anbar,
Anbar, Iraq



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. Invasive weed optimization algorithm (BHA) is one of the recently efficient proposed nature-inspired algorithms that can efficiently be employed for variable selection. In this work, BHA 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.

اختيار المتغيرات في نموذج انحدار بواسون باستخدام خوارزمية الثقب الأسود

زكريا يحيى الجمال
كلية علوم الحاسوب والرياضيات
جامعة الموصل

عبد علي حمد
كلية الادارة والاقتصاد
جامعة الانبار

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 الثقب الأسود ومقارنتها مع طرائق أخرى في اختيار المتغيرات في نموذج انحدار بواسون باستخدام المحاكاة والبيانات الحقيقية وتم استخدام أسلوب مونت – كارلو في المحاكاة لتوليد بيانات تتبع نموذج انحدار بواسون تبعاً 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 كمياري للمقارنة، فقد أظهرت نتائج المحاكاة تفوق خوارزمية الثقب الأسود مقارنة بطرائق اختيار المتغيرات الأخرى. فضلاً عن ذلك، تم التطبيق على بيانات حقيقية تم جمعها من مصابين بمرض العجز الكلوي المزمن والذين يتعالجون بالغسيل الكلوي المستمر، وقد شخّصت حالة المرضى من قبل اطباء مختصين بالتعاون مع مستشفى ابن سينا التعليمي – وحدة الكلية الاصطناعية.

الكلمات المفتاحية: خوارزمية الثقب الأسود، انحدار بواسون، اختيار المتغيرات.

المقدمة

تحليل الانحدار هو أداة احصائية تقوم ببناء أنموذج 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 جداً 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 ان 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، فإنه 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 أصبحت الاساليب التقليدية لاختيار المجموعات الجزئية مثل طريقة الاختيار الامامية (Forward selection)

وطريقة الاختيار إلى الخلف (Backward elimination) وطريقة الاختيار التدريجية (Stepwise selection) غير جيدة في أداء وظيفتها حيث أصبحت أكثر تكلفة في حسابها، فضلاً عن ذلك فإن معايير المعلومات لاختيار المتغيرات مثل معيار أكايكي للمعلومات ((Akaïke information (AIC) ومعيار بيز للمعلومات ((Bayesian information criterion (BIC) أصبحت 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 نموذج انحدار بواسون ((Poisson Regression Model (PRM) الذي يعد أحد النماذج الأكثر شعبية بين النماذج التي لديها متغير استجابة قابل للعد، وقد تم وصفه أولاً عن طريق الباحثين (Nelder and Wedderburn, 1972)، كحالة خاصة من النماذج الخطية المعممة ((Generalized linear models (GLMs)، وللوقوف على مدى أهمية المنهجية مقارنة بالطرائق التقليدية الأخرى سيتم إخضاع النموذج المستخدم ومن ثم توظيف معايير تقييم المعنوية لنتائج كل طريقة.

يهدف هذا البحث إلى توظيف خوارزمية الثقب الأسود ومقارنتها مع طرائق اختيار المتغيرات التوضيحية في نموذج انحدار بواسون أخرى باستخدام المحاكاة والبيانات الحقيقية، من خلال تسليط الضوء على عدد من العوامل التي قد تؤثر على جودة هذه الطرائق ووجوب استخدامها ضمن شروط معينة دون غيرها من الطرائق.

1. 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. وبخلافه من سائر نماذج الانحدار، قد يحتوي النموذج على متغيرات مستقلة كثيرة ما يؤثر سلباً على دقة النموذج وبساطته في تفسير النتائج. يفترض هذا النموذج أن المتغير المعتمد y_i هو متغير استجابة يتبع توزيع بواسون وبمعلمة قدرها (μ) ، كما تتبع الأخطاء العشوائية في النموذج توزيع بواسون بمعلمة قدرها (μ) (Mansson and Kubria (2012) (Hossain And Ahmed, (2012)). ويعرف وفق الدالة الاحتمالية المعرفة بالصيغة الآتية.

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

ويمكن التعبير عنه أيضاً بصيغة المصفوفات وكالتالي:

$$y = \text{Exp}(X\beta + U) \quad \dots (2)$$

إذ أن:

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

لأجل تقدير معلمات نموذج انحدار بواسون باستخدام طرائق الإمكان الجزائية سيتم اللجوء إلى 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 سلفاً بالشكل الآتي:

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

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

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

وبأخذ اللوغاريتم الطبيعي لدالة الإمكان الأعظم للمشاهدات أعلاه نحصل على:

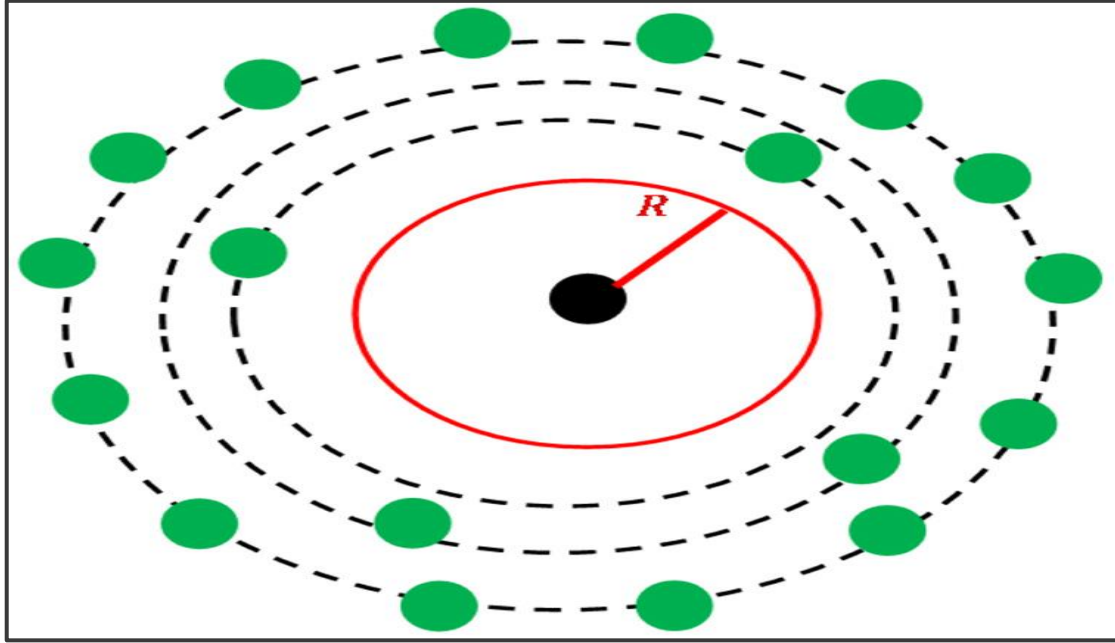
$$\text{Log}L(\mathbf{y}_i | x_i, \boldsymbol{\beta}) = -\sum_{i=1}^n \mu_i + \sum_{i=1}^n y_i (\text{Log}\{\mu_i\}) - \text{Log}\left\{\prod_{i=1}^n y_i!\right\} \quad \dots (4)$$

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

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

2. خوارزمية الثقب الأسود (Black Hole Algorithm (BHA) في السنوات الأخيرة أصبح

الاهتمام متزايد بتصميم خوارزميات التحسين المستوحاة من الطبيعة وتطويرها، وتعد خوارزمية الثقب الأسود (Black Hole Algorithm (BHA واحدة من أحدث الخوارزميات المستوحاة من الطبيعة، وأكثرها شيوعاً وتعود إلى (Hatamlou, 2013)، تحاكي خوارزمية BHA ظاهرة فيزيائية وهي ظاهرة الثقب الأسود في الفضاء؛ لحل مشاكل التحسين من خلال البحث في مساحة المشكلة بطريقة فعالة وبسيطة للغاية، ويتطلب عدداً أقل من المعلومات ووقت أقل، والثقب الأسود: هو أحد أغرب الأجسام الموجودة في الفضاء الخارجي، وأكثرها روعة وهو جسم شديد الكثافة مع امتلاكه لجاذبية قوية، ولأول مرة في عام 1967 أطلق العالم الفيزيائي الأمريكي John Wheeler على ظاهرة انهيار الكتلة اسم الثقب الأسود، حيث يتشكل الثقب الأسود في الفضاء عندما ينهار نجم كبير الحجم وقوة الجاذبية للثقب الأسود تكون عالية جداً لدرجة أنه حتى الضوء لا يستطيع الهروب منه، وتكون الجاذبية قوية جداً لأن المادة قد ضغطت في مساحة صغيرة وأي شيء يعبر حدود الثقب الأسود سيبتلعه ويتلاشى ولا يمكن أن يبتعد عن قوته الهائلة، وتعرف الحدود الكروية للثقب الأسود في الفضاء بأفق الحدث ويطلق على نصف قطر أفق الحدث نصف قطر Schwarzschild عند هذا الشعاع تكون سرعة الهروب مساوية لسرعة الضوء وبمجرد مرور الضوء لا يمكنه الهروب، أي بمعنى آخر إنه لا يمكن لأي شيء أن يهرب من داخل أفق الحدث لأنه لا شيء يمكن أن يكون أسرع من الضوء ويطلق عليه أسود لأنه يمتص كل الضوء ولا يعكس شيئاً، والشكل الآتي يوضح الثقب الأسود، وأفق الحدث ونصف قطر الأفق والنجوم التي حوله (Pashaei & Aydin, 2017).



الشكل (1): مخطط الثقب الأسود (Black Hole) والأفق والنجوم حوله
والخطوات أدناه توضح طريقة محاكاة BHA من ظاهرة الثقب الأسود (Kumar, et al, 2015)
و(Qasim, et.al, 2020):

الخطوة الأولى: تبدأ خوارزمية الثقب الأسود المقترحة (BHA) بمجموعة أولية عشوائية من الحلول المرشحة النجوم (Stars) في مساحة البحث لمشكلة معينة ويتشكل الثقب الأسود في الفضاء الحقيقي عن طريق انهيار النجوم الفردية، ثم تتطور لإيجاد الحل الأمثل ودالة Fitness Function يتم حسابها لكل نجم وفي كل تكرار من BHA يتم اختيار أفضل مرشح في المجتمع بعد تقييم قيم Fitness Function باختيار أفضل قيمة للدالة ؛ ليكون هو الثقب الأسود والباقي يشكل النجوم العادية، وفي المخطط رقم (1) الدائرة السوداء تمثل الثقب الأسود والدوائر الخضراء تمثل النجوم.

الخطوة الثانية: بعد عملية التهيئة تطور خوارزمية BHA الحلول المرشحة نحو الحل الأمثل عبر آلية بسيطة حيث يبدأ الثقب الأسود (المرشح الأفضل) في جذب النجوم (المرشحة الأخرى) من حوله، حيث تبدأ جميع النجوم بالتحرك نحو الثقب الأسود، وعندما يقترب نجم من الثقب الأسود فسيبتلعه الثقب الأسود، ويختفي إلى الأبد، وفي مثل هذه الحالة يتم إنشاء نجم جديد (حل مرشح) بشكل عشوائي ووضعه في مساحة البحث ويبدأ بحثاً جديداً، ويتم حساب الحل المحدث كما في الصيغة الآتية:

$$X_i(t+1) = X_i(t) + \text{rand} * (X_{BH} - X_i(t)) \quad (6)$$

حيث إن:

$X_i(t)$ و $X_i(t+1)$: تمثل مواقع النجم في التكرارات t و $t+1$ ، على التوالي.

X_{BH} : هو موقع الثقب الأسود في مساحة البحث.

rand: هو رقم عشوائي ضمن التوزيع المنتظم [0, 1].

الخطوة الثالثة: يُعرف المجال المحيط بالثقب الأسود في الفضاء الخارجي بأفق الحدث، ويسمى نصف قطر أفق الحدث Schwarzschild radius. توضح الدائرة الحمراء في المخطط (1)

أفق الحدث للثقب الأسود، وفي الفضاء الحقيقي يتم حساب نصف قطر Schwarzschild وفق الصيغة الآتية:

$$R = 2GM / C^2 \quad (7)$$

حيث تشير M و G و C إلى كتلة الثقب الأسود، وثابت الجاذبية، وسرعة الضوء على التوالي. وفي BHA يتم حسابه وفق الصيغة الآتية:

$$R = \frac{f_{BH}}{\sum_{i=1}^N f_i} \quad (8)$$

حيث إن:

$$f_i = \sum_{i=1}^{pop-size} eval(p(t)) \quad (9)$$

$$f_{BH} = \sum_{i=1}^{pop-size} eval(p(t)) \quad (10)$$

حيث إن:

f_{BH} : تمثل قيمة Fitness Function للثقب الأسود.

f_i : تمثل قيمة Fitness Function لكل نجم.

N: هو عدد النجوم (الحلول المرشحة).

الخطوة الرابعة: عندما يعبر النجم أفق الحدث سيبتلعه الثقب الأسود ويختفي؛ بسبب الكثافة الشديدة والجاذبية القوية للثقب الأسود، إذ تكون سرعة الهروب مساوية لسرعة الضوء في منطقة أفق الحدث؛ لذلك لا يمكن ابتعاد أي شيء عن أفق الحدث. في BHA يتم حساب المسافة الإقليدية بين الثقب الأسود والنجم، فإذا كانت هذه المسافة أقل من نصف قطر Schwarzschild، فاستبدل بنجم جديد في الموقع العشوائي في مساحة البحث.

الخطوة الخامسة: في BHA إذا وصل النجم إلى موقع بتكلفة أقل من الثقب الأسود في هذه الحالة يجب استبدال مواقعهم.

الطريقة المقترحة The Proposed Method: من أجل إتمام هدف البحث وتحقيقه، تم اقتراح توظيف خوارزمية الثقب الأسود (BHA) كتقنية مستوحاة من الطبيعة في اختيار المتغيرات، وبالاعتماد على هذه التقنية، فإن كل عنصر (نجم) (Star) في المجموعة سيكون لديه d من المواقع التي تمثل عدد المتغيرات التوضيحية في انموذج انحدار بواسون. بناءً على ذلك، فإن توظيف خوارزمية الثقب الأسود تكون وفق الخطوات التالية:

الخطوة الأولى: تحديد حجم المجموعة (عدد النجوم) وهو 30 نجم، حيث إن كل نجم سيكون له متجه من عدد المتغيرات المستقلة فضلاً عن ذلك تحديد عدد التكرارات داخل خوارزمية الثقب الأسود 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 المثلى، تم الاعتماد على Fitness Function وفق الصيغة الآتية:

$$\text{Fitness Function} = \min \left[\frac{\sum_{i=1}^n (y_i - \hat{m}(\mathbf{X}))^2}{n} \right] \quad (11)$$

الخطوة الرابعة: بالاعتماد على أقل قيمة يحصل عليها أي نجم وفق المعادلة رقم (11) يتم تحديث مواقع باقي النجوم.

الخطوة الخامسة: نستمر بالحل لحين الوصول إلى 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 (2): الية اختيار المتغيرات حسب خوارزمية الثقب الاسود

3. **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: الأول هو تقييم دقة التنبؤ والثاني هو تقييم اختيار المتغيرات.

❖ **معايير تقييم دقة التنبؤ:**

أولاً. خطأ التقدير (EE) (Estimation Error): 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 و β هو متجه المعلمات الحقيقية.
ثانياً. خطأ التنبؤ (PE) (Prediction Error): ويعرف بأنه مربع الفرق بين القيمة الحقيقية لمتغير الاستجابة والقيمة التنبؤية المرافقة له، ويعرف رياضياً بالمعادلة الآتية:

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

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

❖ **معايير تقييم دقة اختيار المتغيرات:** بما إن 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 الآتي:

أولاً. 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.

ثانياً. 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 ثلاثة أحجام من العينات وهي (50,100,150) وذلك لأجل دراسة المقارنة وفق العينات باختلاف أنواعها (صغيرة، متوسطة، كبيرة). 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**

أولاً. تم توليد بيانات المتغير y التي تتبع انموذج انحدار بواسون وكالاتي:

$$y \sim P(\exp(X\beta))$$

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 إن M هي مصفوفة التباين المشترك، حيث إن $M_{ij} = r^{|i-j|}$ ، 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 β الذي ابعاده ($1 \times p$) وكانت قيم متجه معلمات الانحدار β كالاتي

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

حيث إن 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 ومعايير 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 BHA يمكن استخلاص ما يأتي: أ. عندما يتغير معامل الارتباط بين المتغيرات من (0.5) إلى (0.7)، يتبين أن طريقة (BHA) اعطت اقل قيم (EE, PE) حيث بلغ مقدار التحسن بالتنبؤ بالاعتماد على المعيار (PE) بمقدار 68.25% و 2.8.3% عند ($r=0.5$) و 62.57% و 2.91% عند ($r=0.7$) مقارنة بـ (LASSO و SCAD) على الترتيب، كما وبلغ التحسن بخطأ التقدير بالاعتماد على المعيار (EE) بمقدار 98.21% و 46.28% عند ($r=0.5$) و 98.34% و 38.81% عند ($r=0.7$) مقارنة بـ (LASSO و SCAD) على الترتيب.

- ب. عندما يكون معامل الارتباط مساوي إلى (0.9) أعطت طريقة (BHA) أفضل النتائج مقارنة بالطرائق الأخرى حيث تحسن التنبؤ بالاعتماد على المعيار (PE) بمقدار 52.59% و 7.87% مقارنة بـ (LASSO و SCAD) على الترتيب، كما وبلغ التحسن في خطأ التقدير بالاعتماد على المعيار (EE) بمقدار 93.81% و 79.64% مقارنة بـ (LASSO و SCAD) على الترتيب.
- ج. بالاعتماد على معايير اختيار المتغيرات، فقد امتلكت طريقة (BHA) أعلى 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 أنها ذات قيم صفرية عند قيم معامل الارتباط (0.5) و (0.7). في حين أظهرت طريقة (BHA) تبايناً في معايير اختيار المتغيرات عند قيمة معامل الارتباط (0.9).
- د. ظهرت طريقة (LASSO) كأسوأ طريقة في التقدير لأنها تعطي أعلى قيم لـ (PE و EE) وكذلك كأسوأ طريقة في اختيار المتغيرات كونها تميل إلى اختيار متغيرات توضيحية غير مهمة.

الجدول (1): معدل معايير تقييم طرائق الجزاء عندما $n=50$ و $P=10$

r	Method	PE	EE	C	I
0.5	LASSO	32.3507	2.1488	1	0
	SCAD	10.1541	0.0387	4	0
	BHA	8.671	0.0302	5	0
0.7	LASSO	29.9037	2.0302	3.5	0
	SCAD	10.7742	0.0783	4	0
	BHA	9.812	0.0368	5	0
0.9	LASSO	24.1644	1.9384	4	1
	SCAD	12.3546	0.4654	4.5	0
	BHA	10.9257	0.2800	5	0

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

r	Method	PE	EE	C	I
0.5	LASSO	19.3353	2.0341	2	0
	SCAD	6.9986	0.0699	4	0
	BHA	5.9781	0.0392	5	0
0.7	LASSO	18.6220	1.8992	3	0
	SCAD	7.5870	0.1520	4	0
	BHA	6.8209	0.0851	5	0
0.9	LASSO	14.8017	1.7818	5	1
	SCAD	8.5148	0.6986	5	1
	BHA	8.0271	0.5261	6	1

الجدول (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
	BHA	3.1368	0.1215	4	0
0.7	LASSO	7.7433	1.6771	4	0
	SCAD	4.2797	0.4467	4	0
	BHA	3.8914	0.2958	5	0
0.9	LASSO	6.2328	1.8516	5	2
	SCAD	4.7625	1.1938	5	2
	BHA	4.2681	1.2170	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 وآخرون، 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 شخّصت حالة المرضى من قبل أطباء مختصين بالتعاون مع مستشفى ابن سينا التعليمي – وحدة الكلية الاصطناعية، تراوحت أعمارهم بين (20-80) سنة، وتتضمن (38) نموذجاً للذكور و(35) نموذجاً 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.

الجدول (4): وصف المتغيرات المستقلة المستخدمة في الدراسة

رمز المتغير التوضيحي	وصف المتغير التوضيحي	وحدة القياس
X1	الجنس	(ذكر = 1، أنثى = 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 الكلوبولين	غرام/100 ميليلتر

يتم تقدير معلمات نموذج انحدار بواسون بواسطة مقدر الامكان الأعظم (MLE) بغض النظر عن تقدير (B_0) ثم يتم ايجاد قيم (\hat{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 (BHA) على باقي طرائق التقدير المستخدمة الأخرى، حيث إنها أعطت أقل قيمة لمتوسط مربعات الخطأ مما يجعلها أفضل 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.

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

Methods	MSE
MLE	9.358487
LASSO	7.877187
BHA	5.6518
SCAD	6.9741

8. الاستنتاجات لقد أظهرت نتائج المحاكاة والتطبيق العملي أن طريقة BHA هي أفضل من طرائق اختيار المتغيرات الأخرى وتغلبت عليها عندما يكون الارتباط بين المتغيرات (0.5) و (0.7)، حيث امتلكت طريقة BHA 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 أن 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

أولاً. 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. عبد الله، لقاء سعيد وعلوش، ذكرى علي والجراح، إسراء عبد الحق، (2011)، دراسة إنزيم ميتالو إندو ببتايديز 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 weed 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., (2013), *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.
11. Hatamlou, A., (2013), Black hole: A new heuristic optimization approach for data clustering, *Information sciences*, 222, pp [175-184].
12. Pashaei, E., & Aydin, N., (2017), Binary black hole algorithm for feature selection and classification on biological data, *Applied Soft Computing*, 56, pp [94-106].
13. Kumar, S., Datta, D., & Singh, S. K., (2015), *Black hole algorithm and its applications Computational intelligence applications in modeling and control*: Springer, pp [147-170].
14. Qasim, O. S., Al-Thanoon, N. A., & Algamal, Z. Y., (2020), Feature selection based on chaotic binary black hole algorithm for data classification, *Chemometrics and Intelligent Laboratory Systems*, 204.