المجلة العراقية للعلوم الإحصائية (30) 2019
ص ص [39-54]

اختيار المتغيرات في أنموذج انحدار بواسون باستخدام خوارزمية الاعشاب الضارة

غادة يوسف اسماعيل عبدالله د.زكريا يحيى نوري الجّمال
zakariya.algamal@uomosul.edu.iq

المستخلص:

يعد أنموذج انحدار بواسون واحداً من 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 الأنموذج على متغيرات مستقلة كثيرة ما يؤثر سلباً في دقة الأنموذج وبساطته في تفسير النتائج. تهدف هذه الدراسة إلى استخدام خوارزمية الأعشاب الضارة ومقارنتها مع طرائق أُخرى في اختيار المتغيرات في نموذج انحدار بواسون بآستخدام المحاكاة والبيانات الحقيقة 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: الأَول تقييم دقة التنبؤ، والثاني هو تقييم اختيار المتغيرات كمعيار للمقارنة، فقد أظهرت نتائج المحاكاة تفوق خوارزمية الأعشاب الضارة مقارنةَ بطرائق اختيار المتغيرات الأخرى. اضافة الى ذلك، وتم التطبيق على بيانات حقيقية جُمعت من مصابين بمرض العجز الكلوي المزمن، والذين يتعالجون بالغسيل الكلوي المستمر، وقد شخص حالة المرضى من قبل أطباء مختصون بالتعاون مع مستشفى ابن سينا التعليمي – وحدة الكلية الاصطناعية.

……

……

Variable selection in Poisson regression model using invasive weed optimization algorithm

Abstracts:-

Variable selection is a very helpful procedure for improving prediction accuracy by finding the most important variables that are related to the response variable. Poisson regression model has received much attention in several science fields for modeling count data. Invasive weed optimization algorithm (IWO) is one of the recently efficient proposed nature-inspired algorithms that can efficiently be employed for variable selection. In this work, IWO 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.

* مدرس مساعد/ قسم التقنيات الميكانيكية / المعهد التقني
**استاذ مساعد/ قسم الاحصاء والمعلوماتية/ كلية علوم الحاسوب والرياضيات

تاريخ استلام البحث 2019/3/14 تأريخ القبول 4/17 /2019 تاريخ النشر 2019/12/1

1- المقدمة

تحليل الانحدار هو أَداة احصائية تقوم ببناء أُنموذج احصائي، وذلك لتقدير العلاقة بين متغير واحد يدعى المتغير المعتمد ومتغير آخر او عدة متغيرات اخرى 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، وأَصبح دورها مهماً جداً في تطبيقات علوم الحياة المتنوعة خصوصاً في المجال الاقتصادي الذي أخذ على عاتقه اعتماد نماذج الانحدار بالدرجة الأَساس لتكون أَبرز وسائل الدعم العملي للنظريات الاقتصادية، فضلاً عن العلوم الأُخرى كالصحية والحياتية والاجتماعية وغيرها (الراوي،1978) .

يفترض أنموذج الانحدار الخطي الكلاسيكي أَنَّ متغير الاستجابة يعتمد على مجموعة من المتغيرات التوضيحية، إذ يمكن أن تكون هذه المتغيرات عبارة عن متغيرات مستمرة أو متغيرات قابلة للعد، ومع ذلك وعندما 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 كبيراً، فقد أَصبحت الأَساليب التقليدية لاختيار المجموعات الجزئية، مثل: طريقة الاختيار الامامية (Forward selection) ، طريقة الاختيار الى الخلف (Backward elimination) ، طريقة الاختيار التدريجية (Stepwise selection) غير جيدة في أَداء وظيفتها ،فقد أصبحت أكثر تكلفة في حسابها ، إضافة الى ذلك فإن 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 أولاً عن طريق الباحثَين Nelder and Wedderburn (1972)، كحالة خاصة من النماذج الخطية المعممة Generalized linear models (GLMs)، وللوقوف على

مدى اهمية المنهجية مقارنة بالطرائق التقليدية الأُخرى سيتم اخضاع الأنموذج المستخدم ، ثم توظيف معايير تقييم المعنوية لنتائج كل طريقة.

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

2- أُنموذج انحدار بواسون

يعد أُنموذج 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، فإنه يحتوي على متغيرات مستقلة كثيرة تؤثر سلباً في دقة الأنموذج وبساطته في تفسير النتائج. ويفترض 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 ((2012),Hossain And Ahmed) . (2012) ، ويعرف وفق الدالة الاحتمالية المعرفة بالصيغة الآتية.

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

ويمكن التعبير عنه أَيضاً بصيغة المصفوفات وكالآتي:

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

إذ إن:

$\mathbf{y}$: موجه المتغير التابع ذو 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 ذو الدرجة $(n\times1)$ n : حجم العينة P : عدد المتغيرات المستقلة (التوضيحية).

لأجل تقدير معلمات أنموذج انحدار بواسون بآستخدام 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 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 في 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 في أعلاه نحصل على:

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

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

$$LogL(\mathbf{y_i}|x_i, \boldsymbol{\beta}) = -\sum_{i=1}^{n}\left(\mathrm{Exp}\{x_i^T\boldsymbol{\beta}\}\right) + \sum_{i=1}^{n} y_i\left(\mathrm{Log}\left\{\mathrm{Exp}\{x_i^T\boldsymbol{\beta}\}\right\}\right) - \mathrm{Log}\left\{\prod_{i=1}^{n} y_i\,!\right\}$$

$$= \sum_{i=1}^{n}\left(y_i x_i^T\boldsymbol{\beta} - EXP\left(x_i^T\boldsymbol{\beta}\right) - \log y_i!\right) \qquad \dots (5)$$

3- خوارزمية الأَعشاب الضارة

خوارزمية أمثلة الأعشاب الضارة Invasive Weed Optimization Algorithm (IWO)هي خوارزمية التحسين العشوائي العددي المستوحات بيولوجياً من الأعشاب الضارة ،التي اقترحها أَولّ مرة Mehrabian و Lucas في عام (2006 م) . وهذه الخوارزمية ببساطة تحاكي السلوك الطبيعي للأعشاب الضارة في الاستعمار، وايجاد مكان مناسب للنمو والتكاثر. ولمحاكاة السلوك الاستعماري للأعشاب الضارة يجب ان تؤخذ بعض الخصائص الأَساسية لهذه العملية بنظر الاعتبار:

1. يتم نشر عدد محدود من البذور في منطقة البحث (تهيئة عدد السكان).
2. كل البذور تنمو على شكل نباتات مزهرة وتنتج البذور اعتماداً على دالة اللياقة (التكاثر).
3. البذور المنتجة يتم نشرها عشوائياً على منطقة البحث، لتنمو وتصبح نباتات جديدة (التشتت المكاني)
4. تستمر هذه العملية إلى أَنْ يتم الوصول الى الحد الأقصى من عدد النباتات.

النباتات ذات دالة اللياقة العالية وحدها يمكنها البقاء على قيد الحياة وإنتاج البذور، ويجري القضاء على الاخرين (الإقصاء التنافسي). وتستمر العملية إلى أَن يتم الوصول الى الحد الأقصى من التكرارات على أَمل أَن النبات الذي يحمل أفضل دالة لياقة سيكون هو الأَقرب إِلى الحل الأَمثل

تتضمن خوارزمية أمثلة الأعشاب الضارة (IWO) عدداً من الخطوات الاساسية، وهذه الخطوات مترابطة بعضها مع البعض ، ولا يمكن تطبيق هذه الخوارزمية على أية مسالة مالم تطبق هذه الخطوات جميعها والا ستفقد خوارزمية أمثلة الأعشاب الضارة (IWO) قيمتها وفائدتها في إِيجاد وتحسين الحل، ويمكن توضيح خطوات الخوارزمية على النحو الآتي

الخطوة الاولى: تهيئة المجتمع الابتدائي Initialize A Population

يتم توليد مجتمع ابتدائي من الحلول ونشرها على d من الأَبعاد من مساحة المشكلة مع مواقع عشوائية، وحساب قيمة دالة اللياقة لهذا المجتمع.

الخطوة الثانية: التكاثر Reproduction

يسمح للنبات في مجتمع النباتات بإنتاج البذور seed (التكاثر) ، وذلك اعتماداً على قيمة دالة اللياقة الخاصة به، وكذلك الحد الأعلى و الأدنى لدالة اللياقة في المستعمرة، اذ يزداد عدد البذور التي ينتجها النبات خطياً من الحد الأدنى الممكن لإنتاج البذور الى أقصى حد ممكن ، وبعبارة اخرى فان النبات ينتج البذور اعتماداً على قيمة دالة اللياقه الخاصة به ، واقل دالة لياقة للمستعمرة واعلى دالة لياقة للمستعمرة ، وذلك للتاكد من أَنَّ الزيادة تكون خطية

المعادلة في أَدناه توضح عملية التكاثر للأعشاب الضارة :

$$seed_i = floor\left(\frac{f_i - f_{\min}}{f_{\max} - f_{\min}}\left(S_{\max} - S_{\min}\right)\right) + S_{\min} \qquad (6)$$

إذ إن Floor تدل على أَن البذور تقرب لأَقرب عدد صحيح، f_i تمثل دالة اللياقة لـ i من الأعشاب الضارة، $f_{\min} and\ f_{max}$: تمثل الحد الأقصى والأدنى لقيمة دالة اللياقة في لمستعمرة وان $S_{\max} and\ S_{\min}$ تمثل الحد الأقصى والأدنى لعدد البذور التي سوف تنتج في المستعمرة.

تمثل الصيغة في اعلاه العلاقة الرياضية بين عدد البذور وقيمة دالة اللياقة للأعشاب الضارة، اذ ينخفض عدد البذور مع زيادة قيمة دالة اللياقة ،وعدد البذور يتراوح بين ال $S_{\max}$ و $S_{\min}$. وتعدّ الأفراد القابلة للتكاثر هي تلك الأفراد ذوات أفضل قيمة لدالة اللياقة من الأفراد غير الملائمة للاستخدام، وتعني كلمة "أفضل" هنا أَنَّ لهذه الأفراد فرصة أَكبر للبقاء على قيد الحياة والتكاثر . لذا لايسمح للأفراد غير الملائمة للاستخدام بالتكاثر. ومع ذلك فإن وجهة النظر هذه تتجاهل شيئاً مهماً الا وهو أَنَّ الخوارزمية التطورية هي طريقة احتمالية وتكرارية ، فمن الممكن ان بعض الأفراد غير الملائمة للاستعمال تحمل في داخلها معلومات أَكثر فائدة من الأفراد الملائمة خلال عملية التطور . فضلاً عن ذلك وغالبا ما يستطيع النظام الوصول الى النقطة المثلى إذا كان بالإمكان عبور المنطقة غير قابلة للتطبيق (وخاصة في فضاء البحث غير المحدب). لذا اقترحت تقنية التكاثر في أعلاه لإعطاء فرصة أكبر للأفراد غير الملائمة للاستخدام للبقاء على قيد الحياة ، وهذه العملية مماثلة للآلية التي تحدث في الطبيعة.

الخطوة الثالثة: التشتت المكاني Spatial Dispersal

توفر هذه الخطوة لخوارزمية الأعشاب الضارة خاصيتي العشوائية والتكيف، اذ يتم توزيع البذور المتولدة عشوائياً على d من الأَبعاد في فضاء البحث بواسطة أَرقام عشوائية تتوزع توزيعاً طبيعياً بمعدل (μ=0) وتباين متغير. وهذا يعني أَنَّ البذور سيتم توزيعها عشوائياً بحيث أَنها تقع بالقرب من النبات الأَم. الا أَنَّ الانحراف المعياري Standard deviation (SD) (σ) للدالة العشوائية

سيخفض من قيمة اولية محددة مسبقا $(\sigma_{initial})$ إلى قيمة نهائية (σ_{final}) في كل خطوة (كل جيل)، من خلال المعادلة الآتية:

$$\sigma_{iter} = \frac{(iter_{\max} - iter)^n}{(iter_{\max})^n}(\sigma_{initial} - \sigma_{final}) + \sigma_{final} \quad (7)$$

اذ إن σ_{iter} يمثل الانحراف المعياري في الخطوة الحالية، $iter_{\max}$ يمثل الحد الأقصى من التكرارات، وان n يمثل معدل التاشير غير الخطي. يضمن هذا التحويل ان احتمالية اسقاط البذور في منطقة بعيدة في هذا التحويل ينخفض بشكل غير خطي في كل خطوة زمنية، مما يؤدي الى تجميع النباتات المجربة وازالة النباتات غير الملائمة.

ويتم حساب موقع البذور الجديدة باستخدام المعادلة الآتية:

$$x_{son} = x_{parent} + sd = x_{parent} + random * \sigma_{iter} \quad (8)$$

إذ إَن x_{son} يمثل موقع الذرية ، وإن x_{parent} يمثل موقع الاب، في حين Random يمثل توليد اعداد عشوائية من التوزيع الطبيعي القياسي محصورة ضمن الفترة [0,1].

الخطوة الرابعة: الإقصاء التنافسي Competitive Exclusion

إذا كان النبات لايترك أي نسل فسوف ينقرض من الوجود، لذا دعت الحاجة الى نوع من التنافس بين النباتات للحد من العدد الأقصى من النباتات في المستعمرة . وبعد مرور بعض التكرارات فإن عدد النباتات في المستعمرة تصل الى الحد الأقصى عن طريق التكاثر السريع، ومع ذلك فمن المتوقع أَن يتم استنساخ النباتات المجربة أَكثر من النباتات غير الملائمة . عند الوصول الى الحد الأقصى لعدد النباتات في المستعمرة $P_{\max}$ ، فسوف تنشط آالية إقصاء النباتات ذات دالة اللياقة الضعيفة لذلك الجيل. اذ تعمل آلية الإقصاء على النحو الآتي: عندما يتم الوصول الى الحد الأقصى لعدد الأعشاب في المستعمرة يسمح لكل عشب بإنتاج البذور ،وفقا للآلية المذكورة في الخطوة (2) (التكاثر)، ثم يتم السماح للبذور المنتجة بالانتشار في منطقة البحث وفقا للخطوة (3) (التشتت المكاني). وعندما تجد جميع البذور مواقعها في منطقة البحث يتم ترتيبها مع آبائها (كمستعمرة من الأعشاب الضارة). بعد ذلك يتم القضاء على الأعشاب الضارة ذات دالة اللياقة

المنخفضة للوصول الى الحد الأقصى المسموح به للمجتمع في المستعمرة . وبهذه الطريقة ترتب النباتات وذريتها معاً والعنصر الأفضل دالة لياقة سينجو ويبقى على قيد الحياة مع السماح لعملية التكرار داخل الخوارزمية . وكما ذكر سابقاً في الخطوة (2) فإن هذه الآلية تعطي فرصة للنباتات ذات دالة اللياقة المنخفضة لإعادة الإنتاج فإن كانت ذريتها ذات دالة لياقة جيدة في المستعمرة فستنجو وتبقى على قيد الحياة، بعبارة أُخرى لايتم إقصاؤها. وتطبق آلية التحكم بالمجتمع على الذرية ايضاً لحين انتهاء مرحلة معينة مما يحقق الإقصاء التنافسي. والشكل 1 يوضح آلية عمل الخوارزمية في اختيار المتغيرات.

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

الشكل 1: يوضح آلية اختيار المتغيرات حسب خوارزمية الأعشاب الضارة

4- 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. وبشكل عام هناك جانبان من جوانب تقييم أداء الطرائق الجزائية الاول: هو تقييم دقة التنبؤ ، والثاني: هو تقييم اختيار المتغيرات.

4-1 معايير تقييم دقة التنبؤ

أولا: خطأ التقدير (EE) (Estimation Error)

ويعرف بأنه مربع الفرق بين قيمة المعلمات الحقيقية وقيمة المعلمات المقدرة، ويعرف بالشكل الرياضي الآتي :

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

إذ ان $\hat{\beta}$:هو متجه المعلمات المقدرة وفق الطرائق المستخدمة ، 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 رياضيا بالمعادلة الآتية :

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

إذ إن $\hat{y} = \text{Exp}\{\text{X}^T \boldsymbol{\beta}\}$ ، وبالاعتماد على هذين المعيارين يتم تحديد الطريقة الافضل التي تعطي أقل قيمة مقارنة بالطرائق الاخرى.

2-4 معايير تقييم دقة اختيار المتغيرات

بما أَنَّ 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 الآتي:

اولاً : معيار التقييم"C"

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

ثانياً : معيار التقييم "I"

معيار التقييم الذي يرمز له بـ**(I)** ، وهو يعرف بأَنه عدد المعاملات الحقيقية ذات القيم غير الصفرية الذي تم تقديرها بشكل غير صحيح على أَنها ذات قيم صفرية. وتعتمد جودة طرائق الجزاء من ناحية معايير تقييم دقة اختيار المتغيرات على من يعطي أَعلى قيمة لـ **(C)** وأَقل قيمة لـ **(I)** .

5- وصف تجربة المحاكاة Description Simulation Experiment

لقد تم تصميم تجربة ومحاكاتها بآستعمال لغة البرمجة (R) إذ تم توليد المتغير (y_i) في أنموذج انحدار بواسون الذي يتبع توزيع بواسون بمعدل مقداره (μ_i) ، تم استخدام أُسلوب مونت كارلو (Mont Carlo) في المحاكاة فقد تم تعيين قيم حجم العينات (n) تم استخدام ثلاثة احجام من العينات وهي (50,100,150) ، لأجل دراسة المقارنة وفق العينات باختلاف أنواعها (الصغيرة، والمتوسطة، والكبيرة). وسوف تتم المقارنة مع كل من طريقة LASSO ،التي تمثل مختصر Least Absolute Shrinkage and Selection Operator وطريقة SCAD، والتي تمثل مختصر Smoothly Clipped Absolute Deviation .

6- دراسات المحاكاة Simulation Studies

اولا : تم توليد بيانات المتغير y التي تتبع أنموذج انحدار بواسون وكالاتي :

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

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

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

إذ إنَّ $\mathbf{M}$ هي مصفوفة التباين المشترك، وان $_{ij} = \boldsymbol{r}^{|i-j|}$ ، عندما $(i, j = 1,2, \dots, p)$ وان المتغيرات التوضيحية تكون مرتبطة.

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

رابعاً : تم توليد بيانات أُنموذج انحدار بواسون تبعاً لقيم متجه معلمات الانحدار $\boldsymbol{\beta}$ الذي أَبعاده $(1 \times p)$، وكانت قيم متجه معلمات الانحدار $\boldsymbol{\beta}$ كالاتي $\boldsymbol{\beta} = (\mathbf{1.2}, -\mathbf{0.6}, \mathbf{0.8}, -\mathbf{0.4}, \mathbf{1.5}, \mathbf{0}, \dots, \mathbf{0})^{\mathrm{T}}$ ، إذ إن المعلمات غير الصفرية عددها $q = 5$ ، وان المعلمات الصفرية تساوي $p - q$.

7- تفسير نتائج المحاكاة

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

1- عندما يتغير معامل الارتباط بين المتغيرات من (0.5) الى (0.7)، يتبين أَنَّ طريقة (IWO) أَعطت أَقل قيم (EE , PE) فقد بلغ مقدار التحسن بالتنبؤ بالاعتماد على المعيار (PE) بمقدار 69.36% و 2.3% عند (r=0.5) و 63.63% و 2.83% عند (r=0.7) مقارنة بـ(LASSO وSCAD) على الترتيب، كما بلغ التحسن بخطأ التقدير بالاعتماد على المعيار (EE) بمقدار 99.01% و 45.22% عند (r=0.5) و 97.66% و 39.34% عند (r=0.7) مقارنة بـ(LASSO وSCAD) على الترتيب.

2- عندما يكون معامل الارتباط مساوياً لـ (0.9) فقد اعطت طريقة (IWO) أفضل النتائج مقارنة بالطرائق الأُخرى، إذ تحسن التنبؤ بالاعتماد على المعيار (PE) بمقدار 52.59% و 7.28% مقارنة بـ(LASSO و SCAD) على الترتيب ، كما بلغ التحسن في خطأ التقدير بالاعتماد على المعيار (EE) بمقدار 94.90% و 78.77% مقارنة بـ(LASSO و SCAD) على الترتيب.

3- بالاعتماد على معايير اختيار المتغيرات فقد امتلكت طريقة (IWO) أَعلى قيم (C) الذي هو عدد المعاملات الحقيقية ذات القيم الصفرية، والتي تم تقديرها بشكل صحيح على أَنها ذات قيم صفرية، وأَعطت أَقل قيم (I) الذي يعرف بأَنه 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 أَظهرت طريقة (IWO) تبايناً في معايير اختيار المتغيرات عند قيمة معامل الارتباط (0.9).

4- ظهرت طريقة (LASSO) كاسوأ طريقة في التقدير، لأنها تعطي أعلى قيم لـ (PE و EE) وكذلك كاسوأ طريقة في اختيار المتغيرات كونها تميل الى اختيار متغيرات توضيحية غير مهمة.

الجدول (1) : يوضح 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
	IWO	9.9122	0.0212	5	0
0.7	LASSO	29.9037	2.0302	3.5	0
	SCAD	10.7742	0.0783	4	0
	IWO	10.5771	0.0475	5	0
0.9	LASSO	24.1644	1.9384	4	1
	SCAD	12.3546	0.4654	4.5	0
	IWO	11.8954	0.2600	5	0

الجدول (2) :يوضح 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
	IWO	6.7172	0.0432	5	0
0.7	LASSO	18.6220	1.8992	3	0
	SCAD	7.5870	0.1520	4	0
	IWO	7.3276	0.0890	5	0
0.9	LASSO	14.8017	1.7818	5	1
	SCAD	8.5148	0.6986	5	1
	IWO	8.3226	0.5511	6	1

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

r	Method	PE	EE	C	I
0.5	LASSO	8.4941	1.7540	4	0
	SCAD	3.8819	0.1872	4	0
	IWO	3.6292	0.1288	4	0
0.7	LASSO	7.7433	1.6771	4	0
	SCAD	4.2797	0.4467	4	0
	IWO	4.0922	0.3064	5	0
0.9	LASSO	6.2328	1.8516	5	2
	SCAD	4.7625	1.1938	5	2
	IWO	4.8221	1.3040	5	1

8-الجانب التطبيقي

لغرض اتمام الفائدة المرجوة من البحث ، تم التطبيق على بيانات تتبع توزيع بواسون، التي أخذت من بيانات استخدمها من قبل (لقاء سعيد واخرون،2011) حول مرض الفشل الكلوي المزمن فقد 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 (4-3) ساعات، وقد شخص حالة المرضى اطباء مختصون بالتعاون مع مستشفى ابن سينا التعليمي – وحدة الكلية الاصطناعية، وتراوحت أعمارهم بين (80-20) سنة ، وتتضمن (38) أنموذجاً للذكور و (35) أنموذجاً للإناث، ودونت المعلومات الخاصة بالمرضى على وفق استمارة استبيان خاصة لكل مريض اعدت لهذا الغرض لسنة 2013، فقد سجلت الدراسة ثماني متغيرات توضيحية التي يعتقد بأَنَّ لها تأثيراً في متغير الاستجابة الذي يمثل عدد مرات الغسيل الكلوي بالشهر. ويوضح الجدول (4) وصف المتغيرات التوضيحية المستخدمة في الدراسة.

الجدول (4) : يوضح 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 أنموذج انحدار بواسون بواسطة مقدر الإمكان الأَعظم (MLE) بغض النظـر عن تقدير (B_0) ، ثم يتم ايجاد قيم $(\hat{Y})$ لحساب متوسط مربعات الخطأ (MSE) للنموذج، ومن خلال ملاحظة الجدول (5) الذي يوضح نتائج متوسط مربعات الخطأ للأُنموذج المقدر الذي تم الحصول عليها نلحظ تفوق طريقة (IWO) على باقي طرائق التقدير المستخدمة الاخرى، فقد أعطت أقل قيمة لمتوسط مربعات الخطأ مما جعلها أَفضل طريقة للتقدير، ثم تأتي بعده طريقة (SCAD) بالمرتبة الثانية من حيث قيمة متوسط مربعات الخطأ، وكذلك كانت طريقتا (MLE , LASSO) أسوأ طريقتين كونها أَعطت أَعلى قيم لمتوسط مربعات الخطأ.

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

Methods	MSE
MLE	9.358487
LASSO	7.877187
IWO	**5.8744**
SCAD	6.9741

9-الاستنتاجات

أَظهرت نتائج المحاكاة والتطبيق العملي أَنَّ طريقة IWO هي أَفضل من طرائق اختيار المتغيرات الاخرى وتغلبت عليها عندما كان الارتباط بين المتغيرات (0.5) و (0.7)، فقد امتلكت طريقة IWO اقل قيم معايير (EE ,PE, I) وأَعلى قيم (C) لجميع نماذج المحاكاة عندما كان معامل الارتباط بين المتغيرات (0.5) و (0.7). كما أَظهرت نتائج المحاكاة والتطبيق العملي أَنَّ طريقة LASSO هي أسوأ الطرائق ، إذ إن طريقة LASSO أَعطت أَعلى قيم للمعايير (EE PE, I ,) وأقل قيم (C) لجميع النماذج عندما كان معامل الارتباط بين المتغيرات (0.5) و (0.7) و (0.9) .

10-المصادر

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 و الجراح، إسراء عبد الحق ،(2011)، "**دراسة إنزيم ميتالوإندوببتايديز وعلاقته بمرض العجز الكلوي المزمن**"، مجلة علوم الرافدين، المجلد (22)، العدد (4)، ص (71-87).
4. 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, 236-24.
5. 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.

.6 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.

.7 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.

.8 James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). "**An introduction to statistical learning**". Springer, New York.

.9 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.

.10 Mehrabian, A. R., & Lucas, C. (2006). A novel numerical optimization algorithm inspired from weed colonization. Ecological informatics, 1(4), 355-366

.11

.12 Tibshirani, R. (1996). "**Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society**". Series B (Methodological). 58(1), 267–288.

.13 Zou, H. (2006)." **The adaptive lasso and its oracle properties. Journal of the American Statistical Association**". 101(476), 1418–1429.

.14 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.