AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

توظيف الخوارزميات المستوحاة من الطبيعة في اختيار حزمة العرض في مقدر متعدد الحدود

م.م. مروة يحيى مصطفى	أ.د. زكريا يحيى الجمال
marwa.yahya@uomosul.edu.iq	zakariya.algamal@uomosul.edu.iq

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

المستخلص

يلقى موضوع تحليل الانحدار اهتماماً واسعاً وواضحاً في معظم الدراسات وخصوصاً الاقتصادية والطبية منها. ويعد أنموذج الانحدار اللامعلمي بصورة عامة والانحدار اللامعلمي المتعدد بوجه خاص أحد أهم نماذج الانحدار المستخدمة في السنوات الأخيرة التي شهدت توسعاً كبيراً وخصوصاً في الجانب الاقتصادي والبيئي. إذ يعدّ المقدر الممهد واللُبي الموضعي متعدد الحدود (Local Polynomial Kernel Smoother) من أهم المقدرات المستعملة في أنموذج الانحدار اللامعلمي المتعدد. حيث أن هذا المقدر يعتمد بدوره في تقدير أنموذج الانحدار اللامعلمي المتعدد على مصفوفة معلمات تسمى بمعلمات التمهيد والتي لتقديرها أهمية كبيرة في جودة توفيق المنحى المقدر في أنموذج الانحدار اللامعلمي المتعدد. تم في هذا البحث اقتراح توظيف خوارزمية مستوحاة من الطبيعة والمتمثلة بخوارزمية الاعشاب الضارة في عملية تقدير مصفوفة معلمات التمهيد في المقدر الموضعي متعدد الحدود. تم استخدام أسلوب مونت – كارلو في المحاكاة لتوليد بيانات تتبع عدد من نماذج الانحدار اللامعلمي المتعدد. لقد أظهرت نتائج المحاكاة بالاعتماد على متوسط مربعات الخطأ بوضعها معياراً للمقارنة تفوق الطريقة المقترحة مقارنةً بطرائق التقدير الأخرى.

معلومات البحث

تواريخ البحث

تاريخ تقديم البحث: 2022/12/10

تاريخ قبول البحث: 2023/2/26

تاريخ رفع البحث على الموقع: 2023/12/31

الكلمات المفتاحية

مقدرات النواة، مصفوفة عرض الحزمة، المقدر اللُبي الموضعي متعدد الحدود، خوارزمية الاعشاب الضارة

للمراسلة:

م.م. مروة يحيى مصطفى

marwa.yahya@uomosul.edu.iq

https://doi.org/10.55562/jrucs.v54i1.576

المقدمة

يعد تحليل الانحدار(Regression Analysis) أحد الأساليب الإحصائية المهمة لدراسة العديد من الظواهر الطبيعية والاقتصادية والاجتماعية والطبية وغيرها، إذ يستخدم في تمثيل العلاقة بين المتغيرات العشوائية المختلفة بالنسبة إلى عينة معينة أو بالنسبة إلى المجتمع على هيئة معادلة إحصائية لتحقيق الكثير من الأهداف المهمة التي يتوصل إليها من خلال تلك العلاقة. تعد أساليب الانحدار مفيدة في عملية بناء النماذج الإحصائية، حيث صنفت نماذج الانحدار إلى صنفين أساسيين بحسب طبيعة البيانات أولاً: نماذج الانحدار المعلمي (Parametric Regression Models) وثانياً: نماذج الانحدار اللامعلمي (Nonparametric Regression Models) (Rencher, 2002). ان نماذج الانحدار اللامعلمي (Nonparametric Regression Models) تقوم على إيجاد العلاقة بين متغير الاستجابة والمتغيرات التوضيحية من خلال منحنى يصف تلك العلاقة، لذا فإنه في الانحدار اللامعلمي يكون الباحث مهتماً بإعطاء وصف عام للعلاقة وليس دراسة تفاصيل العلاقة الدقيقة. وإن دالة العلاقة تكون غير معروفة وهذه النماذج تكون أكثر مرونة ولا تعتمد على فروض سابقة بل تعتمد بشكل أساسي ومباشر على البيانات (Data) حيث إن نوع البيانات يفسر الشكل الفعلي لمنحنى الانحدار (محمد، 2011).

إن تحديد مصفوفة عرض الحزمة (Bandwidth) أو ما تسمى بمصفوفة معلمات التمهيد (H) في دالة (Kernel) ذات أهمية كبيرة في تحديد وتقريب شكل دالة الانحدار اللامعلمي الى الدالة الأصلية، من خلال إيجاد الطريقة المثلى للموازنة بين التحيز والتباين. فعندما تكون قيمة معلمة التمهيد صغيرة فإن التحيز يكون صغيراً، والتباين يكون كبيراً، وهذا يؤدي إلى خشونة

شكل الدالة وبعكسه نحصل على تعظيم التحيز، وتصغير التباين ويكون شكل الدالة أكثر تمهيداً (نعومة) ويتم الاختيار الجيد لقيم مصفوفة المعلمة التمهيدية (H) من خلال المفاضلة بين التحيز والتباين للحصول على أصغر متوسط مربعات خطأ (MSE).

يهدف هذا البحث الى توظيف تقنية ذكائية وهي خوارزمية الاعشاب الضارة Invasive Weed Optimization Algorithm (IWO) في تقدير قيم مصفوفة معلمة التمهيد (H) لكي تكون أكثر كفاءة مقارنة مع الطرائق الأخرى، وتعمل على تحسين النتائج في تقدير قيم هذه المصفوفة من خلال تجارب المحاكاة والتطبيق على بيانات حقيقية.

المنهجية Methodology

I. المقدرات اللامعلمية Nonparametric estimator

إن المرونة العالية التي تتمتع بها المقدرات اللامعلمية قياساً بالمقدرات المعلمية، وتطور أجهزة الحاسبات أدى إلى ميل الباحثين في العقود الأخيرة للاهتمام بموضوع الانحدار اللامعلمي، وطرائق التمهيد المختلفة، ومن أنواع نماذج الانحدار اللامعلمي هو أنموذج الانحدار اللامعلمي البسيط عندما يكون لدينا متغير توضيحي واحد وصيغته كالآتي (Koláček & Horová, 2017):

$$\mathbf{y}_i = m(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i \qquad i = 1, 2, \ldots, n \tag{1}$$

إذ إن:

$m(\mathbf{x_i})$: تمثل دالة الانحدار اللامعلمي غير المعروفة الواجب تقديرها بالطرائق اللامعلمية.

وكذلك أنموذج الانحدار اللامعلمي المتعدد (Multivariate Nonparametric Regression Model) الذي يقوم على إيجاد العلاقة بين متغير الاستجابة، وعدد من المتغيرات التوضيحية، وهذه العلاقة تكون غير معروفة ويتم تقديرها باستعمال عدة طرائق وهي: طريقة التقدير اللُبي Kernel estimation، وطريقة الشرائح التمهيدية Smoothing Splines، وطريقة المويجة Wavelet estimator ، وتكون صيغة معادلة الانحدار اللامعلمي المتعدد كالآتي (محمد و عبدالحسن، 2018):

$$\mathbf{y}_i = \mathbf{m(x_{1i}, x_{2i}, \ldots, x_{di})} + \boldsymbol{\varepsilon}_i \qquad i = 1, 2, \ldots, n \tag{2}$$

حيث أن:

$\mathbf{X = (x_{1i}, x_{2i}, \ldots, x_{di})}$: تمثل مصفوفة متجهات المشاهدات للمتغيرات التوضيحية.

$m(\mathbf{x}_i)$: تمثل دالة الانحدار غير المعروفة المطلوب تقديرها بالطرائق اللامعلمية.

هناك العديد من الطرائق اللامعلمية لتقدير دالة الانحدار اللامعلمية غير المعروفة والموضحة في المعادلة (1) و(2)، إذ إن هدف التمهيد هو تقريب دالة الانحدار اللامعلمي التقريبية الى دالة الانحدار اللامعلمية الحقيقية، والعمل على تعديل المشاهدات، وهذه الطرائق بنيت على أساس أنموذج مقدر؛ ليعطي أنموذجاً مقارباً للواقع وللتنبؤ بالمستقبل، وإن معظم الطرائق اللامعلمية تفترض أن الخطأ يتوزع بمتوسط مساوٍ للصفر وتباين محدد وإن الدالة هي دالة مستمرة (Continuous Function) وممهدة (Smoothed)، ومن الطرائق اللامعلميةِ الإحصائية التي لاتتطلب توفر فروض بشأن توزيع المجتمع، التي تعد أداة فعالة تعتمد بشكل أساسي على البيانات (Data) بدلاً من الصيغ الدالية المسبقة للبيانات هي طريقة التمهيد بالتقدير (اللُبي) Smoothing Kernel Estimation Method (Hardle, 1994).

تعد طريقة التمهيد بالتقدير اللُبي من الطرائق الأكثر شيوعاً في تقدير نماذج الانحدار اللامعلمي، التي تمثل القاعدة الأولى للتقديرات اللُبية، والفكرة الأساسية التي تقوم عليها طريقة تقدير دالة الانحدار اللامعلمية تكون من خلال أخذ الوسط الحسابي الموزون Weighted Average لقيم متغير الاستجابة من المعادلة (6.2)، إذ إن المقدر اللُبي المستخدم يكون مقدراً تمهيدياً (Smooth) خالٍ من الإضطرابات، ويعتمد على دالة وزن Weighted Function أو ما تسمى بالدالة اللُبية Function Kernel التي تسيطر على شكل المنحنى، فضلاً عن معلمة تمهيدية (عرض الحزمة) التي تسيطر على حجم الأوزان المستخدمة لغرض التقدير، وبالاعتماد على الدالة اللُبية وعلى المعلمة التمهيدية (عرض الحزمة) يعمل هذا المقدر على تعديل أو تدريج المشاهدات وتقريب دالة الانحدار اللامعلمي التقديرية إلى دالة الانحدار اللامعلمي الحقيقية (عنبر،2010).

إن الشروط والصفات الواجب تحققها للاختيار الأمثل لهذه الدالة هي أن تكون الدالة مستمرة وموجبة ومتماثلة حول الصفر، وقابلة للتفاضل إذ أن المشتقة الثانية موجودة (حمزة، 2008). وبالرغم من أن شكل الدالة اللُبية (Ku) مهم للحصول على مقدرات تمتلك صفات تتقارب مع الخواص الإحصائية الاستدلالية إلا أن الباحث (Hardle, 1994) أكد أن اختيار المعلمة التمهيدية (عرض الحزمة) في طريقة التقديرات اللُبية هي أهم من اختيار الدالة اللُبية، ومن المقدرات اللُبية التي تم استعمالها في هذا البحث مقدر Local Polynomial Kernel Smoother.

إن للدوال اللُبية عدة تسميات منها (دالة الوزن Weight Function ، دالة شكل Shape Function ، دالة نافذةWindow Function أو دالة أساسية Basic Function) وهناك سلسلتان من الدوال اللُبية يمكن تمييزهما لتقدير دوال الانحدار اللامعلمي هما: الدوال اللُبية ذات أقل تباين Minimum Variance التي تعمل على تقليل التباين ، والدوال اللُبية المثلى Optimal Kernel التي تعمل على تقليل مجموع مربعات الأخطاء التكاملي Mean Integrated Squared Error (MISE) أي اشتقاق MISE بالنسبة للدالة اللُبية وهذه الدوال تكون دالة حقيقية متماثلة محددة ومستمرة وتكاملها مساوٍ للواحد (متي، 2012).

II. ممهد المقدر اللُبي الموضعي متعدد الحدود Local Polynomial Kernel Smoother (LPKS)

ويسمى أيضا بممهد الانحدار الخطي (الموضعي) وتعتمد هذه الطريقة على فكرة التقريب للدالة m(x) باستخدام متعدد

حدود من درجة معينة، وأساس هذه الفكرة هو توسيع تايلر Taylor Expansion، الذي ينص بدوره على أن أية دالة تمهيدية يمكن أن تتقارب موضعياً من متعدد الحدود من درجة معينة، فإذا كانت تمثل نقطة افتراضية عشوائية بحيث نقوم بتقدير m(x) دالة الموضحة في الصيغة (1) و(2) وعلى افتراض أن يكون لدينا (p+1) مشتقة مستمرة للدالة لبعض قيم p≥1 عند النقطة وباستخدام توسيع تايلر يمكننا تقريب الدالة m(x) موضعياً باستخدام متعدد الحدود من الدرجة p، كما موضح بالصيغة الآتية :(Schucany, 1995)

$$m(x) \approx (x-x_0)^0 m(x_0)/0! + (x-x_0)^1 m^{(1)}(x_0)/1! + \ldots + (x-x_0)^p m^p(x_0)/p! \quad (3)$$

حيث إن $m^r(x_0)$ تمثل المشتقة r للدالة m(x) عند النقطة x_0.

لنفترض هنا أن تكون:

$$\beta_r = m^{(r)}(x_0)/r!$$

حيث إن:

r = 0,1,…,p

وأن $\hat{\beta}_r$ تمثل القيم التي تقلل معيار المربعات الصغرى الموزونة Weighted Least Squared (WLS) كالآتي:

$$\hat{\beta}_r = \min \sum_{i=1}^{n} \left\{ \mathbf{y}_i - \hat{\beta}_0 - \hat{\beta}_1(\mathbf{x}_i - x_0) - \cdots - \hat{\beta}_p(\mathbf{x}_i - x_0)^p \right\}^2 K_h(\mathbf{x}_i - x_0) \quad (4)$$

حيث إن $K_h(\square) = K(\square/h)/h$، التي يتم الحصول عليها بإعادة احتساب الدالة اللُبية $K(\square)$ مع الثابت h > 0، الذي يسمى بدوره بعرض الحزمة، أو معلمة التمهيد التي تستخدم بشكل أساسي في تحديد حجم الجوار الموضعي Local Neighborhood والموضح بالصيغة الآتية (Borrajo, et. al, 2017) :

$$I_h(x_0) = [x_0 - h, x_0 + h] \quad (5)$$

حيث تقوم الدالة اللُبية $K(\square)$ بتحديد كيف تساهم المشاهدات وضمن الفترة $I_h(x_0)$ في الملائمة عند النقطة x_0 ويمكن الإشارة إلى مقدر المشتقة r، $m^{(r)}(x_0)$، على أنه $\hat{m}_h^{(r)}(x_0)$ كما موضح بالصيغة الآتية:

$$\hat{m}_h^{(r)}(x_0) = r!\,\hat{\beta}_r \quad , \; r = 0,1,\ldots,p \quad (6)$$

وبشكل خاص يتم الحصول على مقدر LPKS من الدرجة p للدالة $m(x_0)$ وهو $\hat{m}_h(x_0) = \beta_0$ ويمكن تمثيله بشكل مصفوفة كما موضح في أدناه:

$$\mathbf{X} = \begin{bmatrix} 1 & (x_1 - x_0) & \ldots & (x_1 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - x_0) & \ldots & (x_n - x_0)^p \end{bmatrix}_{n \times (p+1)} \quad (7)$$

وان :

$$\mathbf{W} = \begin{bmatrix} \mathbf{K}_h(x_1 - x_0) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \mathbf{K}_h(x_n - x_0) \end{bmatrix}_{n \times n} \quad (8)$$

تمثل مصفوفة التصميم ومصفوفة الوزن لملائمة المقدر اللُبي الموضعي حول x_0 وهي مصفوفة قطرية وبالتالي يمكن إعادة كتابة معيار WLS كما موضح في الصيغة الآتية:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (9)$$

إذ إن $\mathbf{y} = (y_1, \ldots, y_n)^T$، وأن $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_n)^T$، وهذا يؤدي إلى:

$$\hat{m}_h^{(r)}(x_0) = r!\,\mathbf{e}_{\mathbf{r+1}}^{\mathbf{T}}\, \mathbf{S}_{\mathbf{n}}^{-\mathbf{1}}\, \mathbf{T}_{\mathbf{n}}\mathbf{y} \quad (10)$$

وتشير $\mathbf{e}_{\mathbf{r+1}}$ إلى متجه من الدرجة (p+1) التي تكون عناصره في المجال (r+1) مساوية للواحد، أما باقي عناصره الأخرى فتكون أصفاراً، وأن:

$$\mathbf{S_n} = \mathbf{X^T}\,\mathbf{W}\,\mathbf{X}, \quad \mathbf{T_n} = \mathbf{X^T}\,\mathbf{W} \quad (11)$$

وسيتم التركيز هنا على ممهد المنحنى، كما موضح في أدناه:

$$\hat{m}_h(x_0) = \mathbf{e_1^T}\,\mathbf{S_n^{-1}}\,\mathbf{T_n}\,\mathbf{y} \quad (12)$$

وفي حالة عدم مناقشة تقدير المشتقة، عليه نفترض أن $\hat{y}_i = \hat{m}_h(x_i)$ تمثل القيمة الملائمة Fitted Value للدالة $m(x_i)$ باستخدام الصيغة (13) يتضح أن:

$$\hat{m}_h(\mathbf{x}_i) = a(\mathbf{x}_i)^T \mathbf{y} \quad (13)$$

حيث إن $a(x_i)$ تمثل $\mathbf{e_1^T S_n^{-1} T_n}$ بعد أن يتم استبدال x_0 محل x_i، وبافتراض أن $\hat{\mathbf{y}}_h = [\hat{y}_1,\ldots,\hat{y}_n]$ تمثل القيم الملائمة عند جميع نقاط وقت التصميم Design Time Points عليه يمكننا توضيح قيمة $\hat{y}_h$ كما في الصيغة الآتية:

$$\hat{\mathbf{y}}_h = \mathbf{A}_h \mathbf{y} \quad (14)$$

حيث إن:

$$\mathbf{A}_h = (a(x_1),\ldots,a(x_n))^T \quad (15)$$

وهي ما تعرف بمصفوفة التمهيد لممهد النواة المتعدد الحدود الموضعي LPKS، حيث إن $\mathbf{A}_h$ لا تعتمد على متجه الاستجابة **y**، فإن ممهد النواة المتعدد الحدود الموضعي $\hat{m}_h$ يعرف على أنه ممهد خطي (Kauermann & Opsomer, 2004).

وبتعميم ذلك في حالة وجود أكثر من متغير توضيحي أي d من المتغيرات التوضيحية تكون الصيغ كالآتي (Rady & Ziedan, 2014)،(Schafer & Wasserman, 2013):

$$\mathbf{K_H(X)} = \frac{\mathbf{1}}{|\mathbf{H}|}\mathbf{K(H^{-1}X)} \quad (16)$$

حيث:

K: تمثل الدالة اللُبية.

H: تمثل مصفوفة عرض الحزمة وتكون متماثلة وموجبة وذات بعد d×d.

وإن مجموع المربعات الصغرى الموضعية تكون كالآتي:

$$\sum_{i=1}^{n} \mathbf{W}_i(\mathbf{X})(\mathbf{y}_i - a_0 - \sum_{j=1}^{d} a_j(\mathbf{X}_{ij} - \mathbf{X}_j))^2 \quad (17)$$

حيث

$$\mathbf{W}_i(\mathbf{X}) = \mathbf{K}(\|\mathbf{X}_i - x\|/h) \quad (18)$$

وبفرض أن a_0 و a_j في المعادلة (17) يمثلان حلا لمسألة باستخدام المربعات الصغرى الموزونة في حساب تقدير $\hat{a}$ كما في معادلة (20) وهي قيمة ل a.

ومن المربعات الصغرى الموزونة نحصل على:

$$\hat{a} = (\mathbf{X^T W X})^{-1}\mathbf{X^T W Y} \quad (19)$$

حيث:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11}-x_1 & \ldots & x_{1d}-x_d \\ 1 & x_{21}-x_1 & \ldots & x_{2d}-x_d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}-x_1 & \ldots & x_{nd}-x_d \end{bmatrix}_{d\times d}$$

وإن:

W : تمثل مصفوفة الأوزان وهي مصفوفة قطرية عناصرها $w_i(x)$ وحسب المعادلة (8).

III. عرض الحزمة Bandwidth

إن عملية اختيار عرض الحزمة Bandwidth (المعلمة التمهيديةParameter Smoothing) تعد الخطوة الأكثر أهمية في تقريب دالة الانحدار اللامعلمي إلى الدالة الأصلية، وللحصول على التقريب الجيد والملائم لا بد من إيجاد الطريقة المُثلى لغرض الموازنة بين كل من التحيز والتباين بحيث يكون مقدار الخطأ أقل مايمكن والذي عادة يقاس بمعيار متوسط مربعات الخطأ Mean Squared Error(MSE) أو متوسط مربعات الخطأ التكاملي Mean Integrated Squared Error(MISE) وهذه الموازنة تأتي من خلال استخدام أفضل معلمة تمهيدية، أو ما تسمى بعرض الحزمة، ومعلمة التباين، وحجم النافذة، ومعلمة الأنتشار أو سعة القيد، إن اختيار هذه القيمة التي تستعمل مع دوال Kernel يجب أن يكون بعناية وحذر، إذ إن القيمة الصغيرة جداً تؤثر على تمهيد المنحنى، وتكون منحنى تمهيد منخفض Under Smoothing أما إذا كانت القيمة كبيرة جداً فهي أيضاً تؤثر على تمهيد المنحنى، ولكن تكون في هذه الحالة منحنى تمهيد مرتفع Over Smoothing (Schimek, 2013).

ويرمز للمعلمة الممهدة بالرمز(h) إذا كانت تستخدم لأحادي المتغير، حيث تتضمن اختيار معلمة مفردة، أما في حالة متعدد المتغيرات فتتضمن اختيار مصفوفة عرض الحزمة ويرمز لها بالرمز(**H**)، وتوجد عدة أساليب لاختيار القيمة المُثلى لعرض

الحزمة (المعلمة التمهيدية) التي تحاول تقليل مجموع مربعات الخطأ للأنموذج وهي طريقة التحقق المتقاطع Cross Validation(CV) وطريقة التحقق المتقاطع العام Generalized Cross Validation(GCV) وطريقة التحقق المتقاطع للمربعات الصغرى Least Squared Cross Validation(LSCV) وغيرها من الطرائق الأخرى.

طريقة التحقق المتقاطع Cross Validation(CV)

تعد طريقة معيار التحقق المتقاطع أو ما تسمى بطريقة معيار التقاطع الشرعي من الطرائق الشائعة الاستخدام؛ لإيجاد أنسب قيمة للمعلمة التمهيدية h، إذ تلعب دور مهماً في تحيز وتباين المقدر. إن الفكرة الأساسية لهذه الطريقة تقوم على أساس تقسيم البيانات الى L من المجاميع الجزئية $(g_1, g_2, \ldots, g_l)$ بحيث كل مجموعة تحتوي على عدد متساو من المشاهدات $n_j = (n_1, n_2, \ldots, n_l)$، حيث يتم استبعاد مجموعة واحدة في كل مرة وتكون المجموعة المستبعدة هي g_j بحيث أن $j = 1, 2, \ldots, l$ ، و كحالة خاصة عندما L=1 تسمى Leave One Out Method، ويتم حساب الخطأ حسب الصيغة الآتية (Hardle, 1990):

$$CV(h) = \frac{1}{n - n_j} \sum_{\substack{i=1 \\ i \notin L_j}}^{n} (\mathbf{y_i} - \hat{m}_{-j}(\mathbf{x}_i))^2 \tag{20}$$

حيث ان :

$\hat{m}_{-j}(\mathbf{x}_i)$: تمثل مقدر Local Polynomial في حالة استبعاد مشاهدة.

ولإيجاد قيمة المعلمة التمهيدية المُثلى نستعمل الصيغة الآتية الدالة (Fan, et.al, 2013).

$$\hat{h} = \arg\min_{h \in H} CV(\mathrm{h}) \tag{21}$$

وبتعميم ذلك في حالة وجود أكثر من متغير توضيحي أي d من المتغيرات التوضيحية نستعمل الصيغة كالآتي (Koláček & Horová, 2017):

$$CV(\mathbf{H}) = \frac{1}{n - n_j} \sum_{\substack{i=1 \\ i \notin L_j}}^{n} (\mathbf{y_i} - \hat{m}_{-j}(\mathbf{X_i}))^2 \tag{22}$$

$\hat{m}_{-j}(\mathbf{X_i})$: تمثل مقدر Local Polynomial في حالة استبعاد مجموعة L_j .

ولإيجاد مصفوفة قيم المعلمات التمهيدية المُثلى $\mathbf{H}$ نستعمل الصيغة الآتية:

$$\hat{\mathbf{H}}_{cv} = \arg\min_{\mathbf{H} \in \mathbf{H}} CV(\mathbf{H}) \tag{23}$$

طريقة التحقق المتقاطع العام Generalized Cross Validation (GCV)

تعد طريقة التحقق المتقاطع العام GCV طريقة مقاربة لطريقة معيار التحقق المتقاطع CV، ويمكننا الحصول عليها من صيغة التحقق المتقاطع CV الموضحة في الصيغة (22) مع إجراء تغيير وهو استبدال عناصر القطر الرئيسي m_{tt} لمصفوفة التمهيد $\mathbf{S}_h$ بمعدلها أي إن:

$$\frac{1}{n}\sum_{i=1}^{n} m_{tt} = \frac{1}{n} tr(\mathbf{S}_h) \tag{24}$$

حيث إن:

$tr(\mathbf{S}_h)$: يمثل حاصل جمع عناصر القطر الرئيسي للمصفوفة.

وعليه فإن صيغة معيار التحقق المتقاطع العام GCV تكون كالآتي (Wu & Zhang, 2006):

$$GCV(\mathrm{h}) = \frac{1}{n - n_j} \sum_{\substack{i=1 \\ i \notin L_j}}^{n} \left\{ \frac{\mathbf{y_i} - \hat{m}_{-j}(\mathbf{x_i})}{1 - tr(\mathbf{S}_h)/n} \right\}^2 \tag{25}$$

$$\hat{h}_{GCV} = \arg\min_{h \in H} \mathrm{GCV}_h \tag{26}$$

وبتعميم ذلك في حالة وجود أكثر من متغير توضيحي أي d من المتغيرات التوضيحية تكون الصيغة كالآتي:

$$GCV(\mathbf{H}) = \frac{1}{n - n_j} \sum_{\substack{i=1 \\ i \notin L_j}}^{n} \left\{ \frac{\mathbf{y_i} - \hat{m}_{-j}(\mathbf{X}_i)}{1 - tr(\mathbf{S_H})/n} \right\}^2 \tag{27}$$

ولإيجاد مصفوفة قيم المعلمات التمهيدية المُثلى $\mathbf{H}$ نستعمل الصيغة الآتية:

$$\hat{\mathbf{H}}(\mathrm{GCV}) = \arg\min_{\mathbf{H}\in\mathbf{H}} \mathrm{GCV}(\mathbf{H}) \tag{28}$$

IV. خوارزمية الاعشاب الضارة Invasive Weed Optimization Algorithm (IWO)

في السنوات الأخيرة أصبح الاهتمام متزايدا بتصميم خوارزميات التحسين المستوحاة من الطبيعة وتطويرها، وتعد خوارزمية أمثلة الأعشاب الضارة Invasive Weed Optimization Algorithm (IWO) او ما تسمى بخوارزمية التحسين العشوائي العددي المستوحاة بيولوجيا من الأعشاب الضارة والتي اقترحت لاول مرة من قبل Mehrabian و Lucas في عام (2006 م) . ببساطة تحاكي هذه الخوارزمية السلوك الطبيعي للأعشاب الضارة في الاستعمار وايجاد مكان مناسب للنمو والتكاثر، ولمحاكاة السلوك الاستعماري للأعشاب الضارة يجب ان تؤخذ بنظر الاعتبار الخطوات الاساسية لهذه العملية:

(1) يتم (تهيئة عدد السكان) وذلك نشر عدد محدود من البذور على منطقة البحث.

(2) تتم (عملية التكاثر) بحيث كل البذور تنمو الى نباتات مزهرة وتنتج البذور اعتمادا على دالة اللياقة المعادلة ادناه توضح عملية التكاثر للأعشاب الضارة:

$$seed_i = floor\left(\frac{f_i - f_{\min}}{f_{\max} - f_{\min}}\left(S_{\max} - S_{\min}\right)\right) + S_{\min} \tag{29}$$

(3) (التشتت المكاني) توفر هذه الخطوة لخوارزمية الأعشاب الضارة خاصيتي العشوائية والتكيف, اذ يتم توزيع البذور المتولدة عشوائيا على d من الابعاد في فضاء البحث بواسطة ارقام عشوائية تتوزع توزيعا طبيعيا بمعدل (μ=0) وتباين متغير من خلال المعادلة التالية:

$$\sigma_{iter} = \frac{\left(iter_{\max} - iter\right)^n}{\left(iter_{\max}\right)^n}\left(\sigma_{initial} - \sigma_{final}\right) + \sigma_{final} \tag{30}$$

يتم حساب موقع البذور الجديدة باستخدام المعادلة التالية:

$$x_{son} = x_{parent} + sd = x_{parent} + random * \sigma_{iter} \tag{31}$$

(4) تستمر هذه العملية الى ان يتم الوصول الى الحد الأقصى من عدد النباتات.

والنباتات ذات دالة اللياقة العالية فقط يمكنها البقاء على قيد الحياة وإنتاج البذور, ويجري القضاء على الاخرين (الإقصاء التنافسي). تستمر العملية الى ان يتم الوصول الى الحد الأقصى من التكرارات على امل ان النبات الذي يحمل أفضل دالة لياقة سيكون هو الاقرب الى الحل الامثل، الشكل (1) يوضح الية عمل الخوارزمية في اختيار المتغيرات.

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

شكل (1): الية اختيار المتغيرات حسب خوارزمية الاعشاب الضارة

نتائج المحاكاة

تم تصميم العديد من التجارب ومحاكاتها باستعمال لغة البرمجة (R)، إذ تم الأخذ بنظر الاعتبار ثلاثة أحجام للعينات n=(50,100,250)، وتم إجراء المقارنات للطرق المختلفة المستخدمة CV، GCV، مع الطريقة المقترحة لخوارزمية (IWO) وباستخدام دالة Epanechnikov كدالة أُبية، مع دراسة خمسة نماذج مختلفة هي:

(1) <u>الأنموذج الأول:(Koláček & Horová, 2017)</u>: تم توليد هذا الأنموذج وفق المعادلة الآتية:

$$\mathbf{y_i} = (\mathbf{x_1} - 0.5)^3 + (\mathbf{x_2} - 0.5) + \boldsymbol{\varepsilon_i} \tag{32}$$

حيث تم توليد المتغيرات التوضيحية $\mathbf{x_1}$ و $\mathbf{x_2}$ من التوزيع المنتظم ضمن الفترة [0,1]، أما $\boldsymbol{\varepsilon_i}$ فيتوزع بالشكل الآتي:

$\boldsymbol{\varepsilon_i} \sim N(0, 0.02)$

(2) <u>الأنموذج الثاني (Shang,et.al,2014)</u>: تم توليد هذا الأنموذج وفق المعادلة الآتية:

$$\mathbf{y_i} = \sin\left(2\pi\,\mathbf{x_1}\right) + 4\left(1 - \mathbf{x_2}\right)\left(1 + \mathbf{x_2}\right) + \frac{2\mathbf{x_3}}{1 + 0.8\,\mathbf{x_3^2}} + \boldsymbol{\varepsilon_i} \tag{33}$$

حيث تم توليد المتغيرات التوضيحية $\mathbf{x_1}$ و $\mathbf{x_2}$ و $\mathbf{x_3}$ من التوزيع المنتظم ضمن الفترة [0,1]، أما $\boldsymbol{\varepsilon_i}$ فيتوزع بالشكل الآتي:

$\varepsilon_i \square N(0, 0.02)$

(3) الأنموذج الثالث (Lijian & Rolf, 1999): تم توليد هذا الأنموذج وفق المعادلة الآتية:

$$y_i = \left\{(x_1 - 0.5)^2 + x_2^2\right\}\sin(2\pi x_3) + \varepsilon_i \quad (34)$$

حيث تم توليد المتغيرات التوضيحية x_1 و x_2 و x_3 من التوزيع المنتظم ضمن الفترة [0,1]، أما ε_i فيتوزع بالشكل الآتي:

$\varepsilon_i \square N(0, 0.25)$

(4) الأنموذج الرابع(Lijian & Rolf, 1999) : تم توليد هذا الأنموذج وفق المعادلة الآتية:

$$y_i = \sin\left\{\pi(x_1 + 0.5x_2 + 2x_3 + 0.5x_4)\right\} + \varepsilon_i \quad (35)$$

حيث تم توليد المتغيرات التوضيحية x_1 و x_2 و x_3 و x_4 من التوزيع المنتظم ضمن الفترة [0,1]، أما ε_i فيتوزع بالشكل الآتي:

$\varepsilon_i \square N(0, 0.25)$

(5) الأنموذج الخامس(Goutte,et.al,2000): تم توليد هذا الأنموذج وفق المعادلة الآتية:

$$y_i = 10\sin(\pi x_1 x_2) + 20\left(x_3 - \frac{1}{2}\right)^2 + 10x_4 + 5x_5 + \varepsilon_i \quad (36)$$

حيث تم توليد المتغيرات التوضيحية x_1 و x_2 و x_3 و x_4 و x_5 من التوزيع المنتظم ضمن الفترة [0,1]، أما ε_i فيتوزع بالشكل الآتي:

$\varepsilon_i \square N(0,1)$

إذ تم تكرار اجراء التجربة 500 مرة، وتم الاعتماد على متوسط مربعات الخطأ المتكامل (IMSE) كمعيار للمقارنة بين طرائق التقدير المستخدمة وبيان الطريقة الأفضل وتلخيص نتائج الطرق المستخدمة في الجداول من (1-5).

اظهرت النتائج تفوق طريقة (IWO) على باقي الطرائق الاخرى (CV,GCV) من حيث معايير IMSE. فعلى سبيل المثال جدول (1) اعطت طريقة IWO اقل قيمة عندما n=50، اذ بلغت 0.0130 مقارنة بـــ 0.037 لطريقة CV و 0.036 لطريقة GCV. بالاضافة الى ذلك حافظت الطريقة المقترحة على افضليتها عند تغير حجم العينة ولجميع النماذج المستخدمة.

جدول (1): معدل قيمة MSE وقيمة الانحراف المعياري للأنموذج الأول في مقدر Local Polynomial.

	CV	GCV	IWO
n=50	0.037 ± 0.009	0.036 ± 0.007	0.0130 ± 0.0021
n=100	0.037 ± 0.007	0.034 ± 0.006	0.0013 ± 0.0013
n=250	0.034 ± 0.004	0.003 ± 0.030	0.0012 ± 0.0010

جدول (2): معدل قيمة MSE وقيمة الانحراف المعياري للأنموذج الثاني في مقدر Local Polynomial.

	CV	GCV	IWO
n=50	3.126 ± 0.644	2.978 ± 0.901	0.0131 ± 0.0017
n=100	3.105 ± 0.634	2.814 ± 0.786	0.0074 ± 0.0002
n=250	2.689 ± 0.300	2.084 ± 0.246	0.0045 ± 0.0001

جدول (3): معدل قيمة MSE وقيمة الانحراف المعياري للأنموذج الثالث في مقدر Local Polynomial.

	CV	GCV	IWO
n=50	0.226 ± 0.054	0.218 ± 0.065	0.041 ± 0.030
n=100	0.214 ± 0.044	0.192 ± 0.052	0.029 ± 0.020
n=250	0.208 ± 0.017	0.172 ± 0.019	0.015 ± 0.014

جدول (4): معدل قيمة MSE وقيمة الانحراف المعياري للأنموذج الرابع في مقدر Local Polynomial.

	CV	GCV	IWO
n=50	0.514 ± 0.144	0.153 ± 0.079	0.0039 ± 0.0018
n=100	0.331 ± 0.133	0.135 ± 0.047	0.0036 ± 0.0015
n=250	0.268 ± 0.051	0.119 ± 0.036	0.0025 ± 0.0012

جدول (5): معدل قيمة MSE وقيمة الانحراف المعياري للأنموذج الخامس في مقدر Local Polynomial.

	CV	GCV	IWO
n=50	5.422 ± 1.593	2.520 ± 0.546	0.063 ± 0.030
n=100	4.567 ± 1.356	1.969 ± 0.818	0.038 ± 0.002
n=250	3.801 ± 0.875	1.449 ± 1.406	0.008 ± 0.001

الاستنتاجات

1. من خلال مقارنة طرائق التقدير اللامعلمية المعتمدة والطريقة المقترحة **IWO** لتقدير مصفوفة معلمات التمهيد في أنموذج الانحدار اللامعلمي، وجد أن أفضل طريقة كانت الطريقة المقترحة **IWO** ، أذ حققت أقل قيمة لمعيار متوسط مربعات الخطأ (MSE) ولجميع حجوم العينات الثلاث، ولجميع النماذج المعتمدة في تجارب المحاكاة.
2. جاءت طريقة GCV أفضل طريقة لتقدير مصفوفة معلمات التمهيد بعد الطريقة المقترحة IWO لإعطائها نتائج جيدة في تقدير مصفوفة معلمات التمهيد، وكانت طريقة CV الاخيرة لأنها تعطي دائماً أعلى MSE مقارنة بالطرائق الأخرى.

المصادر

[1] حمزة، سعد كاظم، (2008)، "مقارنة بعض طرائق تقدير نماذج الانحدار اللامعلمي بوجود بيانات غير تامة"، رسالة ماجستير في الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد.

[2] عنبر، جنان عبدالله،(2010)، "بعض طرائق التقديرات اللامعلمية لأنموذج الانحدار التجميعي المجزأ مع التطبيق"، رسالة ماجستير في الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد.

[3] متي، نور صباح، (2012)، "مقارنات لبعض طرائق تقدير دالة الانحدار اللامعلمي"، رسالة ماجستير في الإحصاء، كلية علوم الحاسوب والرياضيات، جامعة الموصل.

[4] محمد، لقاء علي وعبد الحسن، ميسم عبد النبي، (2018)، "مقارنة المقدرات اللامعلمية في تحليل الانحدار المتعدد لدالتي كاما وبيتا" ، مجلة العلوم الاقتصادية والإدارية، المجلد (24)، العدد (108)، الصفحات[488-497]

[5] محمد، محمد عبد الحسين، (2011)، "استخدام مقدر كيرنل Nadaraya-Watson في تقدير دالة الانحدار اللامعلمي"، مجلة القادسية للعلوم الإدارية والاقتصادية، المجلد (13)، العدد (1)، المعهد التقني الديوانية. 15.

[6] Borrajo, M. I., González-Manteiga, W., & Martínez-Miranda, M. D., (2017), "Bandwidth selection for kernel density estimation with length-biased data", Journal of Nonparametric Statistics, 29(3), PP[636-668].

[7] Fan, G. L., Liang, H. Y., & Wang, J.-F., (2013), "Statistical inference for partially time-varying coefficient errors-in-variables models", Journal of Statistical Planning and Inference, 143(3), pp[505-519].

[8] Goutte, C., Larsen, J., & technology, v., (2000), "Adaptive metric kernel regression", Journal of VLSI signal processing systems for signal, image, 26(1-2), pp[155-167].

[9] Hardle, W., (1990), Applied Nonparametric Regression, Cambridge MA : Cambridge University press.

[10] Harldle , W. , (1994)," Applied non parametric regression", Cambridge University Press

[11] Kauermann, G., & Opsomer, J., (2004), "Generalized cross-validation for bandwidth selection of backfitting estimates in generalized additive models", Journal of Computational and Graphical Statistics, 13(1),PP[66-89].

[12] Koláček, J., & Horová, I., (2017), "Bandwidth matrix selectors for kernel regression", Computational Statistics, 32(3), [1027-1046].

[13] Lijian, Y., & Rolf, T., (1999), "Multiveraite bandwidth selection for local linear regression", Statist, 61(4), pp[793-815].

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

[15] Rady, E. H. A., & Ziedan, D., (2014), "Estimation of population total using local polynomial regression with two auxiliary variables", Journal of Statistics Applications & Probability, 3(2), 129.

[16] Rencher, A. C., (2002), "Methods of Multivariate Analysis", Second Edition, John Wiley & Sons, Inc., Canada.

[17] Schafer, C., & Wasserman, L., (2013), Tutorial on Nonparametric Inference, University Carnegie Mellon, pp[1-202].

[18] Schimek, M. G., (2013), Smoothing and regression: approaches, computation, and application, John Wiley & Sons.

[19] Schucany, W. R., (1995), "Adaptive bandwidth choice for kernel regression", Journal of the American Statistical Association, 90(430),PP[535-540].

[20] Shang, H. L., Zhang, X., & Shang, M. H. L., (2014), "Package bbemkr".

[21] Wu, H., & Zhang, J.-T., (2006), "Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches" ,Vol. 515: John Wiley & Sons.

Journal of AL-Rafidain University College for Sciences (2023); Issue 54; 58- 67

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

Employ Nature-Inspired Algorithms in Selecting the Band Width in A Polynomial Estimator

Assist. Lect. Marwa M. Yahya	Prof. Dr. Zakariya Y. Al-Gamal
marwa.yahya@uomosul.edu.iq	zakariya.algamal@uomosul.edu.iq
Department of Statistics and Informatics - College of Computer Science and Mathematics - University of Mosul, Nineveh, Iraq	

Article Information

Article History:
Received: December, 10, 2022
Accepted: February, 26, 2023
Available Online: December, 31, 2023

Keywords:
Kernel estimator; smoothing matrix, Local Polynomial Kernel estimator, Invasive Weed Optimization Algorithm

Correspondence:
Assist. Lect. Marwa M. Yahya
marwa.yahya@uomosul.edu.iq

Abstract

The subject of regression analysis receives broad and explicit attention in the majority of studies, particularly those in the fields of economics and medicine. The nonparametric regression model in general and the multivariate nonparametric regression model in particular is one of the most significant regression models used in recent years which has seen significant expansion, particularly in the fields of economics and the environment. Local Polynomial Kernel estimator is one of the most used estimators in multivariate nonparametric regression model. However, this estimator is fully dependent on the matrix of the smoothing parameter, which is of crucial in the quality of the reconciliation of the estimated shape in the multivariate nonparametric regression model. In this research, employing the nature-inspired technique as Invasive Weed Optimization technique was proposed for the process of estimating the prefix parameter matrix in the estimator. The Mont-Carlo method of simulation has been used to generate data for tracking a number of multivariate regression models. Simulation results based on mean squared error have been shown by developing a benchmark that surpasses the recommended approach in comparison to alternative estimating methods.

https://doi.org/10.55562/jrucs.v54i1.576