AL KUT JOURNAL OF ECONOMIC AND ADMINISTRATIVE SCIENCES
Publisher: College of Economics and Management - Wasit University

مقارنة طريقة Lasso مع طريقة Multi-split Lasso لتقدير انموذج الانحدار متعدد المتغيّرات في ظل بيانات عالية الأبعاد باستخدام المحاكاة

https://doi.org/10.29124/kjeas.1651.04

لمى قيس عواد رضا[1] أ.م. د طارق عزيز صالح[2]

كُلّية الإدارة والاقتصاد/ جامعة واسط

المستخلص

تهدف هذه الدراسة إلى مقارنة بعض طرائق التنظيم المستعملة في تقدير نموذج الانحدار متعدّد المتغيّرات في ظل بيانات عالية الأبعاد. تَم إجراء تحليل شامل للبيانات ودراسة العلاقة بين المتغيّرات المستقلّة والمتغيّر المعتمد. تَمّ تطبيق هذه الطرائق على مجموعة من البيانات الفعلية لتقييم أداؤها وفعاليتها. تَمّتْ المقارنة بين طريقة لاسو (Lasso) التي تستعمل في بيانات عالية الأبعاد، لإيجاد المقدّرات في حالة وجود مشكلة الأبعاد ومقارنتها مع طريقة (Multi-split Lasso) وتَمّ تحليل نتائج التقدير وتقييم الدقة والأداء العام لكلّ طريقة. توصلت الدراسة إلى نتائج مهمة تشير إلى أنّ طريقة (Multi-split Lassoهي الأفضل مقارنة بالطريقة (Lasso) وذلك باستعمال بعض المعايير وأهّمها معيار BIC الذي بنتائجه اعطى الأفضلية للطريقة Multi-split Lasso لأنموذج الانحدار في بيانات عالية الأبعاد.

المصطلحات/ Multivariate Regression، Lasso و Multi-split Lasso، Good fit model.

Abstract

This study aims to compare some of the organizing methods used in estimating a multivariable regression model in the presence of high-dimensional data. A comprehensive analysis of the data was conducted, studying the relationship between the independent variables and the dependent variable. These methods were applied to a set of real data to evaluate their performance and effectiveness. A comparison was made between the Lasso method, which is used for high-dimensional data, to find estimators in the case of dimensionality problem, and the Multi-split Lasso method. The estimation results were analyzed, and the accuracy and overall performance of each method were evaluated. The study reached important results indicating that the Multi-split Lasso method is superior to the Lasso method, based on several criteria, most notably the BIC criterion, which favored the proposed method for the regression model in high-dimensional data.

Terms/Multivariate Regression, Lasso, Multi-split Lasso, Good fit mode

1- المقدمة Introduction

زاد الاهتمام في السنوات الأخيرة بتحليل البيانات ذات الأبعاد العالية لاسيّما عندما يكون عدد المتغيّرات المستقلّة أكبر من حجم العينة، ويعد تحليل الانحدار أحد الطرائق الإحصائية الذي تستعمل في تحليل ودراسة علاقة اثر متغير واحد او اكثر من المتغيّرات المستقلّة (Independent variables) X على متغير معتمد (response variable) Y ، كما تتمتع نماذج الانحدار الخطية التي تشتمل على عدد كبير من المتغيّرات المستقلّة بضعف الأداء ، نتيجة تضخم التباين فيصعب تفسيرها، عند استعمال بيانات عالية الأبعاد فإنّ تقديرات الانحدار القياسية فيما ذلك تلك المرتبطة بالنماذج متعدد المتغيّرات تؤدي إلى تقديرات غير مستقرة مع أخطاء معيارية متضخمة ومِنْ ثَمّ يؤدي إلى انخفاض القوة الإحصائية والاستنتاجات الخاطئة التي توضح العلاقة بين المتغيّرات المستقلّة والمتغيّر المعتمد ونتيجة لذلك سوف يعطينا نتائج غير دقيقة مع البيانات عالية الأبعاد.

توجد هنالك العديد من الطرائق الإحصائية التي تستعمل في تقدير معلمات نموذج الانحدار الخطي أهمّها طريقة المربعات الصغرى الاعتيادية ويرمز لها بالرمز (OLS) والتي تتميز تقديراتها بأفضل التقديرات الخطية غير المتحيزة (BLUE). ولكن يأخذ على هذه الطريقة عدم قدرتها على التعامل مع البيانات عالية الأبعاد، لأنّ مصفوفة المعلومات لا تتمتع برتبة كاملة الأمر الذي يؤدي إلى حصول نتائج غير جيدة.

لذا تَمّ استعمال طريقة (Lasso) للتعامل مع بيانات عالية الأبعاد، والتي يكون مبدؤها تصغير مجموع مربعات الخطأ وفق لقيد معين من المعلمات. ومن اهم المزايا التي تتمتع بها طريقة لاسو هي الحصول على تنبّؤ عالـي الدقّة وتمتعها بعملية التقدير واختيار المتغيّرات في ان واحد إذْ إنّها تقوم بعملية تقليـص بعض المعلمـات وترغـم الأخرى للصفر. و تَمّ في هذا البحث مقارنة طريقة Multi-split مع طريقة (Lasso) للحصول على أفضل تقدير للمعلمات.

2- مشكلة البحث:

هناك مشكلة في السيطرة على الخطأ من النوع الأول في طرائق تقدير انحدار متعدد المتغيّرات والتي شملت طرائق، وذلك بالاعتماد على الجانب التجريبي الذي أظهر إن طريقة Lasso غير مستقرة مع هذا النوع من النماذج، بناء على هذا الاعتقاد يمكن حصر المشكلة في هذه البحث بالسؤال: هل إن الطرائق الجزائية Lasso لها سيطرة على ظاهرة أل Overfit أم لا؟

وهل إن فكرة التجزئة الثنائية المتكررة Multi-split لمجموعة البيانات تسمح باستقرار مسار تقدير المعلمات لطريقة Lasso؟

3- هدف البحث:

يهدف هذا البحث إلى تقدير نموذج الانحدار الخطّي متعدد المتغيّرات في ظلّ وجود مشكلة الأبعاد العالية High Dimansionalوللحالات p>n , p=n , p<n باستعمال طريقة التقدير(Lasso) فضلا عن طريقة Multi-split Lasso وللمقارنة بين الطريقتين باستعمال اسلوب المحاكاة ولتجارب عدّة.

4- تحليل الانحدار **Regression analysis**

تحليل الانحدار هو أسلوب إحصائي يعني بدراسة العلاقة بين المتغيّرات من خلال تقدير معالم النموذج، يتضمن ذلك تقدير معلمات أنموذج الانحدار وإظهار قوة العلاقات واتجاهها وتقييم الأنموذج المقدر، ويعد تحليل الانحدار أحد الأساليب الإحصائية شائعة الاستعمال في نمذجة العلاقة الخطية بين متغير تابع (y) ومتغير واحد أو أكثر من المتغيّرات المستقلّة (x).وبصورة عامة إنّ نموذج الانحدار هو معادلة رياضية يصف العلاقة بين متغيرين أو أكثر، فان العلاقة بين المتغيّرات يمكن إنْ تُقسّم على قسمين هما العلاقة الدالية (Functional Relation) .

إذ يمكن التعبير عن العلاقة الدالية بين المتغيرين بصيغة رياضية محددة، إذ كان (x) يمثل المتغيّر المستقل و(y) يمثل المتغيّر التابع فان العلاقة الدالية تكون وفق الصيغة الأتية [1]:

$$y = f(x)$$

تشير الصيغة المذكورة إنّ التغيّر الحاصل في المتغيّر التابع هو نتيجة التغير الحاصل في المتغيّر المستقل وتكون علاقة بينها علاقة محددة خالية من الأخطاء. يستند الأنموذج الخطي العام إلى افتراض وجود علاقة خطية بين المتغيّر المعتمد (y_i) ، وعدد من المتغيّرات المستقلّة(المتغيّرات المفسرة) $(x_1 , x_2, \dots , x_k)$، وحدا عشوائيا $(\in_m)$، ويمثل لهذه هذه العلاقة بالنسبة لـ (n) من المشاهدات و(k)من المتغيّرات التوضيحية بالشكل الآتي:-

$$y_m = B_{0m} + B_{1m}x_1 + B_{2m}x_2 + \cdots + B_{km}x_k + \in_m \qquad \dots (1)$$

وباستعمال المصفوفات والموجهات يمكن وضع الأنموذج (1) بالشكل الآتي:-

$$Y = XB + \in \qquad \dots (2)$$

إذْ آنّ:-

Y:- موجّه من مرتبة (n×1) لمشاهدات المتغيّر المعتمد.

X:- مصفوفة من مرتبة (n × p) لمشاهدات المتغيّرات التوضيحية.

B:- موجّه من مرتبة (p×1) للمعلمات المطلوب تقديرها.

∈:- موجّه من مرتبة (n×1) للأخطاء العشوائية.

5- نموذج الانحدار الخطّي المتعدّد المتغيّرات

The Multivariate Multiple Liner Regression Model

في نموذج الانحدار متعدّد المتغيّرات، نموذج متعدّد المتغيّرات يصف العلاقة بين عدد من المتغيّرات التوضيحية (التفسيرية) وتأثيرها على متغير معتمد واحد وليس العكس، لذلك لدينا عدد ملاحظات لكل $(y_i \ , \ i = 1 \ldots m)$ يتَمّ إعطاء الصيغة العامة لنموذج الانحدار متعدد المتغيّرات من خلال [2]:

$$y_i = B_{0i} + B_{1i}x_1 + \epsilon_i \qquad \ldots (3)$$

$$\hat{y}_i = \hat{B}_{0i} + \hat{B}_{1i}x_1 \ , \ \ i = 1,2, \ldots., m \qquad \ldots (4)$$

هناك حاجة إلى أربعة مصفوفات للتعبير عن الأنموذج الخطي بشكل مصفوفات وكالآتي:

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \ldots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix}_{n\times m} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}_{n\times 2} \begin{bmatrix} B_{01} & B_{02} & \cdots & B_{0p} \\ B_{11} & B_{12} & \cdots & B_{1p} \end{bmatrix}_{2\times m} + \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \ldots & \epsilon_{1m} \\ \epsilon_{21} & \epsilon_{22} & \ldots & \epsilon_{2m} \\ \vdots & \vdots & \ldots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \ldots & \epsilon_{nm} \end{bmatrix}_{n\times m}$$

Y: مصفوفة (n × m) تعتمد على المتغيّر التابع y.

X: المصفوفة n × 2 التي تتكون من عمود من 1 ، متبوعًا بمتجه العمود للمشاهدات المتغيّر المستقل.

B: مصفوفة m × 2 من المعلمات المراد تقديرها.

∈: مصفوفة n × m للأخطاء العشوائية

في نموذج الانحدار الخطي المتعدد المتغيّرات يتَمّ عدّ العلاقة بين أكثر من متغير تابع وأكثر من متغير مستقل، يبسط نموذج الانحدار إلى حالة (m) من مقاييس الاستجابة $(y_1 , y_2, \ldots , y_m)$ وللمجموعة نفسها إلى (k) من المتغيّرات التوضيحية $(x_1 , x_2, \ldots , x_k)$ وباستعمال حجم عينة (n) لكل من متغير استجابة ومِنْ ثَمّ فإنّ نموذج الانحدار يكتب بالشكل الاتي [3]:

$$y_1 = B_{01} + B_{11}x_1 + B_{21}x_2 + \cdots + B_{k1}x_k + \in_1 \quad \dots (5)$$

$$y_2 = B_{01} + B_{11}x_1 + B_{21}x_2 + \cdots + B_{k1}x_k + \in_2 \quad \dots (6)$$

$$\vdots$$

$$y_m = B_{0m} + B_{1m}x_1 + B_{2m}x_2 + \cdots + B_{km}x_k + \in_m \quad \dots (7)$$

أذ إن:

حد الخطأ $(\in^T = \in_1, \in_2, \dots, \in_m)$ له $(E(\in) = 0)$ ، $(Var\ (\in) = \Sigma)$

B_{ij} : يمثل تقديرات معاملات الانحدار وان $(i = 1,2, \dots, k\ and\ j = 1,2, \dots, m)$ وان (j^{th}) يمثل الاستجابة في تأثير الـ(i^{th}) للتنبؤ.

B_{0j}: يمثل معلمة الحد الثابت لـ (j^{th}) من الاستجابة.

ويمكن تمثّيل الأنموذج بصيغة المصفوفات ونحتاج إلى أربعة أنواع من المصفوفات وكالاتي:

Y: يمثّل مصفوفة من درجة $(n \times m)$ اذ يمثل (m) متجهات عمودية للمشاهدات لكل المتغيّرات المعتمدة

X: يمثّل مصفوفة من درجة $((n \times (k + 1))$ لها عمود من الوحدات و(k) من المتجهات العمودية لمشاهدات المتغيّرات المستقلة.

B: تمثّل مصفوفة من درجة $((k + 1) \times m)$ لها متجهات عمودية إلى المعلمات يتَمّ تقديرها.

$\in$: يمثّل مصفوفة من درجة $(n \times m)$ لها متجهات عمودية للأخطاء العشوائية.

ويمكن كتابة نموذج الانحدار بالشكل الآتي:

$$Y_{(n\times m)} = X_{(n\times(k+1))}B_{((k+1)\times m)} + \in_{(n\times m)} \quad \dots (8)$$

وإنّ $(E(\in_j) = 0)$، $(Cov(\in_j, \in_k) = \sigma_{j.k}\ ,\quad j.k = 1,2, \dots, m)$

وإنّ مصفوفة التباين المشترك $(\Sigma = \sigma_{ik})$ إذْ إنّ:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2m} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nm} \end{bmatrix}$$

وإنّ (σ_{ik}, B) تكون معلمات مجهولة (غير معلومة).

وتكون مصفوفة الاستجابة $(y_{n\times m})$ بالشكل الاتي:

$$y_{n\times m} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix} = (y_1 \quad y_2 \dots y_m) \qquad \dots (9)$$

إذْ أنّ:

y_j: يمثل قيمة عمودية لـ (n) من القياسات إلى (j^{th}) من المتغيّرات.

إذْ أنّ:

$$y_j = [y_{ij}] \qquad for \quad i = 1,2,\dots,n \ , j = 1,2,\dots,m$$

وتكون مصفوفة التصميم (المستقلة) بالشكل الآتي:

$$x_{(n\times(k+1))} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \qquad \dots (10)$$

نلاحظ إنّ الصفوف إلى (x) تتوقف مع المشاهدات للمتغيّرات المستقلّة وتكون مصفوفة المعلمات الأنموذج بالشكل الآتي:

$$B_{((k+1)\times m)} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{km} \end{bmatrix} = [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_m] \qquad \dots (11)$$

إذْ أنّ:

β_{ij}: تمثل $(k+1)$ من معاملات الانحدار في الأنموذج لـ (j^{th}) من المتغيرات.

وان

$$\beta_j = [\beta_{ij}] \quad for\ i = 1,2,\dots,k \quad , j = 1,2,\dots,m$$

وتكون مصفوفة حد الخطأ العشوائي بالشكل الاتي:

$$\epsilon_{n\times m}=\begin{bmatrix}\epsilon_{11} & \epsilon_{12} & \dots & \epsilon_{1m}\\ \epsilon_{21} & \epsilon_{22} & \dots & \epsilon_{2m}\\ \vdots & \vdots & \dots & \vdots\\ \epsilon_{n1} & \epsilon_{n2} & \dots & \epsilon_{nm}\end{bmatrix}=(\epsilon_1 \quad \epsilon_2 \dots \epsilon_m) \qquad \dots(12)$$

إذْ أنّ كل (ϵ_j) تمثل قيمة من الأخطاء العشوائية لكل (m) من متغيرات الاستجابة اذ ان:

$$\epsilon_j=[\epsilon_{ij}]$$

وكذلك (m) من استجابة المشاهدات إلى $((j^{th})$ لمحاولات مصفوفة التباين والتباين المشترك الآتية.

$$\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \dots & \sigma_{1m}\\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2m}\\ \vdots & \vdots & \dots & \vdots\\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nm}\end{bmatrix} \qquad \dots(13)$$

وبالآلية نفسها وبطريقة المربعات الصغرى الخاصة بتقدير نماذج تحليل الانحدار المتعدد والبسيط يتَمّ التقدير لنموذج الانحدار متعدّد المتغيّرات، ففي البداية يتَمّ حساب مجموع مربعات الأخطاء وبعدها يتَمّ إيجاد تقديرات اقل مجموع لمربعات الأخطاء الاتية[3] :

$$SSE=\sum \epsilon_i^2=\epsilon^T\epsilon \qquad \dots(14)$$

ومن خلال حساب المعادلة الطبيعية الآتية:

$$X^TX\hat{B}=X^Ty \qquad \dots(15)$$

نحصل على الحلول في الصيغة الآتية:

$$\hat{B}=(X^TX)^{-1}X^Ty \qquad \dots(16)$$

وباستعمال مقدر المربعات الصغرى لـ (B) يمكن الحصول على القيم التنبؤية الاتية:

$$\hat{y}=X\hat{B} \qquad \dots(17)$$

وعندما $(\hat{B})$ تكون مقدر غير متحيز إلى (B) أي أن:

$$E(\hat{B})=B \qquad \dots(18)$$

$$E(\hat{B})=E((X^TX)^{-1}X^Ty \qquad \dots(19)$$

$$E(\hat{B}) = (X^T X)^{-1} X^T E(XB + \in) \qquad \ldots (20)$$

$$E(\hat{B}) = (X^T X)^{-1} X^T B \qquad \ldots (21)$$

$$E(\hat{B}) = IB = B \qquad \ldots (22)$$

وبذلك يمثل تقدير نموذج الانحدار متعدد المتغيّرات باستعمال طريقة المربعات الصغرى الجزائية

6- مشكلة بيانات عالية الأبعاد

تظهر مشكلة الأبعاد العالية عندما تكون عدد المتغيّرات المستقلّة يفوق على عدد المشاهدات أي $(P > n)$ ويقال عنه بأنموذج الانحدار الخطي ذو الأبعاد العالية (High- dimensional) أما إذا كانت عدد المتغيّرات المستقلّة اقل من عدد المشاهدات أي $(P < n)$ يقال عنه انموذج الانحدار الخطي ذو الأبعاد القليلة (low- dimensional) وفي الأنموذجين نحن نسعى إلى تحقيق عدد من الأهداف منها التقدير، التنبؤ، اختيار المتغيّر [4].

7- دوال الجزاء

اقترح الباحث Tibshirani عام(1996) [5] طريقة تقلص جديدة تسمى "أدنى تقلص مطلق ومشغل الاختيار"، وتُختصر باسم "Lasso". تمت دراسة الخصائص النظرية لطرائق Lasso خلال العقد الماضي وقد، ناقش الباحثان Fan and Li عام(2001) [6] العلاقة بين المربعات الصغرى الجزائية واختيار المجموعات الجزئية، ودرس أيضًا خصائص اختيار المتغيّرات لطرائقLasso. يمكن لـ Lasso أن يقوم باختيار نموذج متسق إذا استوفى شرطًا ضروريًا بشأن مصفوفة التباين المشترك التنبؤية للباحثين Zhao and Yu عام (2006)[7]. تمت استقلالية ظهور هذا الشرط ذاته أيضًا في دراسة الباحث Zou عام(2006) [8].

كما ناقش Fanand Li (2001)[6]، فإن طرائق الانحدار الجزائية مثل Lasso تملك في الواقع خاصتين مثاليتين للأوراكل:

- يتّمّ تقدير المكونات الصفرية (وفقط المركبات الصفرية) على أنها صفر تمامًا بمعدل اقتراب الاحتمال من 1 عندما يتجه الحجم العيني n إلى اللانهائية، إذْ يمثل n حجم العينة.

- يتَمّ تقدير المعلمات غير الصفرية بكفاءة عالية عندما يتَمّ معرفة الأنموذج الجزئي الصحيح.

8- الطرائق الجزائية

ان إجراءات التقدير الإحصائي التقليدية مثل المربعات الصغرى الاعتيادية (OLS) عادة ما يكون أداؤها ضعيف في حالة المشكلات ذات الأبعاد العالية على وجه الخصوص، على الرغم من أن المقدرات OLS عادة ما يكون لها تحيز منخفض، إلا أنها تميل إلى أن تكون لها تباين عالٍ، وقد يكون من الصعب تفسيرها كما بين الباحث Brownعام (1993) [9] . في مثل هذه

الحالات، من المفيد غالبًا استعمال التقليص أي انكماش المقدر نحو الصفر، وهذا ينطوي في الواقع على إدخال بعض الانحراف لتقليل التباين، مع النتيجة النهائية لتقليل متوسط مربعات الخطأ.

وتَمّ استعمال طرائق عدّة للتقليص ومن هذه الطرائق طريقة Lasso المقترحة من قبل الباحث Tibshirani في عام(1996) [5]، اذ تعمل هذه الطريقة على تقدير نماذج البيانات المبعثرة التي تقوم بتحديد المتغيّرات المعنوية وتقليص معامل الانحدار بشكل متزامن

9- *طرائق التقدير* (*Estimation methods*)

تَمّ استعمال بعض طرائق التقدير الحصينة لتقدير انموذج الانحدار الخطّي متعدّد المتغيّرات منها:-

10- طريقة Lasso

تَمّ اقتراح مقدر (Lasso) من قبل الباحث (Tibshirani) في عام 1996، ويُعدّ أحد المقدرات الجزائية التي تستعمل لتقدير معالم نموذج الانحدار الخطّي المتعدّد ويعبر عنها باختصار (Least Absolute Shrinkage and Selection Operator) وتقوم هذه الطريقة على مبدا تصغير مجموع مربعات الخطأ تبعا للقيد الذي يمثل المجموع المطلق للمعلمات [10].

إذْ أنّ دالة الجزاء (lasso) تعمل على جعل بعض المعالم تساوي صفر، عندما تكون معلمـة الجزاء كبيرة وتقليصها تبعا لمقدار معين، وتعتمد على معلمة الجزاء λ للتحكم بمقدار التقليص (Shrinkage) [11]

ويكتب وفق الصيغة الاتية

$$\hat{B}_{lasso} = argmin_{\beta}\left\{\sum_{i=1}^{n}(y - x_i'B)^2 + n\lambda\sum_{i=1}^{\rho}\left|B_j\right|\right\} \qquad \dots (23)$$

λ : تمثّل معلمة الجزاء

$n\lambda\sum_{i=1}^{\rho}\left|B_j\right|$: تمثّل دالة الجزاء

وتعد طريقة ((lasso أكثر جاذبية في اختيار المتغيّر لأنها تتمتع بخصائص منها وضع بعض المعالم الانحدار مساوية للصفر وتقليص الأخرى بمقدار معين مع تقليل دالة الخسارة، وهي بذلك تعطي نموذج قابل للتفسير بسهولة [12] .

وتوجد العديد من طرائق التي تَمّ اقتراحها من قبل الباحثين الهدف منها حساب المقدر (lasso) منها طريقة التي اقترحت من قبل الباحثين (Friedman& Hastie) والتي تسمى بخوارزمية (Cyclic Coordinate Descent) وسيتَمّ توضيحها في الفقرة (1) .

11- طريقة Multi-split Lasso

تتأسس فكرة هذه الخوارزمية على أطروحة Wasserman,2009 [13] الذي سعى إلى توظيف أل Bootstrap في التقسيم ولكن الباحث ,Meinshausen في عام (2009) [14] طور على الفكرة من خلال تقسيم مجموعة البيانات على قسمين، القسم الأول ينفذ مع طريقة لاسو لاختيار أفضل المتغيّرات، أما القسم الثاني فيعمل على تقدير معلمات الأنموذج الصادر من القسم الأول على إنّ يتَمّ تكرار هذا الأجراء عدة مرات (15 مرة على اقل تقدير) في كل مرة يتَمّ حساب قيمة P للمعالم المقدرة.

لقد حظيتْ هذه الخوارزمية باهتمام العديد من الباحثين.

ليكن B العدد الكلي للتقسيمات العشوائية لعينة البيانات الاصلية بحث ان $b = 1, \dots, B$ هي ترقيم لهذه التقسيمات. اذ في كل مرة تقسم العينة إلى مجوعتين منفصلتين ومتساويتين في الحجم. خوارزمية الـ Multi-splitاقترحها (Meinshausen et al.,2009) وحصنها (Uraibi, 2020) يمكن وصفها كالاتي:

1. البداية
2. ليكن $b = 1$
3. ا فرض إن D العينة الاصلية للبيانات،

$$D = \underset{(n \times m)}{\begin{bmatrix} Y_{11} & \dots & Y_{m1} \\ \vdots & \dots & \vdots \\ Y_{n1} & \dots & Y_{mn} \end{bmatrix}}, \underset{(n \times d)}{\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1d} \\ \vdots & \vdots & \dots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nd} \end{bmatrix}}$$

1. قسمت عشوائيا على مجموعتين منفصلتين متساويتين بالحجم $(n/2)$ ولتكن $D_{in}^{(b)}$ و $D_{out}^{(b)}$ كالاتي:

$$D_{in}^{(b)} = \underset{(n/2 \times m)}{\begin{bmatrix} Y_{11}^{*} & \dots & Y_{1m}^{*} \\ \vdots & \dots & \vdots \\ Y_{\frac{n}{2},1}^{*} & \dots & Y_{\frac{n}{2},m}^{*} \end{bmatrix}}, \underset{(n/2 \times d)}{\begin{bmatrix} X_{11}^{*} & X_{12}^{*} & \dots & X_{1d}^{*} \\ \vdots & \vdots & \dots & \vdots \\ X_{\frac{n}{2}1}^{*} & X_{\frac{n}{2}2}^{*} & \dots & X_{\frac{n}{2}d}^{*} \end{bmatrix}}$$

$$D_{out}^{(b)} = \underset{(n/2 \times m)}{\begin{bmatrix} Y_{11}^{**} & \dots & Y_{1m}^{**} \\ \vdots & \dots & \vdots \\ Y_{\frac{n}{2},1}^{**} & \dots & Y_{\frac{n}{2},m}^{**} \end{bmatrix}}, \underset{(n/2 \times d)}{\begin{bmatrix} X_{11}^{**} & X_{12}^{**} & \dots & X_{1d}^{**} \\ \vdots & \vdots & \dots & \vdots \\ X_{\frac{n}{2}1}^{**} & X_{\frac{n}{2}2}^{**} & \dots & X_{\frac{n}{2}d}^{**} \end{bmatrix}}$$

5. افرض إن $\tilde{S}_{\mathcal{H}}^{(b)} = \{j; \hat{\beta}_j^{Lasoo} \neq 0\}$ معاملات الانحدار المقدرة اللاصفرية لـ (λ) $\hat{\beta}_j^{Lasoo}$والمحسوبة من بيانات $D_{in}^{(b)}$ بحيث إنّ $\tilde{S}_{\mathcal{H}}^{c(b)} = \{j; \beta_j^{\mathcal{H}} = 0\}$ هي المعاملات المقدرة الصفرية. هنا لابد من الإشارة إلى ان Lasso ستعمل على تقدير معاملات كل نموذج على حده. بمعنى ان عدد النماذج يكون مكافئا إلى (m).

6. في هذه الخطوة يتَمّ تقدير مَعّلمات نموذج الانحدار للمتغيرات المختارة $\tilde{S}_{\mathcal{H}}^{(b)}$ باستعمال طريقة المربعات الصغرى مع بيانات $D_{out}^{(b)}$ ومن ثم حساب $\tilde{P}_j^{(b)}$ لها، اما بقية المتغيّرات في $\tilde{S}_{\mathcal{H}}^{c(b)}$ اجعل قيمة $\left(\tilde{P}_j^{(b)} = 1\right)$.

$$\tilde{P}_j^{(b)} = \begin{cases} \tilde{P}_j^{(b)} & \text{if} \quad j \in \tilde{S}_{\mathcal{H}}^{(b)} \\ 1 & \text{if} \quad j \notin \tilde{S}_{\mathcal{H}}^{(b)} \end{cases} \qquad \dots (24)$$

7. ثم اختر قيمة P-value لجميع المتغيّرات من خلال حساب $P_j^{(b)}$

$$P_j^{(b)} = \min\left(\tilde{P}_j^{(b)} \times \left|\tilde{S}_{\mathcal{H}}^{(b)}\right|, 1\right), j = 1,2, \dots, p \qquad \dots (25)$$

and then without aggregated, adjusted $\tilde{P}_j^{(b)}$ values as

$$P_j^{(b)} = \min\left(\tilde{P}_j^{(b)} \times \left|\tilde{S}_{\mathcal{H}}^{(b)}\right|, 1\right), j = 1,2, \dots, p \qquad \dots. (26)$$

8. ارفع قيمة العداد $b = b + 1$

9. هل إن $b < B$ إذا كان الجواب نعم اذهب إلى الخطوة (4)

10. ممّا تقدّم نستطيع القول ان الخطوات السابقة انتجت B متجه لـ $P_j^{(b)}$ ولتجميع هذه المتجهات واستخلاص نتيجة واحدة منها اقترح Meinshausen et al. (2009) لاي قيمة ثابتة بين الصفر والواحد مثل $\gamma \in (0,1)$ يكون الحد الأدنى لها على الأقل 0.05 فان

$$Q_j(\gamma) = \min\left\{1, q_\gamma\left(\left\{\frac{P_j^{(b)}}{\gamma}; b = 1, \dots, B\right\}\right)\right\} \qquad \dots (27)$$

إذْ إنّ $q_{\gamma}(.)$ هي الدالة الكمية quantile التجريبية.

إنّ اختيار γ المناسبة يتطلب إضافة تصحيحات أكثر للسيطرة على معدلات عائلة من الأخطاء عند مستوى معين α ذلك من خلال تصحيح العامل $1 - \log(\gamma_{\min})$ بحد اعلى لا يتجاوز (4). بذلك يمكن تعديل قيمة p-value كالاتي

$$P_j^{\text{rob}} = \min\left\{1, 1 - \log(\gamma_{min}) \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma)\right\} \qquad \dots (28)$$

فإنّ معاملات الانحدار فقط التي P_j^{rob} لها لا تساوي (1) ستكون في الأنموذج النهائي او المختار من قبل هذه الطريقة.

12- معيار المعلومات البيزية (Bayesian Information Criterion BIC)

تَمّ تقديم معيار المعلومات البيزية في عام(1978) من قبل الباحث (Schwarz) [15] والذي يعتمد على النظرية البيزية ويتّمّ توضيح هذا المعيار كما يأتي:

$$BIC = -2\,\ell\left(\hat{\beta}_p\right) + log(n)\,(p+2) \qquad \dots (29)$$

وكما هو الحال في AIC، نبحث عن النماذج ذات القيم الصغيرة من BIC.

ولمعرفة الطريقة الأفضل من ناحية التقدير هي التي تعطي اقل قيمة لمعيار BIC.

13- الجانب التجريبي

المقدمة:

يُعدّ اسلوب المحاكاة أداة قوية وفعالة في مجالات عدّة منها الصناعية والهندسية والطبية والاقتصادية وغيرها ، ويهدف هذا الأسلوب إلى تمثيل ومحاكاة العمليات والاحداث الواقعية بشكل افتراضي ممّا يوفر إمكانية لفهم النظم المعقدة وتحليلها وتقييم تأثير المتغيّرات المختلفة بها.

14- مراحل تطبيق المحاكاة: -

أولاً: يتَمّ تحديد الهدف الرئيس للمحاكاة وتحديد المتغيّرات التوضيحية والمعتمدة التي سيتَمّ دراستها.

ثانيا: يتَمّ تطوير انموذج المحاكاة المناسب بناءا على النظام المراد دراسته إذْ يتَمّ تحليل الأنموذج وتصميمه بشكل دقيق.

ثالثا: يتَمّ توليد الأخطاء العشوائية في الأنموذج ويتَمّ تشغيل المحاكاة لمدّة زمنية محددة لتحليل النتائج وتقيمها ويجب تحليل البيانات المستعملة من المحاكاة بعناية لفهم سلوك النظام وتقدير إثر المتغيّرات المختلفة على النتائج.

وبذلك يمكن استعمال أسلوب المحاكاة لتحسين التخطيط وتنظيم العمليات وتحسين الكفاءة وتوفير التكاليف.

15- الجانب المحاكاة

يتّمّ استعمال لغة البرمجة R-package لدراسة طرائق التقدير وتحليلها لأنموذج الانحدار الخطي متعدد المتغيّرات الاتي:-

$$\underset{[n\times m]}{Y}=\underset{[n\times d]}{X}\underset{[d\times m]}{\beta}+\underset{[n\times m]}{e}$$

إذْ أن X هي مصفوفة المتغيّرات التوضيحية ذات بعد $n\times d$ بدون حد ثابت المولدة من توزيع طبيعي متعدد المتغيّرات بمتوسط قيمته صفراً ومصفوفة تباين وتباين مشترك $\sigma=\rho^{|i-j|}$ اي أن

$$\underset{(n\times d)}{X}\sim N\left[\underset{(d\times 1)}{0},\ \underset{(d\times d)}{\rho^{|i-j|}}\right]$$

إذْ $\rho=0.5$ و$n=\{20,30,70,200,500\}$ ، B هي مصفوفة لمعلمات هذا الأنموذج مع الحد الثابت ذات بعد $(d+1)\times m$، إذْ ان $P=\{10,15,30\}$ هي عدد المتغيّرات في الأنموذج، أخيرا e هي مصفوفة الأخطاء العشوائية للنموذج. اما Y فهي مصفوفة مكونة من n مشاهدات لـ $m=2$ من المتغيّرات المعتمدة ولدت كالاتي:

- توليد المصفوفة Z ذات البعد $(n\times m)$ و $\rho=0.5$ لضمان وجود علاقة ارتباط خطية معتدلة بين متغيرات الاستجابة.

$$\underset{(n\times m)}{Z}\sim N\left[\underset{(m\times 1)}{0},\ \underset{(m\times m)}{\rho^{|i-j|}}\right] ,$$

- حددت الباحثة القيم الأولية لـ B لتكون مناسبة لتوليد نموذجين الانحدار متعدد المتغيّرات كالاتي،

$$\underset{(d\times 2)}{\beta^{(1)}}=\begin{pmatrix}1&1\\0&0\\1&1\\0&0\\0&0\\0&0\\1&1\\0&0\\\vdots&\vdots\\0&0\end{pmatrix}_{(d\times 2)}\qquad \underset{(d\times 2)}{\beta^{(2)}}=\begin{pmatrix}1&0\\0&1\\1&0\\0&0\\0&1\\0&0\\1&1\\0&0\\\vdots&\vdots\\0&0\end{pmatrix}_{(d\times 2)}$$

$$\underset{(n\times 2)}{Y^{(1)}}=\underset{(n\times 2)}{Z}+\underset{(n\times d)}{X}\times\underset{(d\times 2)}{B^{(1)}}+\underset{(n\times 2)}{e}$$

$$\underset{(n\times 2)}{Y^{(2)}}=\underset{(n\times 2)}{Z}+\underset{(n\times d)}{X}\times\underset{(d\times 2)}{B^{(2)}}+\underset{(n\times 2)}{e}$$

إذْ إنّ

$$e \sim N_m(0, \Sigma)$$

استعملت الباحثة الطرائق الجزائية Lasso فضلاً عن الأجراء المقترح Multi-split Lasso وتَمّ تكرار التجربة (5000) مرة. لقد عملت الباحثة على بعض المعايير للمفاضلة بين الطرائق، إذا تَمّ حساب حجم الأنموذج في كل تكرار ثم أخذت متوسط هذه الاحجام لـ (5000) دورة وخصصت الرمز Model size لهذا الغرض. من الأنموذج المختار في كل دورة ، كم عدد المتغيّرات الصحيحة (ذات المعالم اللاصفرية في مصفوفة المعلمات B) في الأنموذج الذي تَمّ اختياره بواسطة الطريقة و من ثم اخذ المعدل لهذه المتغيّرات لبيان قوة الطريقة في اختيار المتغيّرات الصحيحة و اطلقنا عليها الـ Power، من اهم المعايير التي تُميّز طريقة جزائية عن أخرى هي سيطرتها على الأخطاء من النوع الأول Type I error و التي يطلق عليها أيضا أل Overfit و هي الحالة التي يحتوي فيها الأنموذج المختار متغيرا واحدا أو اكثر معلمته صفر أثناء التوليد و لكن الطريقة أعطته قيمة لا صفرية و ضمنته بالأنموذج. لقد وضعت الباحثة ذلك في الحسبان وحسبت معدل الأخطاء من النوع الأول Type I error للدورات الـ (5000) ولكل طريقة. بناء على ذلك فالطريقة التي تعطينا اقل عدد من المتغيّرات أل Overfit أو اقل أخطاء من النوع الأول هي طريقة كفوءة إذا اقترنت كفاءتها بقوتها وقدرتها العالية على اختيار المتغيّرات الصحيحة بنسب اعلى من غيرها. أخيرا طالما وجدنا ألOverfit يجب إن نراعي حساب معدل أل Underfit وهي الحالة التي يكون فيها المتغيّر الذي معلمته لا صفرية غير موجود في الأنموذج المختار. مما لا شك فيه إن حجم الأنموذج له علاقة كبيرة ومباشرة بحالتي ألOverfit والـ Underfitووضع في المعايير للتأكد من سلامة التحليل الإحصائي.

فضلاً عن هذه المعايير سعت الباحثة إلى إيجاد معدلات (تحيز المعلمات المقدرة والانحراف المعياري للأنموذج ودرجة الحرية)وأخيرا الـ BIC للأنموذج. بالنسبة لتقدير تحيز المعلمات للأنموذج المختار، فقد اختارت الباحثة الأسلوب الجمعي لتجنب الوقوع في عرض جداول غير متسقة. من إذْ لا يمكن تخمين عدد المتغيّرات المختارة لكل طريقة فبعضها يختار عدد قليل من المتغيّرات وأخرى قد تختار عدد ضخما من المتغيّرات من الصعب جدولتها لأغراض المقارنة بالطرائق الأخرى. لذلك في كل دورة تكرار حسب الباحثة تقدير التحيز كالاتي:

$$b^{(i)} = \left|\sum_{j=1}^{p} \hat{\beta}_j - \sum_{j=1}^{p} \beta_j\right| = \left|\sum_{j=1}^{p} \hat{\beta}_j - 3\right|$$

وبالمفهوم نفسه وجدت معايير الخطأ المعياري للأنموذج ودرجات الحرية فضلاً عن الـ BIC،

$$STD = \frac{\sum_{i=1}^{5000} std^{(i)}}{5000}$$

$$DF = \frac{\sum_{i=1}^{5000} df^{(i)}}{5000}$$

$$BIC = \frac{\sum_{i=1}^{5000} bic^{(i)}}{5000}$$

من هذه المعايير الطريقة التي تعطي اقل قيمة هي الأفضل ما عدى DF فأننا نبحث عن رقم حدد حسب حجم العينة فالعينة التجريبية التي حجمها (100) فأننا نبحث عن (96) كدرجة حرية مثلى في هذه التجارب.

نتائج الجانب التجريبي (المحاكاة)

يتَمّ إجراء نتائج تجارب المحاكاة لمختلف طرائق التقدير الجداول تعرض نتائج طرائق تقدير لأنموذج الانحدار متعدد المتغيّرات وكما مبين في الجدول رقم(1) ولغاية الجدول رقم (6)

الجدول رقم(1)

يمثل حجم الأنموذج وقوته والـOverfit ، Underfit لطرائق التقدير عندما $P = 10$ و $m = 2$

N	Method	Model	Model. Size	Power	Overfit	Underfit	Best
70	Lasso	M1	3.006	1.446	1.56	0	
		M2	2.908	1.446	1.462	0	**
	Multi-split	M1	1.264	1.254	0.01	0.192	
		M2	1.272	1.266	0.006	0.18	*
200	Lasso	M1	6.004	3	3.004	0	*
		M2	6.146	3	3.146	0	
	Multi-split	M1	3.028	3	0.028	0	**
		M2	3.028	3	0.028	0	**
500	Lasso	M1	6.024	3	3.024	0	*
		M2	6.028	3	3.028	0	
	Multi-split	M1	3.026	3	0.026	0	**
		M2	3.026	3	0.026	0	**

المصدر: من عمل الباحث بالاعتماد على نتائج المحاكاة(R)

ابتداء يجب المقارنة بين الأنموذجين الفرعيين لكل متغير استجابة M1, M2 باستعمال اهم معيار هو أل Overfit وأل Underfit، ثم المفاضلة بين النماذج المختارة لاختيار أفضل أنموذج. الملاحظ عندما حجم العينة (70) هناك تفاوت واضح في حجم الأنموذج بشكل عام. فطريقة Lasso اختارت الأنموذج الثاني M2 و الخوارزمية اختارت M2.

و نلاحظ أنّهناك زيادة عن الرقم (3) وهو المعالم اللاصفرية المستعملة في المحاكاة ونقصان عن هذا الرقم. فالأنموذج الثاني لطريقة Lasso حظي بمعدل (2.908) ألا أن التفاوت كان فقط بوجود خطأ من النوع الأول ظهر في الـOverfit تقريبا (1.462) إن الأنموذج الثاني لطريقة Multi-split قدم خطأ من النوع الأول وصل إلى (0.006) وهو الاقل ولكن هناك مشكلة Underfit بلغت 0.18 وهذه تعد مشكلة كبيرة في الأنموذج المختار من إذْ إن حجم الأنموذج ابتداء كان 1.272.

زيادة حجم العينة إلى 200 أحدث فرقا كبيرا، ليس من إذْ اختيار النماذج الفرعية M1,M2. بل من حيث حجم الأنموذج بقي الاختيار للطرائق التقليدية يدور في فلك ما بين الرقمين (6) و (7)، مع المحافظة على ذات القوة وهذا يعني إن معدل الأخطاء من النوع الأول سوف يدور في فلك الرقم (3) او يزيد عليه قليلا. إذا كانت هذه إشارة إلى استقرار هذه الطرائق لكن بالمقابل هناك تحسن بدقة عالية في أداء خوارزمية الـ Multi-split التي اختارت احجام نماذج مكافئة تماما لعدد المعالم اللاصفرية (3) بخطأ مقداره (0.03) تقريبا وللأنموذجين M1,M2.

من الجدول السابق نلاحظ

أن طريقة Lasso اختارت M2 بدلاً من M1 في حجم العينة 200، ويجدر بالذكر أن أداء هذه الطريقة لم يختلف عن أداءاتها السابقة في استقرار اختيار الحجم بين 6 و7 متغيرات دون مشاكل في التحميل الأقل. كما تَمّ توضيح ذلك سابقًا أن اختيار معدل حجم الأنموذج في هذه الأرقام يشير بلا شك إلى وجود مشكلة في التحميل الزائد. فضلاً عن ذلك، فإن الخوارزمية Multi-split حافظت على أداؤها القوي جدًا في التقليل من ظاهرة الـOverfitting أو تحجيم الأخطاء من النوع الأول في أنموذج الاستجابة الأولى إلى 0.024 والأنموذج الثاني إلى 0.018.

بالتأكيد، زيادة حجم العينة تدريجياً إلى 500 سترفع من دقة التقدير واستقراريه الطرائق. ، يلاحظ الاستقرار التام والقوة القوية للخوارزمية Multi-split، إذْ اختارت حجم أنموذج يقترب جدًا من الرقم 3 وهو الهدف واستقرت قوتها عند 3 وهو هدف آخر أيضًا، ومعدل الخطأ من النوع الأول لا يتجاوز 0.03.

من ذلك، يمكننا استنتاج أن الخوارزمية Multi-split تحسن أداؤها بدقة عالية جدًا مع زيادة حجم العينة.

معايير معدل القيمة المطلقة للتحيز ومعدل الانحرافات المعيارية ومعدل درجات الحرية وأخير معيار الـ BIC للمفاضلة بين النماذج واتخاذ القرار الإحصائي حول افضلية نموذج على اخر.

الجدول رقم (2)

يمثل معدل التحيز، الانحراف المعياري، درجة الحرية ومعيار BIC عندما $P = 10$ و $m = 2$

N	Method	Model	Bias	Sigma	Df	BIC	Sig.
70	Lasso	M1	0.304	1.391	67.86	579.423	*
		M2	0.272	1.381	67.96	582.791	
	Multi-split	M1	0.494	1.5	71.4	578.541	**
		M2	0.515	1.489	71.44	580.492	
200	Lasso	M1	0.169	1.407	192.996	1479.01	*
		M2	0.152	1.403	192.854	1483.994	
	Multi-split	M1	0.13	1.415	195.972	1456.014	
		M2	0.128	1.411	195.972	1455.992	**
500	Lasso	M1	0.107	1.409	492.976	3609.807	*
		M2	0.096	1.409	492.972	3612.397	
	Multi-split	M1	0.082	1.412	495.974	3581.046	**
		M2	0.081	1.412	495.974	3581.043	**

المصدر: من إعداد الباحث بالاعتماد على نتائج المحاكاة (لغة R)

ومن النتائج المعروضة في الجدول (2) عندما يكون حجم العينة (70) مشاهدة ان معيار BIC لنموذج M1 لطريقة Lasso اظهر ان الأنموذج الأول معنويا بشكل اكبر من الأنموذج الثاني M2 على الرغم من أنّ معدل التحيز Bias ، Sigma للنموذج الثاني كانا اقل و درجة الحرية Df افضل. هذا ما يؤكد الاضطراب الحاصل في اختيار الطريقة عندما يكون حجم العينة (70) لان قوة اختيارها للمتغيرات الصحيحة كانت ضعيفة كما هو واضح في الجدول السابق. ان معيار BIC لخوارزمية Multi-split بين المعنوية العالية لنموذجه الثاني M2 بأفضلية حتى على طريقة Lasso. مع ذلك لا يمكن الوثوق بهذه النتائج اطلاقا لان الاستدلال الاحصائي من نتائج الجدول (1) كانت تشير إلى ان الدقة في الاختيار والتقدير ترتفع مع ارتفاع حجم العينة.

من الواضح في نتائج الجدول (2) عنده زيادة حجم العينة إلى (200) أكد اختيار الأنموذج الأول M1 لطريقة Lasso, لتتوافق مع نتائجها في الجدول (1) وتقريبا هذا الاتساق كان متفقا مع خيارات الخوارزمية بدرجة كبيرة جدا، مرة أخرى هناك افضلية واضحة لأنموذج الـMulti-split على نظيراتها من خلال قيم BIC.

الجدول (2) يستعرض نتائج زيادة حجم العينة إلى (300) يمكن ملاحظة ان أنموذجي الخوارزمية Multi-split قد تقتربان فيما بينهما بشكل كبير جدا بحيث ان كليهما تفوقا على النماذج المختارة للطريقة الأخرى وبشكل واضح جدا.

زيادة حجم العينة المولدة بأسلوب المحاكاة إلى (500) كان فيصلا في توافق نتائج الجدول (1) والجدول (2) من ناحية أداء كل طريقة. اكدت الخوارزمية الـMulti-split اداءها المتفوق والثابت لكلا نموذجيها M1,M2 اللذان اظهرا أداءً متقارباً فيما بينهما وافضلية على طريقة Lasso.

الجدول رقم (3)

يمثل حجم الأنموذج وقوته والـOverfit ، Underfit لطرائق التقدير عندما $P = 15$ و $m = 2$

N	Method	Model	Model. Size	Power	Overfit	Underfit	Best
70	Lasso	M1	3.4	1.8	1.6	0	**
		M2	3.7	1.8	1.9	0	
	Multi-split	M1	1.54	1.5	0.04	0.3	
		M2	1.56	1.56	0	0.24	*
200	Lasso	M1	7.14	3	4.14	0	*
		M2	7.84	3	4.84	0	
	Multi-split	M1	3.04	3	0.04	0	
		M2	3	3	0	0	**
500	Lasso	M1	7.2	3	4.2	0	
		M2	7.16	3	4.16	0	*
	Multi-split	M1	3.04	3	0.04	0	

**	0	0	3	3	M2		

المصدر: من إعداد الباحث بالاعتماد على نتائج المحاكاة (لغة R)

الملاحظ عندما حجم العينة (70) هناك تفاوت واضح في حجم الأنموذج بشكل عام. فطريقة Lasso اختارت الأنموذج الاول M1 وخوارزمية Multi-split اختارت M2.

و نلاحظ أنّ هناك زيادة عن الرقم (3) وهو المعالم اللاصفرية المستعملة في المحاكاة ونقصان عن هذا الرقم. فالأنموذج الاول لطريقة Lasso حظي بمعدل (3.4) ألا ان التفاوت كان فقط بوجود خطأ من النوع الأول ظهر في الـOverfit تقريبا (1.6) إن الأنموذج الثاني لطريقة Multi-split قدم خطأ من النوع الأول وصل إلى (0) وهو اقل من Lasso ولكن هناك مشكلة Underfit بلغت 0.24 وهذه تعد مشكلة كبيرة في الأنموذج المختار من إذْ إن حجم الأنموذج ابتداء كان 1.56.

زيادة حجم العينة إلى 200 أحدث فرقا كبيرا، ليس من حيث اختيار النماذج الفرعية M1,M2. من إذْ حجم الأنموذج بقي الاختيار لطريقة Lasso يدور في فلك ما بين الرقمين (6) و (7)، مع المحافظة على ذات القوة وهذا يعني إنّ معدّل الأخطاء من النوع الأول سوف يدور في فلك الرقم (3) او يزيد عليه قليلا. كانت هذه إشارة إلى استقرار هذه الطريقة لكن بالمقابل هناك تحسن بدقة عالية في أداء خوارزمية الـ Multi-split التي اختارت أحجام نماذج مكافئة تماما لعدد المعالم اللاصفرية (3) بخطأ مقداره (0.03) تقريبا وللأنموذجين M1,M2.

من الجدول السابق نلاحظ

أن طريقة Lasso اختارت M1 ايضا كما في حجم العينة 200. ويجدر بالذكر أن أداء هذه الطريقة لم يختلف عن أداءاتها السابقة في استقرار اختيار الحجم بين 6 و7 متغيرات دون مشاكل في التحميل الأقل. كما تَمّ توضيح سابقًا أن اختيار معدل حجم الأنموذج في هذه الأرقام يشير بلا شك إلى وجود مشكلة في التحميل الزائد. فضلاً عن ذلك، فإن خوارزمية Multi-split حافظت على أداؤها القوي جدًا في التقليل من ظاهرة الـOverfitting أو تحجيم الأخطاء من النوع الأول في انموذج الاستجابة الأولى إلى 5.46 والأنموذج الثاني إلى 5.82.

بالتأكيد، زيادة حجم العينة تدريجياً إلى 500 سيرفع من دقة التقدير واستقرار الطريقة Lasso. من ناحية أخرى، يلاحظ وجود استقرار واضح في اختيار الأنموذج الفرعي M1 لطريقة Lasso من حجم 70 إلى حجم 500، إذْ تبين أن أخطاء التحميل الزائد من النوع الأول تتغير في طريقة Lasso إذْ تمثلت(4.2)و(4.16). وبالمقابل، يلاحظ الاستقرار التام والقوة القوية للخوارزمية Multi-split، إذْ اختارت حجم نموذج يقترب جدًا من الرقم 3 وهو الهدف واستقرت قوتها عند 3 وهو هدف آخر أيضًا، ومعدل الخطأ من النوع الأول لا يتجاوز 0.04.

من ذلك، يمكننا استنتاج أن خوارزمية Multi-split تحسن أداءها بدقة عالية جدًا مع زيادة حجم العينة

الجدول رقم (4)

يمثّل معدّل التحيز، الانحراف المعياري، درجة الحرية ومعيار BIC عندما $p = 15$ و $m = 2$

N	Method	Model	Bias	Sigma	Df	BIC	Sig.
70	Lasso	M1	0.386	1.358	66.1	590.779	*
		M2	0.316	1.387	66.56	589.851	
	Multi-split	M1	0.433	1.47	71.36	576.689	
		M2	0.467	1.484	71.36	575.331	**
200	Lasso	M1	0.22	1.388	191.86	1484.426	*
		M2	0.178	1.399	191.16	1491.696	
	Multi-split	M1	0.121	1.402	195.96	1453.701	*
		M2	0.113	1.416	196	1453.694	
500	Lasso	M1	0.11	1.406	491.8	3624.226	
		M2	0.09	1.41	491.84	3623.452	
	Multi-split	M1	0.083	1.412	495.96	3584.678	**
		M2	0.077	1.416	496	3584.689	

المصدر: من إعداد الباحث بالاعتماد على نتائج المحاكاة (لغة R)

ومن النتائج المعروضة في الجدول (4) عندما حجم العينة (70) مشاهدة إنّ معيار BIC لأنموذج M2 لطريقة Lassoاظهر إنّ الأنموذج الثاني معنويّ بشكل أكبر من الأنموذج الأول M1 على الرغم من ان درجة الحرية Df ، Sigma للأنموذج الأول كانا اقل و معدل التحيز Bias افضل في الأنموذج الثاني. هذا ما يؤكد الاضطراب الحاصل في اختيار الطريقة عندما حجم العينة (70) لأنّ قوّة اختيارها للمتغيرات الصحيحة كانت ضعيفة كما هو واضح في الجدول السابق. مع ملاحظة ان معيار BIC لخوارزمية Multi-split بين المعنوية العالية للأنموذج الثاني M1 بأفضلية حتى على طريقة Lasso. مع ذلك

لا يمكن الوثوق بهذه النتائج اطلاقا لان الاستدلال الاحصائي من نتائج الجدول (3) كانت تشير إلى ان الدقة في الاختيار والتقدير ترتفع مع ارتفاع حجم العينة.

من الواضح في نتائج الجدول (4) عندما حجم العينة إلى (200) أكد اختيار الأنموذج الأول M1 لطريقة Lasso لتتوافق مع نتائجها في الجدول (3) وتقريبا هذا الاتساق كان متفقا مع خيارات خوارزمية Multi-split بدرجة كبيرة جدا. مرة أخرى هناك افضلية واضحة لنماذج الـMulti-split على طريقة Lasso من خلال قيم BIC.

زيادة حجم العينة المولدة بأسلوب المحاكاة إلى (500) كان فيصلا في توافق نتائج الجدول (3) والجدول (4) من ناحية أداء كل طريقة. بالمقابل اكدت خوارزمية الـMulti-split اداءها المتفوق والثابت لكلا نموذجيها M1,M2 اللذان اظهرا أداءً متقارباً فيما بينهما وافضلية على طريقة Lasso.

الجدول رقم (5)

يمثل حجم الأنموذج وقوته والـOverfit ، Underfit لطرائق التقدير عندما $p = 30$ و $m = 2$

N	Method	Model	Model.Size	Power	Overfit	Underfit	Best
20	Lasso	M1	12.91	6.13	9.91	0	
		M2	12.61	6.13	9.61	0	**
	Multi-split	M1	20.77	6.13	17.77	0	
		M2	20.31	6.13	17.31	0	**
30	Lasso	M1	11.89	5.11	8.89	0	
		M2	11.59	5.11	8.59	0	**
	Multi-split	M1	19.75	5.11	16.75	0	
		M2	19.29	5.11	16.29	0	**
70	Lasso	M1	9.78	3	6.78	0	
		M2	9.48	3	6.48	0	**
	Multi-split	M1	2.9	2.88	0.02	0.12	*

		M2	2.84	2.78	0.06	0.22	
200	Lasso	M1	10.92	3	7.92	0	
		M2	9.02	3	6.02	0	*
	Multi-split	M1	3.04	3	0.04	0	**
		M2	3.06	3	0.06	0	
500	Lasso	M1	8.1	3	5.1	0	*
		M2	8.78	3	5.78	0	
	Multi-split	M1	3.02	3	0.02	0	**
		M2	3.02	3	0.02	0	**

المصدر: من إعداد الباحث بالاعتماد على نتائج المحاكاة (لغة R)

- في طريقة Lasso، يتَمّ استعمال أنموذج M1 وأنموذج M2. يظهر أن الأنموذج M2 يحقق أداءً أفضل من M1 في معظم الحالات، ما عدا عند N=500 إذْ يتفوق M1.

- في طريقة Multi-split، يتَمّ استعمال نموذج M1 ونموذج M2. يظهر أن الأنموذج M1 يحقق أداءً أفضل من M2 في معظم الحالات، إذْ يتميز بعلامة "*" في معظم الحالات.

عندما حجم العينة (70,30,20) اختارت طريقة Lasso الأنموذج الثانيM2 اما خوارزمية Multi-split عند حجم العينة (30,20) اختارت الأنموذج M2 وعند حجم العينة (70) اختارت الأنموذج M1.

ونلاحظ أنّ هناك زيادة عن الرقم (6.13) وهو المعالم اللاصفرية المستعملة في المحاكاة ونقصان عن هذا الرقم. فالأنموذج الثاني لطريقة Lasso حظي بمعدل (12.61) لكن هذا التفاوت كان فقط بوجود خطأ من النوع الأول فظهر في الـ Overfit تقريبا (9.61). وعند زيادة حجم العينة إلى (30) زادت قيمة المعالم اللاصفرية إلى (5.11) وكان معدل الانموذج الثاني لطريقة Lasso يساوي (11.89) اما وجود الخطأ من النوع الأول ظهر عند الـ Overfit بقيمة (8.89). وعند حجم العينة (70) ظهرت المعالم اللاصفرية المستعملة في المحاكاة بالرقم (3). فالأنموذج الثاني لطريقة Lasso حظي بمعدل (9.48) لكن هذا التفاوت كان فقط بوجود خطأ من النوع الأول ظهر في الـ Overfit تقريبا (6.48) .

زيادة حجم العينة إلى 200 أحدث فرقا كبيرا، ليس من إذْ اختيار النماذج الفرعية M1,M2. بل من حيث حجم الأنموذج بقي الاختيار لطريقة Lasso يدور في فلك ما بين الرقمين (6) و (7)، مع المحافظة على ذات القوة وهذا يعني إن معدل الأخطاء

من النوع الأول سوف يدور في فلك الرقم (3) او يزيد عليه قليلا. إذا كانت هذه إشارة إلى استقرار هذه الطريقة الا انه بالمقابل هناك تحسن بدقة عالية في أداء خوارزمية الـ Multi-split التي اختارت احجام نماذج مكافئة تماما لعدد المعالم اللاصفرية (3) بخطأ مقداره (0.04) للأنموذج الاول M1 وللأنموذج M2 (0.06).

من الجدول السابق نلاحظ

أن طريقة Lasso اختارت M1 ايضا كما في حجم العينة 200. ويجدر بالذكر أن أداء هذه الطريقة لم يختلف عن أداءاتها السابقة في استقرار اختيار الحجم بين 6 و7 متغيرات دون مشاكل في التحميل الأقل. كما تَمّ توضيح سابقًا أن اختيار معدل حجم الأنموذج في هذه الأرقام يشير بلا شك إلى وجود مشكلة في التحميل الزائد. فضلاً عن ذلك، فإن خوارزمية Multi-split حافظت على أداؤها القوي جدًا في التقليل من ظاهرة الـ Overfitting أو تحجيم الأخطاء من النوع الأول في انموذج الاستجابة الأولى إلى 5.46 والأنموذج الثاني إلى 5.82.

بالتأكيد، زيادة حجم العينة تدريجياً إلى 500 سيرفع من دقة التقدير واستقراريه الطريقة. لذا من ناحية أخرى، يلاحظ وجود استقرار واضح في اختيار الأنموذج الفرعي M1 لدى طريقة Lasso من حجم 70 إلى حجم 500، إذْ تبين أن أخطاء التحميل الزائد من النوع الأول تتغير في طريقة Lasso إذْ تمثلت (4.2)و(4.16). وبالمقابل، يلاحظ الاستقرار التام والقوة القوية لخوارزمية Multi-split، إذْ اختارت حجم انموذج يقترب جدًا من الرقم 3 وهو الهدف واستقرت قوتها عند 3 وهو هدف آخر أيضًا، ومعدل الخطأ من النوع الأول لا يتجاوز 0.02.

من ذلك، يمكننا استنتاج أن خوارزمية Multi-split تحسن أداؤها بدقة عالية جدًا مع زيادة حجم العينة.

الجدول رقم (6)

يمثل معدل التحيز، الانحراف المعياري، درجة الحرية ومعيار BIC عندما $p = 30$ و $m = 2$

N	Method	Model	Bias	Sigma	Df	BIC	Sig.
20	Lasso	M1	2.618	3.416	66.34	593.382	*
		M2	2.481	3.423	66.64	600.434	
	Multi-split	M1	2.629	3.422	58.94	639.165	
		M2	2.513	3.420	58.48	622.355	*
30	Lasso	M1	1.608	2.406	65.33	592.372	*
		M2	1.471	2.413	65.63	599.424	

	Multi-split	M1	1.619	2.412	57.93	638.155	
		M2	1.503	2.410	57.47	631.345	*
70	Lasso	M1	0.498	1.296	64.22	591.262	*
		M2	0.361	1.303	64.52	598.314	
	Multi-split	M1	0.266	1.407	71.1	564.072	**
		M2	0.332	1.419	71.16	564.925	
200	Lasso	M1	0.241	1.354	188.08	1505.385	
		M2	0.186	1.381	189.98	1494.335	*
	Multi-split	M1	0.116	1.393	195.96	1450.24	
		M2	0.119	1.411	195.94	1450.039	**
500	Lasso	M1	0.182	1.396	490.9	3633.056	*
		M2	0.141	1.414	490.22	3634.945	
	Multi-split	M1	0.093	1.407	495.98	3587.359	**
		M2	0.092	1.426	495.98	3587.393	

المصدر: من إعداد الباحث بالاعتماد على نتائج المحاكاة (لغة R)

ومن النتائج المعروضة في الجدول (6) عندما يكون حجم العينة (30,20) مشاهدة فإنّ معيار BIC لأنموذج M1 لطريقة Lasso اظهر أنّ الأنموذج الاول معنويّ بشكل اكبر من الأنموذج الثاني M2 ،وان درجة الحرية Df، وSigma للأنموذج الاول كانا اقل و معدل التحيز Bias افضل في الأنموذج الثاني. اما خوارزمية Multi-split عندما حجم العينة (30,20) مشاهدة ان معيار BIC لأنموذج M2 أظهر ان الأنموذج الثاني معنويّ بشكل اكبر من الأنموذج الاول M1 ،وان درجة الحرية Df، وSigma للأنموذج الثاني كانا اقل و معدل التحيز Bias افضل في الأنموذج الأول. وعندما حجم العينة (70) مشاهدة فإنّ معيار BIC لأنموذج M1 لطريقة Lasso أظهر ان الأنموذج الاول معنويّ بشكل اكبر من الأنموذج الثاني M2 ،وان درجة الحرية Df، وSigma للأنموذج الأول كانا اقل و معدل التحيز Bias افضل في الأنموذج الثاني. هذا ما يؤكد الاضطراب الحاصل في اختيار الطريقة عندما حجم العينة (70) لان قوة اختيارها للمتغيرات الصحيحة كانت ضعيفة كما هو واضح في الجدول السابق. مع ملاحظة ان معيار BIC لخوارزمية Multi-split بين المعنوية العالية للأنموذج الاول M1

بأفضلية على طريقة Lasso. مع ذلك لا يمكن الوثوق بهذه النتائج اطلاقا لان الاستدلال الاحصائي من نتائج الجدول (5) كانت تشير إلى ان الدقة في الاختيار والتقدير ترتفع مع ارتفاع حجم العينة.

من الواضح في نتائج الجدول (6) عنده زيادة حجم العينة إلى (200) أكدت ان خوارزمية الـ Multi-split وطريقة Lasso فقد فضلت الأنموذج الثاني M2 على الأنموذج الاول M1. مرة أخرى اثبتت الـ Multi-split افضلية واضحة للنماذج على نظيرتها من خلال قيم BIC.

زيادة حجم العينة المولدة بأسلوب المحاكاة إلى (500) كان فيصلا في توافق نتائج الجدول (5) والجدول (6) من ناحية أداء كل طريقة. اكدت خوارزمية الـ Multi-split اداءها المتفوق والثابت لكلا نموذجيها M1,M2 اللذان اظهرا أداء متقاربا فيما بينهما وافضلية على طريقة Lasso.

16- الاستنتاجات:

1. تبين أن تحليل البيانات ذات الأبعاد العالية يشكل تحديًا في تقدير نماذج الانحدار المتعددة المتغيرات، إذْ يؤدي حجم البيانات الكبير وعدم توافر عدد كافٍ من العينات إلى تقديرات غير مستقرة وتضخم في الأخطاء القياسية.

2. طريقة Lasso تعد أداة فعالة لتقدير نماذج الانحدار في بيانات عالية الأبعاد، إذْ تساعد في تحسين الأداء وتقليل التباين والتحكم في التعقيد. تظهر هذه الطريقة قدرة على اختيار المتغيّرات المهمة والتخلص من غير الضرورية.

3. اكدت خوارزمية الـ Multi-split اداءها المتفوق والثابت لكلا نموذجيها اللذان اظهرا أداء متقاربا فيما بينهما وافضلية على طريقة Lasso ولاسيّما عند حجوم العينات الكبيرة.

17- التوصيات:

1. ينصح بمواصلة البحث والتطوير في طرائق تقدير نماذج الانحدار متعددة المتغيّرات في بيانات عالية الأبعاد، لاسيّما فيما يتعلق بتحسين أداء طريقة Lasso وتعديلها لمعالجة أي قيود أو تحسينات محتملة.

2. يوصى بدراسة واستعمال طرائق أخرى مثل طريقة SCADوطريقة MCP وغيرها، وذلك لتقييم أداؤها وفعاليتها في تحليل بيانات عالية الأبعاد وتقدير نماذج الانحدار.

3. ينبغي أيضًا مواصلة البحث في تقنيات تحليل البيانات الحديثة والذكاء الاصطناعي، مثل شبكات التعلم العميق وتقنيات التحسين العالية الأبعاد، لتحسين القدرة على التنبؤ وتقدير النماذج في بيانات عالية الأبعاد.

المصادر

1- Breiman, L. (1995). Better subset selection using the non-negative garotte. Technometrics, 37(4):373–384.

2- Burnham, Kenneth P, & Anderson, David R. (2002). Model selection and multimodel inference: a practical information-theoretic approach: Springer Science & Business Media.
3- Jenan Nasha and Mohammad Ass'ad(2015), Estimation of Multivariate Multiple Linear Regression Models and Applications, An-Najah National University, Thesis Master of Mathematics
4- Bakin, S. (1999). Adaptive regression and model selection in data mining problems. PhD Thesis, School of Mathematical Sciences, The Australian National University, Canberra.
5- Tibshirani, Robert. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.
6- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360.
7- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. The Journal of Machine Learning Research, 7:2541–2563.
8- Zou, Hui. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.
9- Brown, J. (1993). Measurement, Regression and Calibration. Oxford University Press: Oxford, UK.
10- Jatherine, Stuart (2011) " Robust Regression", Springer
11- Zhang, C. H. and Zhang, S. (2014) Confidence intervals for low dimensional parameters in high-dimensional linear models. Journal of the Royal Statistical Society, Series B, 76.
12- Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014) A significance test for the lasso. Annals of Statistics,42(2).
13- Wasserman, L. and Roeder, K. (2009),High Dimensional Variable Selection. Annals of Statistics, 37(5A)
14- Meinshausen, N., Meier, L. and Bu¨hlmann, P. (2009),P-values for Highdimensional Regression. Journal of the American Statistical Association, (104);
15- Schwarz, Gideon. (1978). Estimating the dimension of a model. The annals of statistics, 6(2), 461-464.