تحليل البيانات عالية الأبعاد باستعمال نماذج انحدار

islasso ،adaptive lasso ،lasso

الباحث سهيل سهيل كريم أ.م. د. أسماء نجم عبد الله

قسم الإحصاء - كلية الإدارة والاقتصاد - جامعة بغداد

الملخص: اكتسبت طرق تحليل الانحدار (*regression analysis*) سمعة جيدة لكونها من أكثر تقنيات النمذجة المعروفة على نطاق واسع في العديد من المجالات العلمية والتطبيقية، اذ تسمح بتقدير تأثير المتغيرات التوضيحية على المتغير المعتمد وذلك من خلال إرجاع تقديرات المعلمات والأخطاء المعيارية الموثوقة لحساب فترات الثقة وقيم (p-value)، وتظهر مشكلة البحث عندما تكون البيانات عالية الأبعاد ذات تركيبة معقدة تضم العديد من المتغيرات المشتركة غير المعلوماتية (*uninformative variables*)، عندئذ ستكون المعالم (*parameters*) أكبر من حجم العينة أي إن (p>n)، مما يصعب عملية اختيار المتغيرات المشتركة المناسبة التي يجب تضمينها في النموذج، وتقدير معالم النموذج، في إن واحد.

ويهدف البحث الى تحليل البيانات العالية الأبعاد بإطار تجريبي جديد مكافئ تقاربياً لانحدار (*lasso*)، ولكنه قادر في حالة العينات المحدودة والمتوسطة الحجم على تقدير المعالم بدقة، يسمى هذا الإطار بأنموذج انحدار التمهيد المستحث ذو اقل انكماش مطلق لاختيار العامل، ويشار اليه اختصارا (*islasso*)، كما يهدف البحث الى المقارنة بين انحدار (*lasso*) وطريقة بين انحدار (*adaptive lasso*)، وطريقة انحدار (*islasso*) المقترحة من قبل (*Cilluffo* وآخرون، عام ٢٠١٩). ومن اهم الاستنتاجات التي تم التوصل اليها، أن عملية تقليص المعاملات وتقدير المعالم باستعمال طريقة (*islasso*) أفضل من طريقة (*lasso*) وطريقة (*adaptive lasso*) كونها تعطي متوسط مربعات خطأ (*MSE*) اقل في حالة العينات الصغيرة، كما يمكن من خلالها الحصول على قيم إحصاءه (*Wald -Chi Squared*) بسهولة نسبياً.

الكلمات المفتاحية: انحدار *islasso*، إحصاءه وايلد، البيانات عالية الأبعاد.

Abstract: Regression analysis has gained a good reputation for being one of the most widely known modeling techniques in many scientific and applied fields. (p-value). The research problem appears when the high-dimensional data includes many uninformative variables, then the parameters will be larger than the sample size, i.e. (p>n), which makes it difficult to selection variables that should be included in the model, and estimating the parameters of the model, at the same time. The research aims to analyze high-dimensional data with a new experimental framework that is asymptotically equivalent to (lasso) regression, but it is able, in the case of limited and medium-sized samples, to accurately estimate the parameters.

The research also aims to compare between (lasso) and (islasso) method proposed by (Cilluffo et al., 2019). Among the most important conclusions that have been reached, the process of reducing coefficients and estimating parameters using the (islasso) method is better than (lasso) and (adaptive lasso) method as it gives less mean error squares (MSE) in case of small samples, also can be Through it, obtaining the values of his statistic (Wald - Chi Squared) is relatively easy.

Keywords: islasso regression, wild chi-statistic, high-dimensional data.

١- المقدمـــــــــة Introduction

اجتذب موضوع اختيار المتغيرات في المساحات عالية الأبعاد (غالباً ما تكون بالمئات أو الآلاف من الأبعاد) اهتماماً كبيرًا في أبحاث استخراج او تعدين البيانات (*Data Mining*) في السنوات السابقة، وهو شائع في العديد من المشكلات الحقيقية، إذ أن عملية اختيار مجموعة فرعية مثالية من المتغيرات أو البيانات وفقًا لمعيار معين وخصائص وميزات معينة من مجموعة من البيانات، يجب أن يتم وفقًا للغرض من اختيار الميزة أو الخاصية، وعادةً ما تهدف هذه العملية الى تحسين دقة التنبؤ لخوارزمية استخراج البيانات المستعملة. بشكل عام، يكون الهدف هو تحديد الميزات المهمة في مجموعة البيانات وتجاهل الميزات الأخرى باعتبارها زائدة عن الحاجة أو غير ذات صلة. مع ذلك، فإن تحليل البيانات عالية الأبعاد والبيانات ذات التركيبة المعقدة المستويات باستعمال نماذج الانحدار عالية الأبعاد (*high dimensional regression*) يطرح بعض المشاكل المرتبطة بتعقيد النموذج او وجود متغيرات غير مفيدة او غير معلوماتية (*uninformative variables*). كما إن قرار التحكم في المتغيرات المشتركة، وكيفية اختيار المتغيرات المشتركة التي يجب تضمينها في النموذج، يمكن أن يؤدي الى الفشل في اختيار المتغيرات المشتركة الصالحة، وبالتالي يتم الحصول على تقديرات معلمات متحيزة (*biased parameters*) في التجارب العشوائية. **[3, pp. 3-4]**

أن الحل المناسب لتسهيل عملية تحليل ودراسة بيانات كهذه هو استعمال أسلوب الانحدار ذو اقل انكماش مطلق لاختيار العامل (*least absolutes shrinkage and selection operator*) والذي يشار اليه اختصاراً بالرمز (*lasso*)، والذي تم اقتراحه من قبل الباحث (*Robert Tibshirani*) فــي عــام (١٩٩٦) كطريقة جديدة للتقدير في النماذج الخطية، حيث يعتبر هذا الأسلوب واسع الانتشار نسبياً، إذ تم تطبيقه في العديد من الأبحاث البيولوجية والطبية لاكتشاف الارتباطات المحتملة بين عوامل الخطر والأمراض ذات الصلة، فضلاً عن تحسين التنبؤ والتحقق من صحة النتائج.

مؤخرا تم تطوير أسلوب الانحدار ذو اقل انكماش مطلق لاختيار العامل (*lasso*) الى إطار تجريبي جديد مكافئ تقاربياً لانحدار (*lasso*)، ولكنه قادر في حالة العينات المحدودة والعينات المتوسطة الحجم على تقدير المعالم بدقة وسهولة من خلال استعمال خوارزميات نيوتن (*Newton-algorithms*) ومصفوفة التغاير (*covariance matrix*)، يسمى هذا الإطار بأنموذج انحدار التمهيد المستحث ذو اقل انكماش مطلق لاختيار العامل (*Induced Smoothing least absolutes shrinkage and selection operator*)، ويشـــار اليـــه اختصـــارا (*islasso*)، وتـــم اقتراحـــه مـــن قبـــل البـــاحث (*Cilluffo*) واخـــرون عـــام ٢٠١٩، للتعامل مع النماذج الإحصائية وتقدير الدوال والمعادلات غير ممهدة التي تمنع تطبيق خوارزميات التقدير والتقارب المعتادة. **[2, p. 348]**

في هذا البحث سيتم التطرق إلى طرق تحليل البيانات عالية الأبعاد والنماذج المعقدة، باستعمال طريقة انحدار (*lasso*)، فضلاً عن توضيح طريقة انحدار (*adaptive lasso*)، وطريقة انحدار (*islasso*)، والمقارنة بين الطرق واستعراض اهم مميزات استخدام كل منهم في تحليل البيانات عالية الأبعاد، خصوصاً عندما تكون البيانات ذات تركيبة معقدة، تضم متغيرات غير معلوماتية (*uninformative variables*).

٢- البيانات عالية الأبعاد *High-dimensional Data*

تشير الأبعاد في الإحصائيات إلى عدد السمات أو الميزات التي تحتوي عليها مجموعة البيانات، ويمكن تمثيل هذه البيانات في جدول بيانات، مع عمود واحد يمثل كل بُعد، لكن من الناحية العملية من الصعب القيام بذلك، ويرجع ذلك إلى أن العديد من المتغيرات تكون مترابطة مع بعضها. أما الأبعاد العالية (*High-Dimensional*) فتشير الى الارتفاع المذهل في عدد الأبعاد، إذ يمكن أن يتجاوز عدد الميزات عدد المشاهدات بحيث تصبح العمليات الحسابية صعبة للغاية، على سبيل المثال، يمكن أن تحتوي المصفوفات الدقيقة التي تقيس التعبير الجيني، على عشرات المئات من العينات، وكل عينة تضم عشرات الآلاف من الجينات، في هذه الحالة عادةً ما يتم تمثيل أي مجموعة بيانات بواسطة مصفوفة حيث تمثل الصفوف العينات التي تم تسجيلها وتمثل الأعمدة السمات أو الميزات المطلوبة لتمثيل المشكلة المطروحة، ثم يتم تلخيص مجموعة البيانات من خلال إيجاد مصفوفات أضيق (أو أصغر) تكون قريبة إلى حد ما من الأصل، اذ تحتوي المصفوفات الضيقة (*narrow matrices*) على عدد صغير من العينات و/ أو عدد صغير من السمات، وبالتالي يمكن استخدامها بشكل أكثر كفاءة من المصفوفة الكبيرة الأصلية، وتسمى عملية العثور على المصفوفات الضيقة بتقليل الأبعاد (*dimensionality reduction*). **[7, pp. 1-2]**

وبسبب التطور السريع في تقنيات المعلومات وتطبيقاتها في التجارب العلمية أصبحت الإحصاءات عالية الأبعاد شائعة بشكل متزايد، ففي النظام عالي الأبعاد، يكون تجميع المتغيرات واستغلال بنية المجموعة أمرا طبيعيا تماماً. ومن وجهة نظر عملية، يبدو أنه لا غنى عن تجاوز نهج استنتاج معاملات الانحدار الفردية خصوصاً عندما لا يتعلق الاهتمام بمتغير واحد فقط بل مجموعة من المتغيرات. مؤخرا ومع تزايد تعقيد وحجم البيانات المتاحة في مجال التعلم الآلي، تم توظيف

تقنيات تقليل الأبعاد لمعالجة استخراج السمات أو الميزات (الحصول على مجموعة الميزات الأكثر احكاماً والأغنى بالمعلومات لمشكلة معينة) لتكون قادرة على وضع نموذج أفضل للعملية الأساسية لتوليد البيانات. **[8, p. 2]**

٣- انحدار (*Lasso*) *Lasso Regression Model*

أحد أساليب الانحدار الذي يستخدم الانكماش (المكان الذي تتقلص فيه قيم البيانات باتجاه نقطة مركزية (*central point*))، تم اقتراحه من قبل الباحث (*Robert Tibshirani*) فـي عــام (١٩٩٦)، ويعتبر مناسب تماماً للنماذج التي تعرض مستويات عالية من التعددات الخطية (*multicollinearity*) أو عندما نريد اختيار أجزاء معينة من النموذج، مثل اختيار متغير (أو حذف المعلمة).

ولنفرض لدينا نموذج الانحدار الخطي بحيث أن $y = (y_1, \dots, y_n)^T$ ، هو متجه متغيرات الاستجابة للمشاهدات i^{th} ، $i = 1, 2, \dots, N$ والمتجه $x_j = (x_{1j}, \dots, x_{nj})^T$ لكل $j = 1,2, \dots, p$ يمثل المتجه المتغيرات التنبؤية (التوضيحية)، وان $\hat{\beta} = (\beta_1, \dots, \beta_p)^T$ هو متجه المعلمات، فان تقديرات (*Lasso*) تعرف بالصيغة التالية: **[4, p. 268]**

$$\hat{\beta}_{(lasso)} = min \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad \dots .(1)$$

$$S.t. \quad \sum_j |\beta_j| \le \lambda$$

حيث أن $\lambda \ge 0$ هي معلمة الضبط، وتتحكم في قوة دالة الجزاء، وفي مقدار الانكماش الذي يتم تضمينه في التقديرات. فعندما $\lambda = 0$، فأنه لا يتم حذف أي معلمة، وعند زيادة قيمة λ، يتم تعيين المزيد من المعاملات نحو الصفر. ولنفرض أن $\hat{\beta}_j$ هي تقديرات المربعات الصغرى، وان $\lambda_0 = \sum |\hat{\beta}_j|$، فأن القيم $\lambda < \lambda_0$، سوف تسبب انكماش في الحلول نحو الصفر، وقد تكون بعض المعاملات مساوية تماماً للصفر ويتم إزالتها من النموذج. حيث تؤدي دوال الجزاء إلى جعل قيم المعاملات أقرب إلى الصفر مما يؤدي إلى إنتاج نماذج انحدار ابسط.

أن أنموذج انحدار (*lasso*) مع دالة الجزاء $l1$ مبني على فكرة تصغير غاروت (*garotte*) غير السلبي (*non-negative garotte minimizes*) والذي يجد مجموعة من عوامل التحجيم غير السلبية $c = \{c_j\}$ لتقليل المقدار:

$$\sum_{i=1}^{N} \left(y_i - \sum_j c_j \hat{\beta}_j x_{ij} \right)^2 \quad s.t \quad c_j \ge 0 , \quad \sum_j c_j \le \lambda \qquad \dots\dots (2)$$

إذ يعتمد تصغير (*garotte*) على حجم تقديرات (OLS) ويقلصها بعوامل غير سلبية يكون مجموعها مقيداً. ودائما يمتلك تصغير غاروت خطأ تنبؤ أقل من انحدار الحرف، ماعدا الحالة التي يضم فيها النموذج العديد من المعاملات الصغيرة غير الصفرية. **[6, p. 269]**

ومن الناحية العملية فإن نموذج انحدار (*lasso*) يخلق تحيزات مفرطة عند اختيار المتغيرات الهامة ويكون غير متسق من حيث اختيار المتغير، وهذا يعني أن مجموعة المتغيرات التي تم اختيارها بواسطة (*lasso*)، لا تتكون بشكل ثابت من مجموعة حقيقية من المتغيرات المهمة، ولجعل (*lasso*) مقاوما للقيم المتطرفة وتوزيعات الخطأ ذات التطرف الثقيل (*heavy-tailed errors*)، تم إيجاد طرق جديدة لتقدير المعلمات واختيار متغير في وقت واحد. **[5, pp. 273-274]**

وبما أن تقدير (*lasso*) هو دالة غير خطية وغير قابلة للتفاضل لقيم الاستجابة حتى بالنسبة للقيمة الثابتة لـ λ ، فمن الصعب الحصول على تقدير دقيق لخطأها المعياري (*standard error*). ولإجراء ذلك فأن أحد الأساليب المعتمدة هو استعمال التمهيد (*bootstrap*)، إما أن يكون λ ثابتًا، أو قد نقوم بتحسين أكثر من λ لكل عينة ممهدة، ثم استخدام الخطأ المعياري للمربعات الصغرى لتلك المجموعة الفرعية، اذ يمكن التعبير عن تقديرات $\tilde{\beta}$ بالصيغة التالية:

$$\tilde{\beta} = (X^t X + \lambda W^-)^{-1} X^t Y \qquad \dots\dots (3)$$

كما يمكن التعبير عن مصفوفة التغاير (*covariance matrix*) للتقديرات بالصيغة التالية:

$$\hat{\sigma}^2 \, (X^t X + \lambda W^-)^{-1} \, X^t X \, \, (X^t X + \lambda X^-)^{-1} \qquad \dots\dots (4)$$

حيث يمثل الرمز W المصفوفة القطرية ذات العناصر القطرية $|\tilde{\beta}_j|$، ويشير الرمز W^- الى معكوس المصفوفة W العام (*general inverse*)، وان $\hat{\sigma}^2$ هو تقدير التباين الخطأ (*error variance*). وتكمن الصعوبة في الصيغة (4) أنها تعطي تبايناً تقديرياً بقيمة ٠ عندما $\hat{\beta}_j = 0$. **[6, pp. 272-273]**

٤- انحدار (*Lasso*) التكيفي *Adaptive Lasso Regression*

تعتبر طريقة الانحدار ذو أقل انكماش مطلق تكيفي واختيار العامل (*adaptive lasso*) التي تم اقتراحها من قبل الباحث (*Hui Zou*)، في عام (٢٠٠٦) نسخة جديدة من أنموذج انحدار (*lasso*)، فمن المعروف أن تقديرات (*lasso*) تكون منحازة للمعاملات الكبيرة، بينما في انحدار (*lasso*) التكيفي يتحكم في انحياز التقديرات من خلال استعمال الأوزان التكيفية (*adaptive weights*) لاستبعاد المعاملات المختلفة غير الهامة (أو غير المعنوية) في انحدار الجزاء (*penalize regression*)، بالتالي فأن اختيار المتغير باستخدام أسلوب (*adaptive lasso*) يكون متسقاً (*consistent*). وبتعيين أوزان مختلفة لمعاملات انحدار (*lasso*) المختلفة فان: **[1, p. 10]**

$$min \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} w_j \, |\beta_j| \qquad \ldots .(5)$$

حيث يمثل الرمز w متجه الاوزان. فإذا كانت الأوزان تعتمد على البيانات وتم اختيارها بدقة، فيمكن أن يمتلك (*lasso*) الموزون خصائص أوراكل (*oracle properties*). بالتالي فان تقديرات (*adaptive lasso*) يمكن أن تعرف بالصيغة التالية:

$$\hat{\beta}_{(adaptive\ lasso)} = min \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \, |\beta_j| \qquad \ldots .(6)$$

حيث أن $\hat{w} = 1/|\hat{\beta}|^{\gamma}$ هو متجه الاوزان المقدر، وان $\hat{\beta}$ هو مقدر (OLS) متسق لـ $\hat{\beta}_{(adaptive\ lasso)}$
أن طريقة انحدار (*lasso*) التكيفي هي أساساً طريقة انحدار جزاء $l1$. حيث يمكن استخدام الخوارزميات الفعالة المستعملة لحل (*lasso*) لحساب تقديرات (*lasso*) التكيفية. **[9, p. 1419]**

ويمكن الحصول على تقديرات $\tilde{\beta}$ عن طريق الحساب المتكرر (*iteratively computing*) لانحدار الحرف (*ridge regression*) كما في الصيغة التالية:

$$\tilde{\beta} = (X_d^t \, X_d + \lambda_n \sum \beta_0)^{-1} \, X_d^t \, Y \qquad \ldots\ldots(7)$$

حيث يمثل الرمز X_d أول d من الأعمدة في المصفوفة X، كما يمكن التعبير عن مصفوفة التغاير (*covariance matrix*) المقدرة للمكونات غير الصفرية لتقديرات (*lasso*) التكيفية بالصيغة التالية:

$$\sigma^2 \left(\mathbf{X}_{\mathcal{A}_n^*}^T \mathbf{X}_{\mathcal{A}_n^*} + \lambda_n \Sigma\left(\hat{\beta}_{\mathcal{A}_n^*}^{*(n)}\right)\right)^{-1} \times \mathbf{X}_{\mathcal{A}_n^*}^T \mathbf{X}_{\mathcal{A}_n^*} \left(\mathbf{X}_{\mathcal{A}_n^*}^T \mathbf{X}_{\mathcal{A}_n^*} + \lambda_n \Sigma\left(\hat{\beta}_{\mathcal{A}_n^*}^{*(n)}\right)\right)^{-1} \qquad \ldots .(8)$$

فاذا كان التباين σ^2 غير معلوم، فانه يمكن استبداله بتقديراته من النموذج الكلي، اما بالنسبة للمتغيرات ذات $\hat{\beta}_j^{*(n)} = 0$ فإن الأخطاء المعيارية المقدرة هي ٠. **[6, p. 269]**

٥- انحدار (*Lasso*) الممهد المستحث *Induced Smoothing Lasso Regression*

يعتبر أنموذج انحدار (*islasso*)، المقترح من قبل الباحث (*Cilluffo*) وأخرون في عام (٢٠١٩)، نسخة جديدة من أنموذج انحدار (*lasso*)، لتقدير الدوال والمعادلات غير ممهدة التي تمنع تطبيق خوارزميات التقدير المعتادة، كما انه يسمح بالحصول على نتائج جيدة من الناحية العملية، لأنها تؤدي إلى خطأ قياسي غير صفري لتقدير المعلمات الصفرية، مما يساهم في تحديد المتغيرات غير الهامة التي تركت خارج النموذج، ويعتمد أنموذج انحدار (*islasso*) على فكرة التمهيد المستحث (*induced smoothing*) التي اقترحها الباحثان (*Brown & Wang*) في عام (٢٠٠٥).

ولنفرض أن $y = X\beta + \epsilon$ هو نموذج الانحدار الخطي قيد البحث، وان ϵ يمثل متجه الأخطاء ذات المتوسط الصفري والأخطاء المتجانسة وان $y = (y_1, y_2, \ldots, y_n)^t$ يمثل متجه المتغيرات المعتمدة، وان X هي مصفوفة المتغيرات التوضيحية ذات الابعاد $(n * p)$ وان β يمثل متجه معاملات الانحدار، لكي يتم تحقيق هدف انحدار (*lasso*) وهو

تصغير المقدار $\frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$ عند ثبات λ فأن معادلة التقدير الزائفة (*pseudo*) يمكن التعبير عنها بالصيغة التالية:

$$U(\beta) = -X^T(y - X\beta) + \lambda\{2\,\mathrm{I}(\beta > 0) - \mathbf{1}_p\} \qquad \ldots\ldots(9)$$

وبتطبيق طريقة التمهيد المستحث (*Induced Smoothing*) بالاعتماد على التوزيع الخليط للعينات للمقدر β فأن معادلة تقدير التمهيد المستحث يمكن التعبير عنها كما يلي:

$$\widetilde{U}(\beta) = -X^T(y - X\beta) + \lambda\mathcal{P}(\beta, v; c) \qquad \ldots\ldots(10)$$

حيث يمثل الرمز $p(\beta, v; c)$ متجه الجزاء ذي البعد p، والذي يمتلك المركب العام (*generic component*)، $p = c_j\left\{2\phi\left(\beta_j/v_j^{1/2}\right) - 1\right\} + (1 - c_j)\left\{2\phi_\epsilon\left(\beta_j/v_j^{1/2}\right) - 1\right\}$، ويمثل الرمز v القطر الرئيسي للمصفوفة $V = var\,(\hat{\beta})$، ويتضمن c أوزان مخاليط p الأساسية، ويمثل الرمز $\widetilde{U}(\beta)$ المصفوفة الممهدة، بالتالي فأن مصفوفة الانحدار يمكن الحصول عليها بأخذ المشتقة الأولى لمعادلة تقدير التمهيد المستحث $\widetilde{U}(\beta) = \frac{\partial}{\partial\beta}\,\widetilde{U}(\beta)$ وتكون بالصيغة التالية:

$$\widetilde{U}'(\beta) = X^TX + \lambda\,\mathcal{P}'(\beta, v; c) \qquad \ldots\ldots(11)$$

حيث أن مشتق الجزاء هو المصفوفة القطرية ذات الأبعاد $(j * j)$ التي يكون مركبها العام $c_j\left\{2\phi\left(\beta_j/v_j^{1/2}\right) / v_j^{1/2}\right\} + (1 - c_j)\left\{2\phi_\epsilon\left(\beta_j/v_j^{1/2}\right) / v_j^{1/2}\right\}$. ومن خلال المصفوفة الممهدة $\widetilde{U}(\beta)$ يمكن حساب مصفوفة التغاير (*covariance matrix*) كما في الصيغة التالية:

$$V = \widetilde{U}'(\widehat{\beta})^{-1}\,\mathcal{I}\,\widetilde{U}'(\widehat{\beta})^{-1} \qquad \ldots\ldots(12)$$

حيث يمثل الرمز $\hat{\beta}$ القيمة النهائية عند التقارب، وان I هي مصفوفة المعلومات (*Information matrix*) أي أن $I = X^TX$، وتكون مستقلة عن $\hat{\beta}$. **[1, pp. 3-5]**

٦- اختبـار ويلـد - مربع كاي *Wald Chi-Squared Test*

اختبار تقريبي لاختبار نسبة الإمكان (*Likelihood Ratio Test*) يستخدم على نطاق واسع في حالة العينات الكبيرة لمعرفة ما إذا كانت المتغيرات التوضيحية في النموذج معنوية أم لا، ويمكن حذف المتغيرات التي لا تضيف شيئاً دون التأثير على النموذج بأي طريقة ذات معنى. كما يمكن استخدام اختبار (*Wald*) للعديد من النماذج المختلفة بما في ذلك النماذج ذات المتغيرات الثنائية أو المتغيرات المستمرة. أن الفرضية الصفرية لاختبار (*Wald*) هي: بعض المعلمات = بعض القيم، أي أن:

$$\mathrm{H_0}: \beta = 0$$

فاذا تم رفض الفرضية الصفرية، فإنها تشير إلى أنه يمكن إزالة المتغيرات المعنية دون الإضرار كثيراً بملاءمة النموذج قيد الدراسة. وإذا أظهر اختبار (*Wald*) أن معلمات بعض المتغيرات التوضيحية هي صفر، فيمكنك إزالة المتغيرات من النموذج. أما إذا أظهر الاختبار أن المعلمات ليست صفرية، فيجب تضمين تلك المتغيرات في النموذج. ويمكن كتابة صيغة اختبار (*Wald*) كما يلي:

$$W_T = \frac{[\hat{\alpha} - \alpha]^2}{1/I_n(\hat{\alpha})} = I_n(\hat{\alpha})[\hat{\alpha} - \alpha]^2 \qquad \ldots\ldots\ldots(13)$$

حيث يمثل الرمز $\hat{\alpha}$ مقدر الإمكان الأعظم، ويمثل $I_n(\hat{\alpha})$ مصفوفة معلومات فيشر. ونظراً لتقدير $\hat{\beta}$ باستخدام أنموذج انحدار (*islasso*) وبما أن الخطأ المعياري $\mathrm{SE}(\hat{\beta})$ المحسوب على أنه الجذر التربيعي للعناصر القطرية الرئيسية لمصفوفة التغاير فانه يمكن تعريف اختبار (*Wald*) بالصيغة التالية:

$$\mathrm{w_0} = \frac{\hat{\beta}}{\mathrm{SE}(\hat{\beta})} \qquad \ldots\ldots\ldots(13)$$

ويتم الحصول على قيم (*p-value*) تحت $w_0 \rightarrow N(0,1)$. بالتالي فأن الأداء الجيد لاختبار (*Wald*) يعتمد على مدى معولية التقريب العادي (*Normal approximation*) لـ w_0. علاوة على ذلك، تميل صيغة الشطيرة الى التضخيم في تقدير تباين توزيع العينات، مما يجعل إحصاءه (*Wald*) أداة فعالة لاختبار المعاملات غير الصفرية في انحدار (*lasso*). **[1, p. 5]**

٧- الجانب التطبيقي *An Application*

تم تطبيق أنموذج انحدار (*lasso*) ونموذج انحدار (*adaptive lasso*) وأنموذج انحدار (*islasso*)، على مجموعة بيانات حقيقية تهتم بدراسة عوامل الإصابة والتشخيص المبكر لمرض سرطان الثدي، تم الحصول على البيانات من مركز السرطان – وزارة الصحة العراقية، كما تمت المقارنة بين الطرق بالاعتماد على معيار أفضلية التقدير لكل أنموذج باستعمـــــــال جذر متوسط مربعات الخطأ (RMSE)، إذ تتألف عينة البحث من (٢٣٩) مشاهدة و (٣٢) متغير (*variable*) يتضمن:

١. عمر المريض *Increasing age*.
٢. التاريخ الوراثي العائلي *Significant Family History*.
٣. أصغر سن عند الحيض *Younger Age At Menarche*.
٤. قلة النشاط البدني *Lack Of Physical Activity*.
٥. كبار السن في الحمل الأول *Older Age At First Pregnancy*.
٧. استخدام العلاج بالهرمونات البديلة *Use Of Hormone Replacement Therapy*.
٨. اتباع نظام غذائي يفتقر إلى الخضار *Diet Lacking In Vegetable*.
٩. استخدام وسائل منع الحمل عن طريق الفم *Use Of Oral Contraceptives*.
١٠. نوع التشخيص (خبيث M، او حميد B).
١١. زيادة تناول الكحول *Increased Alcohol Intake*.
١٢. التدخين *Smoking*.

كما تم حساب الوسط الحسابي (*mean*) ومعيار الاسوأ (*worst*) أو الأكبر (متوسط أكبر ثلاث قيم) لعشرة ميزات حقيقية القيمة لكل نواة خلية:

أ. نصف القطر *radius*، ويمثل متوسط المسافات من المركز إلى المعلمات على المحيط.
ب. النسيج *texture*، ويمثل الانحراف المعياري لقيم التدرج الرمادي (أي مقياس الرمادية).
جـ. المحيط *perimeter*، الحدود الخارجية للمنطقة.
د. المساحة *area*.
هـ. التمهيد *smoothness*، الاختلاف المحلي في أطوال نصف القطر.
و. حجم التراص أو الاكتناز *compactness*. (*compactness = (perimeter^2 / area-1.0)*).
ز. التقعر أو التجويف *concavity*، شدة الأجزاء المقعرة من الكفاف (*contour*).
حـ. النقاط المقعرة *concave points*، عدد الأجزاء المقعرة من الكفاف.
ط. التناظر *symmetry*.
ي. البعد الكسري *fractal dimension*.

تم تطبيق طرق تحليل الانحدار المستعملة في الدراسة (*lasso*، *adaptive lasso*، *islasso*) والتي تساهم في عملية تقليص عدد المتغيرات التوضيحية والتقدير في أن واحد، باستعمال لغة البرمجة الإحصائية (R)، إذ تم الحصول على النتائج التالية:

جدول (١) قيم متوسط مربعات الخطأ والخطأ المعياري لنماذج انحدار (*lasso, adaptive lasso, islasso*)

Model	*n*	λ	*RMSE*	*Nonzero parameter.*
lasso	239	0.00371	0.22996	27
islasso	239	0.03699	0.23009	23
adaptive lasso	239	0.08191	0.23110	16

ولمعرفة أي من نماذج تحليل الانحدار المستخدمة (*lasso*، *adaptive lasso*، *islasso*) أفضل في تقليص المتغيرات التوضيحية المشتركة وتقدير المعلمات في أن واحد، تم اعتماد معيار المقارنة متوسط مربعات الخطأ Mean Squared Error (MSE)، إذ نلاحظ من الجدول (١) ما يلي:

١. لنماذج الانحدار كافة، أظهرت النتائج أن التقدير باستعمال طريقة (*islasso*) يعطي نتائج متقاربة جداً مع النتائج المستحصل عليها باستعمال انحدار (*lasso*)، وانحدار (*adaptive lasso*)، إذ ما تمت المقارنة على أساس معيار متوسط مربعات الخطأ (MSE)، مع وجود أفضلية نسبية لطريقة انحدار (*lasso*) على بقية الطرق المستعملة قيد الدراسة، حيث بلغت قيمة الجذر التربيعي لمتوسط مربعات الخطأ باستعمال انحدار (*islasso*) بلغت ($RMSE = 0.23009$)، وبلغت قيمة الجذر التربيعي لمتوسط مربعات الخطأ باستعمال انحدار (*lasso*) بلغت ($RMSE = 0.22996$)، أما قيمة الجذر التربيعي لمتوسط مربعات الخطأ باستعمال انحدار (*adaptive lasso*) فقد بلغت ($RMSE = 0.23110$)، مما يعطي أفضلية للتقدير باستعمال انحدار (*lasso*)، وبلغت قيمة معلمة الضبط باستعمال انحدار (*islasso*) بلغت ($\lambda = 0.03699$)، أما قيمة معلمة الضبط باستعمال انحدار (*lasso*) فقد بلغت ($\lambda = 0.00371$)، وبلغت قيمة معلمة الضبط باستعمال انحدار (*adaptive lasso*) بلغت ($\lambda = 0.08191$)، أما عدد المعاملات غير الصفرية فقد بلغ (27) معامل باستخدام أنموذج (*lasso*) وبلغ (23) معامل باستخدام نموذج (*islasso*)، وبلغ (16) معامل باستخدام نموذج (*islasso*).

كما نلاحظ أيضا وجــود أفضلية نسبية للتقديــر باســتعمال طــريقة (*islasso*) على حســاب طــريقة (*adaptive lasso*)، إذ أن قيمة الجذر التربيعي لمتوسط مربعات الخطأ كانت اقل من قيمة الجذر التربيعي لمتوسط مربعات الخطأ باستعمال انحدار (*adaptive lasso*)، مما يعطي أفضلية للتقدير باستعمال انحدار (*islasso*). وعند تطبيق أنموذج انحدار (*islasso*) تم الحصول على التقديرات وحساب الأخطاء المعيارية وقيم إحصاءه (*Wald -Chi Squared*) للمتغيرات التوضيحية فضلاً عن حساب قيم (*p-value*)، كما موضح في الجدول (٢) التالي:

جدول (٢) قيم اختبار Wald -Chi Squared باستخدام أنموذج انحدار (*islasso*) عندما (n=239).

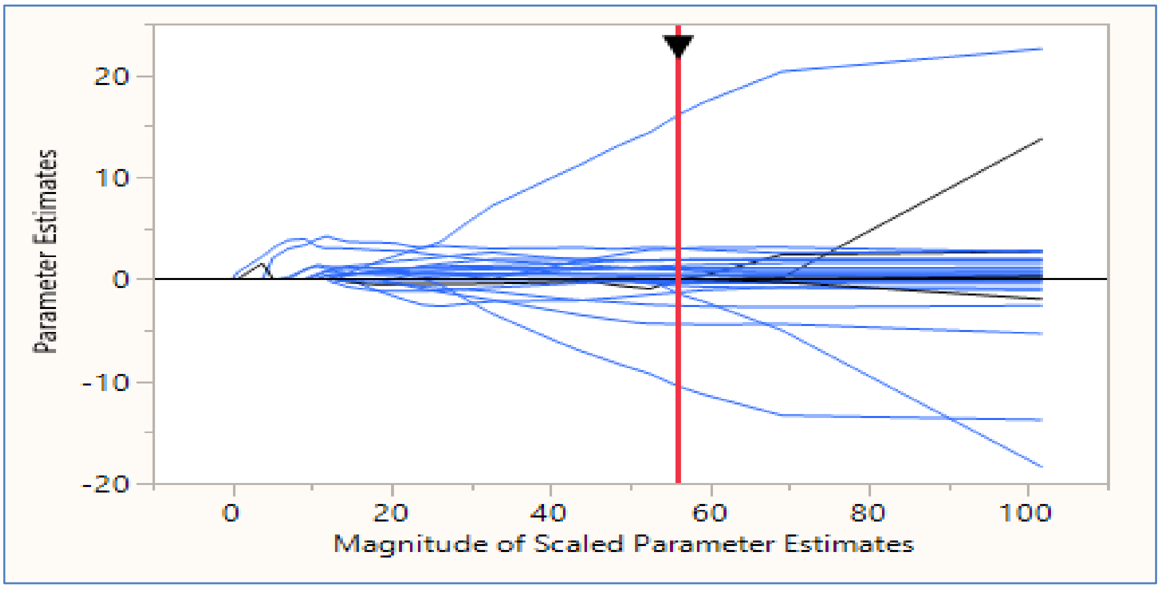
Term	Estimate	Std Error	Wald x^2	P-value
intercept	-0.652	0.337	3.751	0.0528
id	0.000	0.000	0.537	0.4636
radius_mean	0.000	0.000	0.000	1.000
texture_mean	0.005	0.007	0.571	0.4500
perimeter_mean	0.000	0.000	0.000	1.000
area_mean	0.000	0.000	0.000	1.000
smoothness_mean	0.000	0.000	0.000	1.000
compactness_mean	-2.366	0.789	9.002	0.0027*
concavity_mean	0.624	1.033	0.365	0.5455
concave points_mean	2.312	1.774	1.699	0.1925
symmetry_mean	0.000	0.000	0.000	1.0000
fractal_dimension_mean	-3.259	4.674	0.486	0.4857
increasing age	0.462	0.168	7.593	0.0059*
significant family history	0.000	0.000	0.000	1.000
younger age at	0.000	0.000	0.000	1.000
lack of physical activity	-0.002	0.001	4.939	0.0263*
older age at first	14.105	6.752	4.363	0.0367*
use of hormone	-0.736	1.660	0.197	0.6574
diet lacking in vegetable	-2.493	1.033	5.821	0.0158*
use of oral	5.175	4.772	1.176	0.2781
increased alcohol intake	0.178	2.446	0.005	0.9421
smoking	0.000	0.000	0.000	1.000
radius_worst	0.085	0.020	17.802	<0.0001*
texture_worst	0.007	0.005	1.761	0.1844

perimeter_worst	0.000	0.000	0.000	1.0000
area_worst	0.000	0.000	8.436	0.0037*
smoothness_worst	0.850	0.965	0.777	0.3780
compactness_worst	0.000	0.000	0.000	1.0000
concavity_worst	0.401	0.229	3.052	0.0806
concave points_worst	1.009	0.855	1.392	0.2380
symmetry_worst	0.760	0.357	4.529	0.0333*
fractal_dimension_worst	3.428	1.687	4.132	0.0421*
Loglikelihood	٢٣,٧٠٨			

من الجدول (٢) نلاحظ وجود (٢٢) متغير توضيحي من بين (٣٢) متغير يمتلكون تقديرات غير صفرية، كما نلاحظ أيضا وجود (١٠) متغيرات توضيحية تمتلك تقديرات صفرية، بالتالي يمكن إزالة هذه المتغيرات من النموذج كونها لا تضيف شيئاً الى نموذج الانحدار، وبلغ الحد الأدنى لقيمة الاحتمالية (*Loglikelihood*=23.708).

ويوضح الشكل (١) التالي مخطط مسار الحل عندما (n=239)، وكيف يوفر تطبيق أنموذج انحدار (*islasso*) مع توزيع طبيعي قياسي أداة موثوقة لاختبار فرضية عدم وجود تأثير، إذ يتم تمييز مسارات المتغيرات التي لها معاملات غير صفرية باللون الأزرق، كما نلاحظ أن هناك عدد من المتغيرات لها مسارات تم تقليصها إلى الصفر مبكراً تم تمييزها باللون الأسود، ويمثل المحور الرأسي في مخطط مسار الحل قيم تقديرات المعلمات (*parameter estimates*) للتنبؤات المعيارية (*standardized predictors*)، كما يشير الخط الأحمر العمودي إلى قيمها عند الانكماش الأمثل (*optimal shrinkage*)، على النحو المحدد من خلال توظيف معيار العبور الشرعي (*cross validation*).

شكل (١) يوضح أنموذج انحدار (*islasso*) عندما (n=239).



٨- الاستنتاجات *Conclusions*

في هذه البحث، تم تقديم ثلاث طرق حديثة لتحليل انحدار (*lasso*, *adaptive lasso*, *islasso*) لها أهمية بالغة في تحليل البيانات عالية الأبعاد والنماذج المعقدة التي تضم متغيرات مشتركة غير معلوماتية، إذ أنها تساهم في تقليص المتغيرات المشتركة والتقدير في أن واحد، كما تم تطبيق طريقة انحدار (*islasso*) المقترحة من قبل (*Cilluffo* وآخرون، في عام ٢٠١٩)، التي يمكن من خلالها الحصول على قيم إحصاءه (Wald -Chi Squared) بسهولة نسبياً، فضلاً عن سهولة تحديد عرض الحزمة (*bandwidth*) بواسطة الخطأ المعياري (*standard error*) المقابل المحسوب للبيانات. كما نستنتج من خلال نتائج الجانب التطبيقي أن التقدير باستعمال طريقة (*islasso*) يعطي نتائج متقاربة جداً مع النتائج المستحصل عليها باستعمال انحدار (*lasso*)، أي انه كلما زاد حجم العينة وانخفض مقدار الخطأ المعياري، فإن أنموذج انحدار (*islasso*) يقترب من أنموذج انحدار (*lasso*)، مما يجعل أنموذج انحدار (*islasso*) مكافئاً لأنموذج انحدار

(*lasso*). كما نلاحظ أن هنالك أفضلية نسبية للتقدير باستعمال طريقة (*lasso*) على حساب طريقة (*adaptive lasso*)، لأنها تعطي متوسط مربعات الخطأ (MSE) اقل عند المقارنة.

المصادر

1. Cilluffo, G., Sottile, G., La Grutta, S., & Muggeo, V. M. (2020). The Induced Smoothed lasso: A practical framework for hypothesis testing in high dimensional regression. *Statistical Methods in Medical Research*, *29*(3), 765-777.
2. Frost H and Amos C. Gene set selection via lasso penalized regression (SLPR). Nucleic Acids Res 2017; 45: e114.
3. Gavrishchaka, V. V., & Ganguli, S. B. (2003). Volatility forecasting from multiscale and high-dimensional market data. *Neurocomputing*, *55*(1-2), 285-305.
4. Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical statistics*, *9*(2), 319-337.
5. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc: Ser B 2011; 73: 273–282.
6. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Series B 1996; 58: 267–288.
7. Yang Y. Statistical inference for high dimensional regression via constrained lasso. arXiv:1704.05098 [math.ST], Apr. 2017.
8. Zhang X and Cheng G. Simultaneous inference for high-dimensional linear models. J Am Stat Assoc 2017; 112: 757–768.
9. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006; 101: 1418–1429.