ISSN (1681- 6870)

اختبار فرضيات أنموذجي انحدار Islasso وlasso بأستخدام أسلوب المحاكاة

أ. د. دجلة إبراهيم العزاوي [1]،
[1] كلية الإدارة والاقتصاد – جامعة بغداد، بغداد، العراق
dr.dejela.mahdi@coadec.uobaghdad.edu.iq

م. م. رائد فاضل الحسني [2]
[2] كلية الإدارة والاقتصاد – جامعة المستنصرية، بغداد، العراق

المستخلص

في هذا البحث تم اختبار الفرضيات الخاصة بمعاملات معادلة الانحدار الخطي في أنموذج انحدار اقل انكماش مطلق لاختيار العامل (lasso). كما تم تطبيق أنموذج انحدار التمهيد المستحث ذو اقل انكماش مطلق لاختيار العامل (islasso)، المقترح من قبل Cilluffo وآخرون، 2019). ويعتبر أنموذج انحدار (islasso) بديل عن تطبيق أسلوب انحدار الشرائح (regression splines)، إذ يتم تحديد عرض الحزمة (bandwidth) بواسطة الخطأ المعياري (standard error) المقابل المحسوب بالبيانات، كما انه يسمح بالحصول على مصفوفة التغاير وإحصاءه (Wald) بسهولة نسبيًا.

أظهرت نتائج تجارب المحاكاة أفضلية لمقدرات أنموذج انحدار (islasso) في حالة العينات الصغيرة والمتوسطة (n<50)، كما تبين انه كلما زاد حجم العينة وانخفضت قيمة الخطأ المعياري، فإن مقدرات أنموذج انحدار (islasso) تقترب من مقدرات انحدار (lasso)، مما يجعل انحدار (islasso) مكافئاً لانحدار (lasso).

الكلمات المفتاحية: العامل lasso، العامل Islasso ، التجانس المستحث، إحصاءة Wald

Testing Hypotheses in High Dimensional Regression

Prof. Dr. Dijlah Ibrahim Al-Azzawi
University of Baghdad / College of Administration and Economics / Department of Statistics / Iraq.
dr.dejela.mahdi@coadec.uobaghdad.edu.iq

Raed Fadel Mohamed Al-Hassani
Mustansiriyah University / College of Administration and Economics / Iraq.
raad@uomustansiriyah.edu.iq

Abstract:

In this paper, we test the hypotheses concerning the regression coefficients once by using the lasso model, then by using islasso model (proposed by Giovanna et al, 2019), where the islasso model is considered to be an alternative procedure for the regression splines model, where the bandwidth is determined by the corresponding standard error calculated by the data and allows the covariance matrix and Wald statistic to be obtained relatively easily.

The results of the simulation showed better importance for the islasso regression model in the case of small and medium samples (n <50). also, it was found when the sample size increases and the standard error decreases, the islasso gets closer to the original lasso, making the islasso asymptotically equivalent to the lasso.

Keywords: lasso, Islasso, Induced Smoothing, Wald Statistic.

ISSN (1681- 6870)

1. المقدمة: Introduction

تُستخدم نماذج الانحدار على نطاق واسع إذ تعتبر أداة إحصائية فعالة وراسخة في العديد من مجالات البحوث التطبيقية، كما إنها تسمح بتقدير تأثير المتغيرات التوضيحية على المتغير المعتمد وذلك عن طريق إرجاع تقديرات النقطة والأخطاء المعيارية لحساب فترات الثقة والقيم الاحتمالية (p-value). أن نماذج الانحدار عالية الأبعاد (high dimensional regression) تطرح بعض المشاكل المرتبطة بتعقيد النموذج بوجود متغيرات غير معلوماتية (uninformative variables). ولعل الحل المناسب لتحديد المعامل أو المتغير وتقدير المعلمات في وقت واحد هو استخدام أنموذج انحدار اقل انكماش مطلق لاختيار العامل (least absolutes shrinkage and selection operator) والذي يشار اليه اختصاراً بالرمز (lasso)، 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. [12] [11]

في عام 2014 ناقش الباحث (Musoro) وآخرون، أداء نموذج الجزاء (penalized model) في حالة وجود بيانات متعددة للتنبؤ بمرض الانسداد الرئوي المزمن; واستخدم الباحث (Pripp) وآخرون في عام 2017، أنموذج انحدار (lasso) لتقييم الارتباط بين العديد من المؤشرات الحيوية للالتهاب والورم الدموي المزمن; كما استخدم الباحث (Khanji) وآخرون في عام 2018 أنموذج انحدار(lasso) لبناء نماذج التنبؤ المتعلقة بأمراض القلب والأوعية الدموية. في هذا البحث سنناقش أنموذج انحدار التمهيد المستحث ذو اقل انكماش مطلق لاختيار العامل (islasso)، والذي يعتمد على فكرة التمهيد المستحث (induced smoothing) الذي قدمه (Brown & Wang) في عام 2005، للتعامل مع النماذج الإحصائية التي تمتلك دوال تقدير غير ممهدة. ويعتبر أنموذج انحدار (islasso) مكافئًا لنموذج انحدار (lasso)، كما انه قادرًا على حساب التقديرات في حالة العينات المحدودة باستخدام خوارزميات نيوتن (Newton-type algorithms) ومصفوفة التغاير (covariance matrix). [6][1]

2. أنموذج انحدار لاسو: Lasso Regression Model

أحد أنواع الانحدار الخطي الذي يستخدم الانكماش (shrinkage). والانكماش هو المكان الذي تقلصت فيه قيم البيانات باتجاه نقطة مركزية (central point). هذا النوع من الانحدار مناسب تماماً للنماذج التي تعرض مستويات عالية من التعددات الخطية (multicollinearity) أو عندما نريد اختيار أجزاء معينة من النموذج، مثل اختيار متغير / إلغاء المعلمة. [11]

في هذا الأنموذج يتم إضافة دالة جزاء (penalty) تساوي القيمة المطلقة لحجم المعاملات، وهذا النوع من التنظيم يمكن أن يؤدي إلى نماذج متفرقة مع عدد قليل من المعاملات، إذ يمكن أن تصبح بعض المعاملات صفرية ويتم إزالتها من النموذج، حيث تؤدي دوال الجزاء الأكبر إلى جعل قيم المعامل أقرب إلى الصفر، وهذا مثالي لإنتاج نماذج أبسط. وتهدف خوارزمية أنموذج انحدار لأسو الى تقليل المقدار:

$$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\,\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \qquad \ldots\ldots (1)$$

يتم تقليص بعض القيم إلى الصفر تماماً، مما ينتج عنه نموذج انحدار يسهل تفسيره. وتتحكم معلمة الضبط λ في قوة دالة الجزاء، حيث أن λ هي في الأساس كمية الانكماش، فعندما $\lambda = 0$، فأنه لا يتم حذف أي معلمة، وعند زيادة قيمة λ، يتم تعيين المزيد من المعاملات على الصفر (من الناحية النظرية عندما $\lambda = \infty$، يتم حذف جميع المعاملات)، بالتالي يزداد مقدار التحيز، أما عند نقصان قيمة λ فأن التباين يزداد.

ISSN (1681- 6870)

3. طريقة التمهيد المستحث: Induced Smoothing Method

تم تقديم أسلوب التمهيد المستحث (Induced Smoothing) من قِبل (Brown and Wang) في عام 2005، وذلك لتقدير الدوال والمعادلات غير الممهدة التي تمنع تطبيق خوارزميات التقدير والتقارب المعتادة، حيث يتم تحجيم الاضطرابات الطبيعية للمعلمة والحصول على معادلة تقديرية جديدة عن طريق حساب متوسط الدرجة غير الممهدة. من الناحية التقنية، يتم تحديد أنموذج انحدار (lasso) وفقًا لشروط (Karush-Kuhn-Tucker)، ولا تمتلك معادلة التقدير. ولنفرض أن $U(\beta)$ هي معادلة التقدير غير الممهدة العامة التي نحصل من خلالها على المقدّر $(\hat{\beta})$، وأن $v = \text{var}(\hat{\beta})$. فأن مقدر (lasso) يمكن أن يكتب:[7]

$$v^{-1/2}\left(\hat{\beta}-\beta\right)\sim v \quad \text{and} \quad f(v) \approx c\,\emptyset\,(v) + (1-c)\,\emptyset_{\in}(v) \qquad \ldots\ldots (2)$$

حيث أن $\emptyset\,(.)$ هي دالة الكثافة الاحتمالية تتبع التوزيع الطبيعي القياسي، وان $\emptyset_{\in}(.)$ تمثل دالة الكثافة الاحتمالية تتبع التوزيع الطبيعي بمتوسط صفر وتباين صغير جدا يقترب من الصفر، وان $c \in [0,1]$ هو الخليط الموزون غير المعلوم. [5]

$$\begin{aligned}\tilde{U}(\beta) &= E_v[U(\beta + v^{1/2}\,v)] \\ &= \int U(\beta + v^{1/2}\,v) f(v)\mathrm{d}v \\ &= \int U(\beta + v^{1/2}\,v)\{c\phi(v) + (1-c)\phi_{\epsilon}(v)\}\mathrm{d}v \\ &= c\int U(\beta + v^{1/2}\,v)\phi(v)\mathrm{d}v + (1-c)\int U(\beta + v^{1/2}v)\phi_{\epsilon}(v)\mathrm{d}v \end{aligned} \qquad \ldots\ldots (3)$$

وبما أن النتيجة ممهدة وغير مجزأة (unpenalized)، وإن أسلوب التمهيد المستحث فعال فقط في حالة دوال الجزاء غير الممهدة $I(\beta > 0)$. وبافتراض خليط المكونين المذكورين أعلاه، فإن دالة الجزاء الممهدة يمكن أن تكون بالصيغة التالية:

$$p(\beta, v; c) = c\left\{2\emptyset\left(\frac{\beta}{\frac{1}{v^2}}\right) - 1\right\} + (1-c)\left\{2\emptyset\left(\frac{\beta}{\frac{1}{v^2}}\right) - 1\right\} \qquad \ldots\ldots (4)$$

إذ أن $\emptyset\,(.)$ و $\emptyset_{\in}(.)$ هي دوال توزيع تراكمية تتبع التوزيع الطبيعي. وان c هو الخليط الموزون الذي يعتمد على عدة عوامل لإيجاده، بما في ذلك الإشارة الحقيقية، وتباين الخطأ، ومقياس التغاير. سيتم استعراض أدناه طريقتين بديلتين لمعالجة المشكلة مع وجود المجهول c. حيث نهدف إلى الحصول على مصطلح جزاء مستقل عن الضوضاء c. [5] [4]

1.3 أسلوب بيزين المستعار (الزائف): A pseudo Bayesian approach

يتم تحديد مقدار الخليط الموزون c ودالة التوزيع السابقة F(c)، باستخدام طريقة (Bayesian) مما يؤدي إلى الحصول على دالة متوسط الجزاء، كما مبين في الصيغة التالية:

ISSN (1681- 6870)

$$p(\beta, v) = \int_0^1 p(\beta, v; c)\, d\, F(c) \qquad \ldots\ldots (5)$$

حيث يمثل $p(\beta, v; c)$ متجه الجزاء عالي الأبعاد، ولعكس حالة عدم اليقين (uncertainty) في الخليط الموزون c، فإن متوسط الجزاء المذكور في العادلة (4) يمكن التعبير عنه بالصيغة التالية:

$$\bar{p}(\beta, v) = \sum_{k=1}^{k} p(\beta, v; c_k)\, /K \qquad \ldots\ldots (6)$$

إذ أن c_k , ... , c_1, c_2 وان k تمثل القيم المتوازنة الواقعة بين (1,0). كما نلاحظ أن دالة متوسط الجزاء مستقلة عن c، إذ تعتمد على v و β فقط. بينما نلاحظ من المعادلة (4) ونظيرتها التجريبية أنها تمتلك خصائص بيزية بشكل واضح، إذ لا تحتاج الى أي معلمات مسبقة. بالتالي يمكن التعبير عن معادلة تقدير التمهيد المستحث كما يلي:[8]

$$\bar{U}(\beta) = -\sum_i (y_i - x_i\beta)x_i + \lambda \bar{p}\,(\beta, v)$$

وبإخذ المشتقة الأولى لمعادلة تقدير التمهيد المستحث نحصل على:

$$\bar{U}'(\beta) = -\sum_i x_i^2 + \lambda \bar{p}'\,(\beta, v)$$

بالتالي فأن صيغة دالة مشتق الجزاء يمكن التعبير عنها بالصيغة التالية: [3]

$$\bar{p}'(\beta, v) = \frac{\partial}{\partial \beta}\bar{p}(\beta, v) = \bar{p}'(\beta, v) = \frac{1}{k}\sum_k \{c_k(2\emptyset(\frac{\beta}{v^{\frac{1}{2}}})/v^{\frac{1}{2}} + (1 - c_k)(\frac{2\emptyset_\epsilon\left(\frac{\beta}{v^{\frac{1}{2}}}\right)}{v^{\frac{1}{2}}})\} \ldots (7)$$

2.3 اختيار c التكيفي: Adaptive c selection

في هذه الطريقة تم اتباع نهجاً إرشادياً وذلك عن طريق حساب الخليط الموزون c كدالة من النسبة الحالية $w = \hat{\beta}^2/v$، بدلاً من أخذ معادلة متوسط الجزاء قيم محتملة من c. إذ تتمثل الفكرة في أن دالة كثافة الخليط $\emptyset_\epsilon(v)$ في المعادلة (1) يجب أن تقترب من الصفر عندما تصبح قيمة w صغيرة جداً، وبالمثل فأن دالة الكثافة المستمرة $\emptyset\ (v)$ يجب أن يقترب من w عندما تكون w كبيرة. حيث يتم تعيين قيم c المحصورة بين (0،1) في كل خطوة من خطوات الخوارزمية التكرارية.

ولنفرض أن $y = x\beta + \epsilon$ هو نموذج انحدار خطي متعدد، حيث أن $y = (y_1, \ldots., y_n)^T$ هو متجه المتغيرات المتأثرة، ويمثل الرمز x مصفوفة المتغيرات التوضيحية ذات الأبعاد $n * p$، وان $\beta = (\beta_1, \ldots., \beta_p)^T$ هو متجه معاملات الانحدار. أن المعادلة التقديرية "الزائفة" بثبات λ تكون بالصيغة التالية:[13] [4]

$$U(\beta) = -x^T(y - x\beta) + \lambda(2\, I(\beta > 0) - I_p)$$

وبتطبيق أسلوب التمهيد المستحث والمعتمد على التوزيع الخليط (mixture distribution) في حالة العينات الكبيرة للمقدر $\hat{\beta}$ نحصل على:

ISSN (1681- 6870)

$$\tilde{U}(\beta) = -x^T(y - x\beta) + \lambda p(\beta, v; c)$$

حيث يحتوي متجه الجزاء $p(\beta, v; c)$ ذي الأبعاد p على المركب العام (generic component) والذي يعبر عنه بالصيغة التالية:

$$c_j(2\emptyset(\frac{\beta_j}{v_j^{\frac{1}{2}}})/v_j^{\frac{1}{2}}) + (1 - c_j)(2\emptyset_{\in}(\frac{\beta_j}{v_j^{\frac{1}{2}}})/v_j^{\frac{1}{2}}) - 1) \qquad \ldots\ldots (8)$$

إذ يمثل الرمز v في دالة الجزاء القطر الرئيسي للمصفوفة $V = \text{var}(\hat{\beta})$، ويتضمن c الأوزان التابعة للمتغير p. أما مصفوفة الميل (slope matrix) فيمكن التعبير عنها بالصيغة التالية:

$$\tilde{U}(\beta) = x^T x + \lambda p'(\beta, v; c) \qquad \ldots\ldots (9)$$

ويسمح وجود $\tilde{U}(\beta)$ بتطبيق صيغة (sandwich formula) لحساب مصفوفة التغاير المقدرة، حيث أن:

$$V = \bar{U}(\hat{\beta})^{-1} I \bar{U}(\hat{\beta})^{-1} \qquad \ldots\ldots (10)$$

إذ أن قيمة $\hat{\beta}$ هي القيمة النهائية عند التقارب، وان ($I \propto x^T x$) هي مصفوفة المعلومات المستقلة عن $\hat{\beta}$. ومن خلال أجراء أنموذج (islasso) يتم استبدال دالة القيمة المطلقة غير الممهدة بدالة تقريبية ممهدة بالاعتماد على تقدير الخطأ القياسي، إذ أن أصغر خطأ قياسي هو الأقرب. فكلما زاد حجم العينة فأن الخطأ القياسي يقترب من الصفر، $\text{SE}(\hat{\beta}) \rightarrow 0$ و $\emptyset(\beta/\text{SE}(\hat{\beta})) \rightarrow I(\beta > 0)$

مع التأكيد على أن أنموذج انحدار (islasso) يكافئ تقريباً أنموذج (lasso). [10] [3]

4. اختبـار والـد: Wald test

يعد اختبار Wald (يُطلق عليه أيضًا اختبار Wald Chi-Squared Test) طريقة لمعرفة ما إذا كانت المتغيرات التوضيحية في النموذج مهمة أم لا. ويمكن من خلاله حذف المتغيرات التي لا تضيف شيئًا دون التأثير على النموذج. ويمكن استخدام الاختبار في العديد من النماذج المختلفة بما في ذلك النماذج ذات المتغيرات الثنائية أو المتغيرات المستمرة.

أن فرضية العدم لاختبار (Wald test) هي:

$$H_0: \beta = 0$$

إذا تم رفض فرضية العدم، فأن هذا يشير إلى معنوية الفروق، أي أنه يمكن إزالة المتغيرات المعنية دون إلحاق ضرر بالنموذج قيد البحث. ونظراً لتقدير $\hat{\beta}$ باستخدام أنموذج انحدار (islasso) وبما أن الخطأ القياسي $\text{SE}(\hat{\beta})$ المحسوب على أنه الجذر التربيعي للعناصر القطرية الرئيسية لمصفوفة التغاير المبينة في المعادلة (9)، بالتالي يمكن تعريف اختبار (Wald) بالصيغة التالية:

$$w_0 = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} \qquad \ldots\ldots (11)$$

ISSN (1681- 6870)

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) لـw_0.[12] [6]

5. وصف تجربة المحاكاة: Simulation experiments

يمكن تعريف المحاكاة بأنها عملية تمثيل أو تقليد للواقع الحقيقي باستعمال نماذج معينة، وكثيرا ما نجد في الواقع الحقيقي أن هناك عمليات تكون معقدة الفهم والتحليل لذلك فمن الأفضل أن نوصف هذه العمليات بصورة مشابهة للصور الحقيقية بنماذج معينة. في هذه التجربة تم أجراء بعض عمليات المحاكاة بطريقة مونت كارلو (Monte Carlo) لتقييم سلوك العينة المحدود باستخدام أنموذج انحدار (lasso) وكذلك أنموذج انحدار (islasso). [2]

في القسم الأول من عملية المحاكاة، نهدف إلى تقييم توزيع العينة المحدود لإحصائه (Wald) المبينة في المعادلة (8) تحت الفرضية H_0 (عدم وجود تأثير). حيث تم توليد (1000) تكرار لنموذج الانحدار الخطي لعينة بحجم (n=50)، وعدد المعلمات (p=20) في كل تكرار، كما تم تنفيذ خوارزمية أنموذج انحدار (islasso) بمعلمة ضبط (λ) يتم اختيارها باستخدام طريقة معيار العبور الشرعي ذو الخمسة أضعاف (fivefold cross-validation)، حيث يتم بناء قيم معلمة ضبط (λ) من خلال أخذ 100 قيمة ذات فواصل زمنية متساوية ضمن الفترة $[\log(\lambda_{min}), \log(\lambda_{max})]$.[9] [13]

أن القيمة العظمى (λ_{max}) هي أصغر قيمة، حيث تكون فيها جميع المعاملات صفر، ويتم حسابها من خلال الصيغة ($\max\{x^T y\}$)، في حين أن القيمة الصغرى (λ_{min}) التي تؤدي إلى تقديرات غير صفرية، ويتم حسابها على أنها ($0.0001\ \lambda_{max}$) أو ($0.01\ \lambda_{max}$) إذا كانت $n < p$.

في القسم الثاني من عملية المحاكاة، تم تقييم أداء إحصائه (Wald) الناتجة من أجراء انحدار (islasso) لاختبار فرضية العدم $H_0: \beta_j = 0$ لكل معامل β_j في معادلة الانحدار. حيث تم توليد (500) تكرار لنموذج الانحدار الخطي $y \sim N(x\beta, I_n)$ ولسيناريوهات مختلفة، إذ تم اختيار عينتين بالأحجام (n=100)، و(n=200) وعدد معلمات (p=0.5, 0.8, 1.2, 2) لكل n، وتأتي المتغيرات المشتركة p من توزيع الطبيعي المتعدد (multi-normal distribution) مع مصفوفة تغاير موحدة (identity covariance matrix)، ثم يتم تنفيذ خوارزمية أنموذج انحدار (islasso) بمعلمة الضبط (λ).[8] [7]

6. تحليل النتائج: Analysis of the results

من خلال نتائج تجارب المحاكاة تبين أن إحصائه (Wald) الناتجة من تطبيق أنموذج (islasso) مع توزيع طبيعي قياسي، هي أداة موثوقة لاختبار فرضية عدم وجود تأثير. كما نلاحظ توزيع المعاينة باستخدام إحصاءه (Wald)، الناتجة من تطبيق أنموذج (lasso) معطى بواسطة تقدير النقطة والأخطاء القياسية المحسوبة. أن القيم المختلفة من n وp (أيضًا مع $n < p$) تؤدي إلى عدم وجود فروق معنوية في توزيعات أخذ العينات، وبالتالي يتم حذف المخططات المتعلقة بها.

تم الحصول على النتائج باستعمال الحاسب الإلكتروني وذلك بالاعتماد على لغة البرمجة الإحصائية (R) كونها تعد حالياً من أفضل برامج الحوسبة الإحصائية والرسوم البيانية. وبعد تطبيق أنموذج (islasso) تم الحصول على التقديرات وقيم إحصائه (Wald)، والأخطاء المعيارية، فضلاً عن قيم (p-value)، وكما في الجدول (1) التالي:

ISSN (1681- 6870)

جدول (1): قيم اختبار Wald باستخدام أنموذج انحدار (islasso) في حالة العينات الصغيرة.

xi	Estimate	Std.Error	Df	Wald Stat.	P-value
X1	0.010326	0.009473	0.017	1.090	0.2757
X2	-0.005037	0.007297	0.023	-0.690	0.4900
X3	-0.013103	0.011566	0.020	-1.133	0.2573
X4	0.000733	0.055027	0.001	0.013	0.9894
X5	0.000069	0.008669	0.000	0.008	0.9937
X6	-0.000246	0.022142	0.000	-0.011	0.9911
X7	0.000016	0.006332	0.000	0.002	0.9980
X8	-0.007380	0.007934	0.017	-0.930	0.3523
X9	0.000641	0.057478	0.000	0.011	0.9911
X10	0.000105	0.011737	0.000	0.009	0.9929
X11	0.005505	0.007832	0.022	0.703	0.4821
X12	-0.000234	0.024492	0.000	-0.010	0.9924
X13	0.004349	0.008509	0.040	0.511	0.6093
X14	0.005657	0.007398	0.020	0.765	0.4445
X15	0.007719	0.007908	0.016	0.976	0.3291

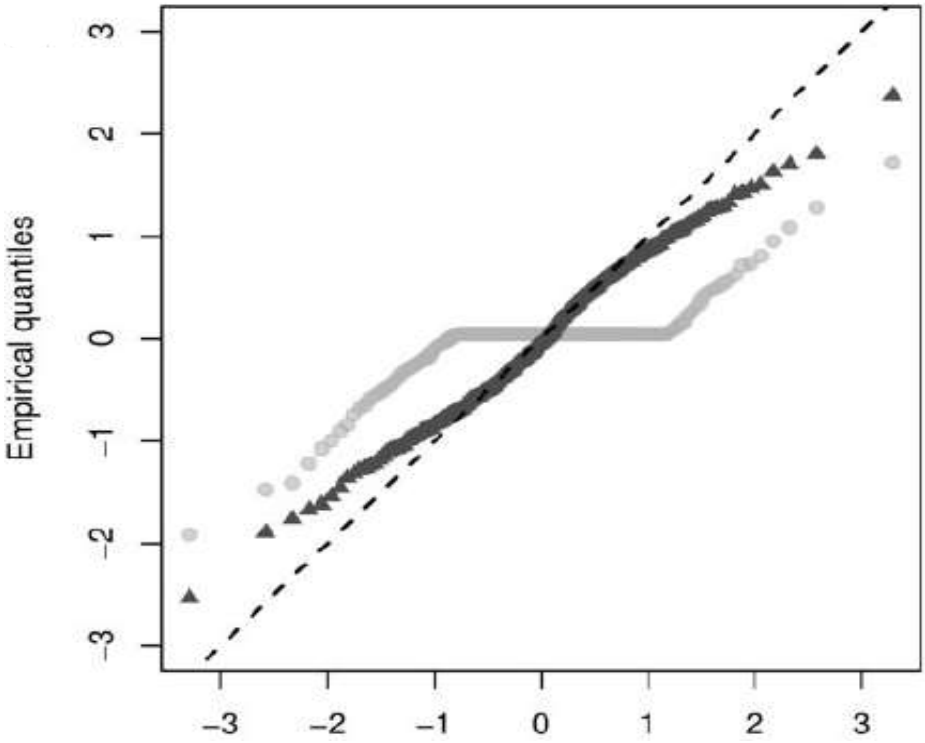
Dispersion parameter: 33.315

Residual deviance : 1575.4
AIC: 321.83
Lambda: 103.35

نلاحظ من الجدول (1) أن جميع قيم اختبار (Wald) كانت غير معنوية عند مستوى معنوية (0.05) بما يدل على عدم وجود فروق معنوية، بالتالي قبول فرضية العدم H_0 (عدم وجود تأثير). كما نلاحظ أن قيمة مقياس الجودة النسبية (AIC) قد بلغت (321.83)، وبلغت قيمة معلمة التشتت (33.315)، وبلغت قيمة الانحراف المتبقي (1575.4)، أما قيمة معلمة الضبط (103.35)، ويوضح الشكل (1) التالي توزيع المعاينة لإحصائه (Wald) المبينة في المعادلة (8)، إذ تم تصوير توزيع أخذ العينات باستخدام أنموذج (lasso) لإحصائه (Wald)، والذي يعطى بنسبة تقدير النقاط والأخطاء المعيارية.

ISSN (1681- 6870)

يوضح الشكل (1) التالي كيف توفر إحصائه (Wald) الناتجة من تطبيق أنموذج انحدار (islasso) مع توزيع طبيعي قياسي أداة موثوقة لاختبار فرضية عدم وجود تأثير. أن الخط ذي الدوائر الرمادية يوضح إحصاءات (Wald) لأنموذج انحدار (lasso)، ويوضح الخط ذي المثلثات السوداء إحصاءات (Wald) لأنموذج انحدار(islasso). أن الخطوط المتقطعة تمثل التوزيع الطبيعي القياسي. وفي كل تكرار، تم الحصول على معلمة الضبط (λ) الأمثل من خلال طريقة معيار العبور الشرعي ذو الخمسة أضعاف (fivefold cross-validation).



شكل (1) نماذج انحدار (lasso) و (islasso)، لعينة بحجم n=50

نلاحظ من الجدول (2) التالي انه عند زيادة حجم العينة (n=100)، فأن جميع قيم (.Wald Stat) كانت معنوية عند مستوى معنوية (0.05)، بالتالي رفض فرضية العدم H_0 (أي وجود تأثير معنوي) كما نلاحظ أن قيمة مقياس الجودة النسبية (AIC) قد بلغت (282.82)، وبلغت قيمة معلمة التشتت (0.776)، وهي قيمة صغيرة جدا إذ ما تمت مقارنتها بمعلمة التشتت في حالة العينات الصغيرة، وبلغت قيمة الانحراف المتبقي (25.048)، بما يعطي أفضلية لنماذج الانحدار (lasso) و(islasso) في حالة العينات الكبيرة، أما قيمة معلمة الضبط (3.221).

جدول (2) قيم اختبار Wald باستخدام أنموذج انحدار (islasso) في حالة العينات الكبيرة.

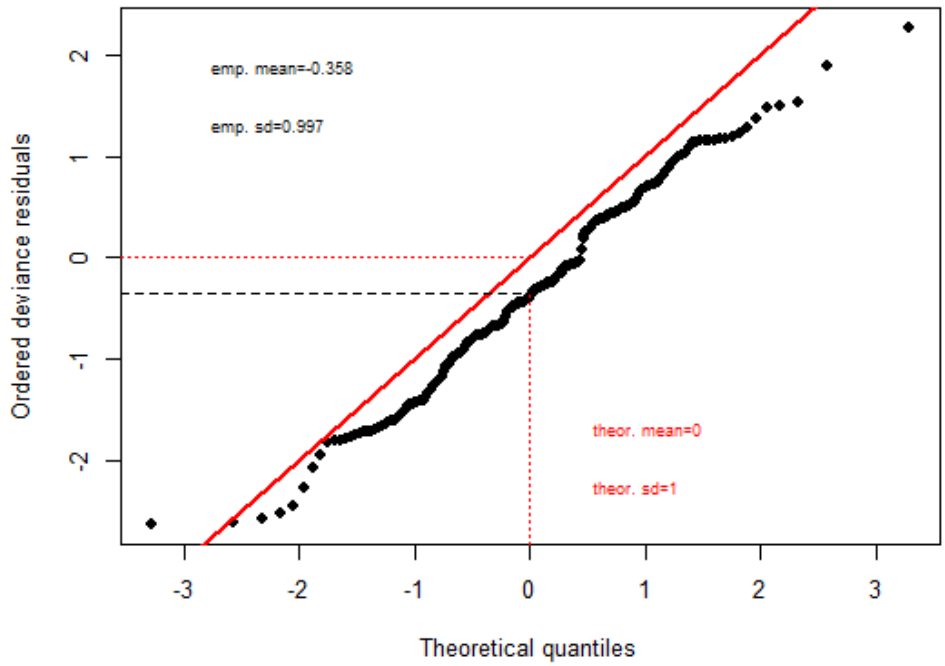
xi	Estimate	Std.Error	Df	Wald Stat.	P-value
X1	-2.186376	0.149669	1.000	-14.61	< 2e-16 ***
X2	-1.619296	0.174946	1.000	-9.256	< 2e-16 ***
X3	-1.749319	0.147400	1.000	-11.87	< 2e-16 ***
X4	-1.348380	0.163548	1.000	-8.245	< 2e-16 ***
X5	-1.667539	0.176731	1.000	-9.435	< 2e-16 ***
X6	-1.302098	0.131480	1.000	-9.903	< 2e-16 ***
X7	-0.775307	0.150950	1.000	-5.136	2.8e-07 ***
X8	-1.172093	0.136702	1.000	-8.574	< 2e-16 ***
X9	-1.047668	0.158984	1.000	-6.590	4.4e-11 ***
X10	-0.821793	0.139023	1.000	-5.911	3.4e-09 ***
X11	0.421064	0.114865	1.000	3.666	0.00025 ***
X12	1.054584	0.187329	1.000	5.630	1.8e-08 ***
X13	0.936192	0.138704	1.000	6.750	1.5e-11 ***
X14	1.210470	0.155113	1.000	7.804	6.1e-15 ***
X15	1.644972	0.185565	1.000	8.865	< 2e-16 ***

Dispersion parameter: 0.776

Residual deviance : 25.048
AIC: 282.82
Lambda: 3.2213

ويبين الشكل (2) التالي نتائج أجراء أنموذج انحدار (islasso) عندما (n=100)، لتوزيع المعاينة المحدود لإحصائه (Wald). كما يوضح الشكل (2) كيف توفر إحصائه (Wald) الناتجة من تطبيق أنموذج انحدار (islasso) مع توزيع طبيعي قياسي أداة موثوقة لاختبار فرضية العدم وجود تأثير معنوي. أن الخط ذي المثلثات السوداء إحصاءات (Wald) للنماذج (islasso, lasso). أن الخطوط المتقطعة تمثل التوزيع الطبيعي القياسي. وفي كل تكرار، تم الحصول على معلمة الضبط (λ) الأمثل من خلال طريقة معيار العبور الشرعي ذو الخمسة أضعاف (fivefold cross-validation).

ISSN (1681- 6870)



شكل (2) نماذج انحدار (lasso) و (islasso)، لعينة بحجم n=100

7. الاستنتاجات: Conclusions

في هذا البحث، تم تطبيق أنموذج اقل انكماش مطلق لاختيار العامل (lasso) وأنموذج التمهيد المستحث ذو اقل انكماش مطلق لاختيار العامل (islasso)، كما تم اختبار الفرضيات الخاصة بمعاملات معادلة الانحدار لكل أنموذج باستخدام اختبار (Wald)، حيث يعتبر أنموذج (islasso) إطار جديد لنماذج الانحدار مع دوال الشرائح.

تم توظيف طريقة التمهيد المستحث (الجديدة نسبياً) بنجاح في بعض السياقات للتعامل مع معادلات التقدير غير الممهدة، إذ أن تطبيق طريقة التمهيد المستحث في أنموذج انحدار (lasso) يؤدي إلى استبدال دالة الجزاء بنظرائها السلسة حيث يتم ضبط "عرض الحزمة"، أي مقدار التنعيم الذي يعمل على كل معامل بواسطة الخطأ المعياري المقابل الذي يتم حسابه من البيانات.

كما نلاحظ من خلال الجانب التجريبي أن كلما زاد حجم العينة وانخفض الخطأ المعياري، فإن أنموذج انحدار (islasso) يقترب من أنموذج انحدار (lasso)، مما يجعل أنموذج انحدار (islasso) مكافئًا لأنموذج انحدار (lasso).

References

[1] Brown B and Wang Y. Standard errors and covariance matrices for smoothed rank estimators. Biometrika, 2005; 92:149–158.

[2] Diane, I. G & Lomnie, C. V, (1981) " A Simulation Study Of estimators for the 2-Parameter Weibull Distribution ", IEEE Transaction on Reliability, Vol. R-30, No. 1.

[3] Fahrmeir L, Kneib T, Lang S, et al. Regression: models, methods, and applications. Berlin: Springer, 2013.

[4] Friedman J, Hastie T, and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010; 33: 1–22.

[5] Knight K and Fu W. Asymptotics for lasso-type estimators. Ann Stat 2000; 28: 1356–1378.

[6] Meinshausen N and Bu¨ hlmann P. High-dimensional graphs and variable selection with the lasso. Ann Stat 2006; 34: 1436–1462.

[7] Muggeo VMR. Interval estimation for the breakpoint in segmented regression: a smoothed score-based approach. Aust NZ J Stat 2017; 59: 311–322.

[8] Owen AB. A robust hybrid of lasso and ridge regression. Contemporary Math 2007; 443: 59–72.

[9] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Series B 1996; 58: 267–288.

[10] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc: Ser B 2011; 73: 273–282.

[11] Tutz G and Gertheiss J. Regularized regression for categorical data (with discussion). Stat Model 2016; 16: 161–200.

[12] Wang Y, Fu L, and Bai Z. Rank regression for analysis of clustered data: a natural induced smoothing approach. Compute Stat Data Anal 2010; 54: 1036–1050.

[13] Zou H, Hastie T, and Tibshirani R. On the ‘degrees of freedom’ of the lasso. Ann Stat 2007; 35: 2173-2192.

ISSN (1681- 6870)

R CODE

```
library(Matrix)
library(glmnet)
library(islasso)

# Simulation (1)
set.seed(1000)
n <- 50
p <- 100
p1 <- 20  #number of nonzero coefficients
coef.ver <- sort(round(c(seq(.5, 3, l=p1/2)
            , seq(-1, -2, l=p1/2)), 2))

sigma <- 1
coef <- c(coef.ver, rep(0, p-p1))
X <- matrix(rnorm(n*p), n, p)
eta <- drop(X%*%coef)
mu <- eta
y <- mu + rnorm(n, 0, sigma)
zls <- islasso(y~-1+X, family=gaussian)

summary(zls)
logLik(zls)
predict(zls, type="response")
plot(zls)

# Simulation (2)
set.seed(1000)
n <- 100
p <- 100
p1 <- 20  #number of nonzero coefficients
coef.ver <- sort(round(c(seq(.5, 3, l=p1/2)
            , seq(-1, -2, l=p1/2)), 2))

sigma <- 1
coef <- c(coef.ver, rep(0, p-p1))
X <- matrix(rnorm(n*p), n, p)
eta <- drop(X%*%coef)
mu <- eta
y <- mu + rnorm(n, 0, sigma)
zls <- islasso(y~-1+X, family=gaussian)

summary(zls)
logLik(zls)
predict(zls, type="response")
plot(zls)

q()   # ----- end ------
```