مقارنة بين بعض المقدرات الجزائية الحصينة باستخدام المحاكاة

أ.م.د. عماد حازم عبودي / جامعة بغداد / كلية الإدارة والاقتصاد
م.م. علي حميد يوسف/ جامعة واسط / كلية الإدارة والاقتصاد

تاريخ التقديم:2017/3/5
تاريخ القبول:2017/4/16

المستخلص:

تعد طريقة المربعات الصغرى الجزائية طريقة ملائمة وشائعة للتعامل مع البيانات ذات الابعاد العالية ولاسيما التي يكون فيها عدد المتغيرات التوضيحية اكبر من حجم العينة ، ومن ضمن المزايا التي تتمتع بها طريقة المربعات الصغرى الجزائية هي ضمان الحصول على تنبؤ عالي الدقة وكذلك قيامها بعملية التقدير واختيار المتغيرات في ان واحد ، فهي تقوم بتقليص بعض المعاملات وجعلها مساوية للصفر . حيث انها تعطي نموذجاً متبعثراً (Sparse Model) اي النموذج الذي يتضمن اقل عدد ممكن من المتغيرات ومن ثم يكون قابلاً للتفسير بسهولة. وعلى الرغم من تلك المزايا التي تتمتع بها طريقة المربعات الصغرى الجزائية الا انها تعد طريقة غير حصينة بمعنى انها تتأثر بالقيم الشاردة ، وللتغلب على هذه المشكلة يتم استبدال دالة خسارة المربعات الصغرى بدالة خسارة حصينة ليتم الحصول على طريقة المربعات الصغرى الجزائية الحصينة ، ويكون المقدر الناتج يدعى بالمقدر الجزائي الحصين الذي يتعامل مع مشكلتي الابعاد والقيم الشاردة .

وفي هذا البحث تمت عملية المقارنة بين مقدري (Sparse LTS) و(MM Lasso) باستعمال المحاكاة وقد تم التوصل الى افضلية مقدر (MM Lasso) في معظم التجارب وذلك بالاعتماد على معيار متوسط مربعات الخطأ ، ومعدل الايجابية الزائف ومعدل السلبية الزائف.

المصطلحات الرئيسة في البحث/ المربعات الصغرى الجزائية ، Lasso، LTS ، MM

مجلة العلوم
الاقتصادية والإدارية
العدد 100 المجلد 23
الصفحات 490ـ 504

*البحث مستل من اطروحة دكتوراه

1- المقدمة

زاد الاهتمام بموضوع تحليل البيانات ذات الابعاد العالية في السنوات الأخيرة ولاسيما التي يكون فيها عدد المتغيرات التوضيحية اكبر من حجم العينة ، فهي اجتذبت العديد من الباحثين وضمن مجالات مختلفة منها الرياضيات التطبيقية ، الهندسة الالكترونية والجينات الوراثية وغيرها .

ان نماذج الانحدار الخطية التي تتضمن عدداً كبيراً من المتغيرات التوضيحية تكون ذات أداء ضعيف وذلك بسبب كبر التباين فضلاً عن ذلك فإنها تكون صعبة التفسير . وفي العديد من الدراسات يكون فيها عدد المتغيرات التوضيحية اكبر من حجم العينة $(p > n)$ الأمر الذي يؤدي إلى فشل النموذج التقليدي في التعامل مع البيانات عالية الإبعاد .

ان إحدى المسائل المهمة في الإحصاء هي اختيار المتغيرات في الانحدار ، ومع الزيادة في عدد المتغيرات التوضيحية مع صغر حجم العينة يصبح هناك تحدي رئيس في عملية تقدير المعلمة واختيار المتغير في النموذج .

ففي نموذج الانحدار التقليدي فأن أسلوب اختيار المتغيرات (الاختبارات الأمامية ، الاختبارات الخلفية، الانحدار المتدرج ..الخ) تكون غير ملائمة في التعامل مع البيانات عالية الإبعاد .

وعلاوة على ذلك فان طريقة المربعات الصغرى الاعتيادية (Ordinary Least Square) والتي هي شائعة الاستخدام في الانحدار تكون غير ملائمة في التعامل مع البيانات عالية الأبعاد ، حيث لا يمكن ان تكون مصفوفة المعلومات ذات رتبة كاملة الأمر الذي يؤدي الى عدم الحصول على حل وحيد .

ان الاسلوب الشائع للتعامل مع البيانات ذات الابعاد العالية هو طريقة المربعات الصغرى الجزائية والتي تستند الى مبدأ تصغير مجموع مربعات الخطأ وفقاً لقيد معين على المعلمات .

ان من ضمن المزايا التي تتمتع بها طريقة المربعات الصغرى الجزائية هي ضمان الحصول على تنبؤ عالي الدقة وكذلك قيامها بعملية التقدير واختيار المتغيرات في ان واحد ، حيث تقوم بتقليص بعض المعاملات وجعل الاخرى مساوية للصفر . حيث انها تعطي نموذجاً متبعثراً (Sparse Model) أي النموذج الذي يتضمن اقل عدد ممكن من المتغيرات ومن ثم يكون قابلاً للتفسير بسهولة.

كما ان طريقة المربعات الصغرى الجزائية هي غير حصينة بمعنى تتأثر بالقيم الشاذة ، وللتغلب على هذه المشكلة يتم استبدال دالة خسارة المربعات الصغرى الجزائية بدالة خسارة حصينة لنحصل على طريقة المربعات الصغرى الجزائية الحصينة ، ويكون المقدر الناتج يدعى بالمقدر الجزائي الحصين الذي يستطيع التعامل مع مشكلتي الابعاد والقيم الشاذة .

2- هدف البحث

يهدف هذا البحث الى المقارنة بين المقدرات الجزائية الحصينة لنموذج الانحدار الخطي في ظل وجود مشكلتي الابعاد والقيم الشاذة والحصول على افضل مقدر من بين المقدرات الاخرى ومن ثم الحصول على افضل تقدير باستعمال المحاكاة وذلك بالاعتماد على معيار متوسط مربعات الخطأ ، معدل الايجابية الزائف ومعدل السلبية الزائف.

3- الجانب النظري

3-1 الانحدار الخطي وطريقة المربعات الصغرى[10][9][5]

(Linear Regression and Least Square Method)

يعد الانحدار احد التقنيات الإحصائية المستعملة على نطاق واسع في مختلف العلوم لتحديد العلاقة الخطية بين متغيرين او اكثر من المتغيرات بحيث انه من الممكن ان يتم التنبؤ بأحد المتغيرات عن طريق الاخر . حيث المتغير المراد التنبؤ به يدعى بمتغير الاستجابة (Response Variable) او المتغير المعتمد (Dependent Variable) . اما المتغير التنبؤي يدعى بالمتغير التوضيحي (explanatory Variable) او المتغير المستقل (Independent Variable) .

ان نموذج الانحدار الخطي العام يكتب وفق الصيغة الاتية :-

$$Y = X\beta + \varepsilon \qquad (1)$$

حيث ان :-

Y: متجه المتغير التابع من الدرجة **(nx1)**

X : مصفوفة المتغيرات التوضيحية من الدرجة **(nxp)**

ε : متجه حد الخطأ العشوائي من الدرجة **(nx1)** ، وحيث ان موجه الأخطاء يفترض ان يتوزع توزيعاً طبيعياً بمتوسط صفر وتباين ثابت $\sigma^2 I_n$ بمعنى اخر ممكن كتابته كالاتي :-

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

ان مقدر المربعات الصغرى يكون وفقاً للصيغة الاتية :-

$$\hat{\beta}_{LS} = (\acute{X}X)^{-1}\acute{X}Y \qquad (2)$$

اما مصفوفة التباين والتباين المشترك للمقدر $(\hat{\beta}_{LS})$ تعطى بالصيغة الاتية :-

$$Var - Cov(\hat{\beta}_{LS}) = \sigma^2(\acute{X}X)^{-1} \qquad (3)$$

ان موجه معلمات نموذج الانحدار الخطي المقدر بطريقة المربعات الصغرى **(OLS)** يمتلك أفضل تقدير خطي غير متحيز **(Best Linear Unbiased Estimator)** .

ان من ضمن شرط تطبيق المربعات الصغرى الاعتيادية ان لا يكون هناك تعدد ارتباط عالي بين المتغيرات التوضيحية (التعدد الخطي) ، كما ينبغي ان يكون عدد المشاهدات اكبر من عدد المعلمات المطلوب تقديرها $p < n$ ، وهذا يعني ان رتبة المصفوفة X في النموذج (1) يجب ان تكون اقل من عدد المشاهدات اي ان :-

$$rank(X) = p < n$$

2-3 نموذج الانحدار الخطي ذو الابعاد الكبيرة[11][4]

High-dimensional linear regression model

ان نموذج الانحدار الخطي (1) يكون ذو ابعاد عالية **(High-dimensional)** اذا كان عدد المتغيرات التوضيحية اكبر من حجم العينة $(p > n)$. اما اذا كان عدد المتغيرات التوضيحية اقل من حجم العينة $(p < n)$ فأن الأنموذج يدعى بأنه ذو ابعاد قليلة **(low dimension)**. وفي كلا الحالتين الإبعاد العالية والقليلة نحن نرغب في تحقيق الأهداف الاتية :-

1- التقدير **Estimation**
2- التنبؤ **Prediction**
3- اختيار المتغير **Variable Selection**

ان مقدر المربعات الصغرى $\hat{\beta}_{LS}$ الذي نحصل عليه من خلال مجموع مربعات الخطأ ، والحصول على المعادلة الطبيعية الاتية :-

$$\acute{X}Y = \acute{X}X\beta_{ls}$$

وبحل المعادلة المذكور آنفاً نحصل على **(nxp)** من أنظمة المعادلات ل $\hat{\beta}_{LS}$.

وطالما معكوس $\acute{X}X$ يكون موجود ، فأنه يوجد حل وحيد للنظام . ولكن هذه ليست الحالة العامة للانموذج عالي الابعاد . فأذا كان **(p>n)** فأن المشكلة تتجلى بعدم وجود حل وحيد .

وبكلام أخر فأن البيانات ذات الابعاد العالية لا يمكن ان تعامل بالأسلوب نفسه الذي يتم بالنسبة للبيانات ذات الابعاد القليلة .

كما ان الطرائق التقليدية في اختيار المتغيرات كطريقة (All Possible Regression) ، (Forward Regression) ، (Backward Regression) و (Stepwise Regression) . إلى غيرها من هذه الطرائق لا يمكن ان تستعمل في حالة البيانات ذات الابعاد العالية.
مما تقدم يتضح ان هذه الأساليب ليست ذات جدوى أمام الزيادة في عدد المتغيرات التوضيحية وبهذا نكون بحاجة إلى أسلوب بديل للمربعات الصغرى الاعتيادية.

3-3 طريقة المربعات الصغرى الجزائية [8][7]

Penalized Least Square Method

تعد طريقة المربعات الصغرى الجزائية طريقة ملائمة وشائعة للتعامل مع البيانات عالية الابعاد ، اي التي يكون فيها عدد المتغيرات التوضيحية اكبر من حجم العينة ، حيث انه لا يمكن في هذه الحالة استخدام طريقة المربعات الصغرى الاعتيادية .
ان طريقة المربعات الصغرى الجزائية تستند الى مبدأ تصغير مجموع مربعات الخطأ مع بعض القيود على المعلمات ، حيث يتم الحصول على تقديرات المربعات الصغرى الجزائية من خلال تقليل دالة الهدف (Object Function) والتي تتألف من جزئين هما دالة الخسارة (loss function) ودالة الجزاء (penalty function) والتي تكون وفقاً للصيغة الاتية :-

$$p_{ls}(\lambda,\beta) = (y - X\beta)'(y - X\beta) + n\sum_{j=1}^{p} p_{\lambda}(|\beta_j|) \qquad (3)$$

حيث ان:-
$p(\cdot)$: تمثل دالة الجزاء (penalty function).
λ : تمثل معلمة الجزاء (penalty parameter)
عليه فأن المقدر الجزائي يتم الحصول عليه وفقاً للصيغة الاتية :-

$$\hat{\beta} = \operatorname{argmin}\{p_{ls}(\lambda,\beta)\} \qquad (4)$$

ان طريقة المربعات الصغرى الجزائية تقوم بعملية التقدير واختيار المتغيرات في ان واحد .
وبالمقارنة مع الطريقة التقليدية في اختيار المجاميع الفرعية (subset selection) ، فالانحدار الجزائي له ثلاثة مزايا منها:-
1) عدد المجاميع الفرعية الممكنة (possible subsets) يتزايد اسياً مع p ، والتي لا يمكن تحقيقها حسابياً لاختيار المجموعات الفرعية (subset selection) .
2) الانحدار الجزائي له مقدر أكثر استقرارية (stable estimator) .
3) المقدر الجزائي ينجز الاختيار والتقدير في ان واحد ، في حين اختيار المجاميع الفرعية (subset selection) هو اجراء من خطوتين ، حيث الأخطاء في الخطوة الأولى ممكن ان تضخم في الخطوة الثانية .
ان دالة الجزاء الجيدة يجب ان تؤدي إلى ان مقدر المربعات الصغرى الجزائية يكون له ثلاث من الخصائص والتي تم اقتراحها من قبل (Fan &Li) [7] وهي :-
1) عدم التحيز (Unbiasedness) : مقدر المربعات الصغرى الجزائية يكون غير متحيز تقريباً عندما تكون المعلمة الحقيقية المجهولة كبيرة لتجنب تحيز النمذجة غير الضروري .
2) التبعثر (Sparsity) : مقدر المربعات الصغرى الجزائية يكون قاعدة مستوى العتبة والتي تضع التقديرات ذات المعاملات الصغيرة إلى الصفر .
3) الاستمرارية (Continuity) : مقدر المربعات الصغرى الجزائية يجب ان يكون دالة مستمرة في البيانات لتجنب عدم الاستقرارية في تنبؤ النموذج .
كما ذكر (Fan&Li) بأن المقدر من الناحية المثالية يتمتع بخصائص الاوراكل (Oracle Properties) والتي تعني :-

1) احتمال تمييز (identifying) النموذج الحقيقي يكون واحد عندما ($n \to \infty$) . ان هذه الخاصية تدعى بخاصية التبعثر (Sparsity) .
2) المقدر يمتلك توزيع طبيعي محاذي (asymptotical normal) .

1-3-3 مقدر Lasso [1][2][8][14]

اقترح (Tibshirani) عام 1996 دالة جزاء لنموذج الانحدار الخطي تعرف ب Lasso وهي مختصر ل (Least Absolute Shrinkage and Selection Operator) لتقدير معلمات نموذج الانحدار الخطي وإجراء اختيار المتغير (Variable Selection) بشكل اني .
ان مبدأ هذه الطريقة هو تصغير مجموع مربعات البواقي وفقآ إلى قيد يمثل المجموع المطلق للمعاملات والتي تكون اصغر من ثابت معين . فلنموذج الانحدار الخطي (1) ، فأن مقدر Lasso ، $\hat{\beta}_{Lasso} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ يتم الحصول عليه وفق الصيغة التالية :-

$$\hat{\beta}_{Lasso} = \operatorname{argmin}_{\beta}\left\{\sum_{i=1}^{n}(y_i - \acute{x}_i\beta)^2\right\} \ \ s.t \sum_{i=1}^{p}|\beta_j| \le t \qquad (5)$$

t : تمثل معلمة التناغم (Tuning Parameter) وان $t \geq 0$. ومن الممكن التعبير عن الصيغة (5) بالصيغة التالية والتي تكون مكافئها لها وكالاتي:-

$$\hat{\beta}_{Lasso} = \operatorname{argmin}_{\beta}\left\{\sum_{i=1}^{n}(y_i - \acute{x}_i\beta)^2\right\} + n\lambda\sum_{j=1}^{p}|\beta_j| \qquad (6)$$

حيث ان :-
λ : تمثل معلمة الجزاء (Penalty Parameter) و تدعى بمعلمة الضبط (Regularization Parameter) .
$\lambda\sum_{j=1}^{p}|\beta_j|$: تدعى بدالة الجزاء (Penalty Function) وتدعى ايضاً بدالة الضبط (Regularization Function) ويرمز لها بالرمز L_1 norm
ان الجزاء L_1 norm يكون singular عند الأصل (origin) ، و Lasso تعد أكثر ملاءمة من حيث اختيار المتغير لاحتفاظها بخصائص جيدة حيث يتم من خلالها وجعل بعض معاملات الانحدار مساوية للصغر وتقليص الأخرى بمقدار معين مع التقليل من دالة الخسارة ومن خلال ذلك يعني ان تقديرات Lasso يمكن ان تنتج مجموعة مبعثرة (Sparse Set) من معاملات الانحداروبذلك تعطينا نموذج أكثر تفسيراً .
ان مقدر انحدار الجزاء Lasso ممكن ان يختار المتغيرات ويحسن من ملاءمة نموذج الانحدار والتي لا يكون لها حل شامل لجميع أنواع البيانات .
ان مقدر Lasso في حالة كون المصفوفة X متعامدة (Orthogonal) اي ان $\acute{X}X = I$ يكون وفقاً للصيغة الاتية :-

$$\hat{\beta}_{Lasso} = \operatorname{sign}(\hat{\beta}_j^{o})\left(|\hat{\beta}_j^{o}| - \lambda\right)_+ \quad , \qquad j = 1,2,..,p \qquad (7)$$

حيث ان :-
+: تشير الى الجزء الموجب داخل القوسين .
$\hat{\beta}_j^{o}$: مقدر اولي

ان دالة الجزاء $\sum_{j=1}^{p}|\beta_j|$ ممكن ان تقرب الى $\sum_{j=1}^{p}\frac{\beta_j^2}{|\beta_j|}$ ، وعليه فأن الحل الى Lasso ممكن ان يقرب الى انحدار الحرف ، ومن خلال استعمال الحل التكراري يمكن حل على مقدر Lasso التقريبي والذي يكون وفقاً للصيغة الاتية :-

$$\hat{\beta}^{(i+1)}{}_{lasso} = \left(\acute{X}X + \lambda\,\Lambda^{(i)}\right)^{-1}\left(\acute{X}y\right) \qquad (8)$$

حيث ان :-

$\Lambda^{(i)}$: تمثل المعكوس للمصفوفة القطرية التالية :-

$$diag\left\{\left|\beta_{lasso,1}^{(i)}\right|, \left|\beta_{lasso,2}^{(i)}\right|, \dots, \left|\beta_{lasso,p}^{(i)}\right|\right\}$$

4-3 المقدرات الجزائية الحصينة

ان طرائق الانحدار الحصينة تكون بديلة عن طريقة المربعات الصغرى عندما تكون بعض الفرضيات الاساسية منتهكة كأن يكون توزيع البواقي غير طبيعي ، كون الاخطاء ذات ذيل ثقيل او هناك بعض القيم الشاذة تؤثر في النموذج . حيث ان الطرائق الحصينة ستكون ملائمة لمعالجة هذه المشكلة وان المقدرات الناتجة عن هذه الطرائق تدعى بالمقدرات الحصينة والتي تكون غير متأثرة بالقيم الشاذة.

ان خصائص الكفاءة، ونقطة الانهيار ونقاط الرفع العالية تستعمل لتعريف تقنية اداء الانحدار الحصين في المعنى النظري .

ان احدى أهداف المقدرات الحصينة هو الحصول على نقطة انهيار عالية $(\varepsilon_n{}^*)$ والتي عرفت من قبل (Donoho & Huber 1983) . ان نقطة الانهيار ممكن ان تعرف بأنها نقطة او النسبة المئوية المحددة من التلويث في البيانات في اول اختبار إحصائي يصبح (swamped) . وعليه فأن نقطة الانهيار تكون ببساطة النقطة الأولية في اي اختبار إحصائي تصبح (swamped) بسبب تلويث البيانات .

بعض مقدرات الانحدار لديها اقل نقطة انهيار ممكنة $\left(\frac{1}{n} \text{ or } \frac{0}{n}\right)$ وبمعنى اخر ان قيمة واحدة من الشواذ ستؤدي إلى ان تكون معادلة الانحدار عديمة الفائدة ، كما توجد مقدرات أخرى لديها أعلى نقطة انهيار ممكنة $\left(\frac{n}{2} \text{ or } 50\%\right)$.

فاذا كان أسلوب التقدير الحصين لديه نقطة انهيار(50%) فأن (50%) من البيانات يمكن ان تحتوي على الشواذ وستبقى المعاملات قابلة للاستعمال ان مقدرات المربعات الصغرى الجزائية ليست حصينة بمعنى تتأثر بالقيم الشاذة ، وللتغلب على هذه المشكلة يتم استبدال دالة خسارة المربعات الصغرى الجزائية بدالة خسارة حصينة .

ان الصيغة العامة لطريقة المربعات الصغرى الجزائية الحصينة تكون وفقاً للصيغة الاتية :-

$$\sum_{i=1}^{n}\rho(y_i - x_i') + n\sum_{j=1}^{p}p_\lambda(|\beta_j|) \qquad (9)$$

حيث ان:-

$\rho(.)$:تمثل دالة الخسارة الحصينة العامة كأن تكون دالة خسارة (MM) او (LTS) وغيرها.

$p_\lambda(.)$: تمثل دالة الجزاء والتي سبق وان تم تعريفها .

ومن خلال اشتقاق المعادلة (9) بالنسبة الى β ومساواتها للصفر نحصل على المقدرات الجزائية الحصينة والتي تكون ذات اداء افضل من مقدرات المربعات الصغرى الجزائية في حالة وجود نسبة من الشواذ في البيانات

1-4-3 مقدر Sparse LTS [3][12]

اقترح (Rousseew) عام (1984)طريقة المربعات الصغرى المشذبة هي واحدة من من أكثر الطرائق الحصينة شيوعاً في تقدير معلمات نموذج الانحدار الخطي هي توسيع من المتوسط المشذب وان المقدر الناتج من هذه الطريقة يدعى بمقدر المربعات الصغرى المشذبة ويرمز له بالرمز (LTS) .

ان متجه البواقي المربعة يكون كالاتي :-

$$r^2(\beta) = (r_1^2, r_2^2, \ldots, r_n^2)' \;;\; r_i^2 = (y_i - x_i'\beta)^2 \,, i = 1,2,..,n$$

ويتم حساب مقدر (LTS) يتم حسابة وفقاً للصيغة الاتية :-

$$\hat{\beta}_{LTS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{h} \left(r^2(\beta)\right) i{:}n \qquad (10)$$

h : ثابت ومداه $\left(\frac{n}{2} < h < n\right)$

كما ان :-

$$\left(r^2(\beta)\right)1{:}n \le \left(r^2(\beta)\right)2{:}n \le \cdots \le \left(r^2(\beta)\right)n{:}n$$

ان طريقة المربعات الصغرى المشذبة (LTS)تفشل في التعامل مع مشكلة الابعاد ، اي لا يمكن تطبيقها في حالة كون عدد المتغيرات التوضيحية اكبر من حجم العينة لذلك اقترح الباحثون (Alfons et al) عام 2013 إضافة حد الجزاء Lasso من النوع (L1-Penalized) إلى دالة الهدف ليتم الحصول على طريقة المربعات الصغرى المشذبة المتبعثرة (Sparse LTS) وهي طريقة فعالة للتعامل مع مشكلة الإبعاد والقيم الشاذة وان المقدر الناتج عن هذه الطريقة يدعى بمقدر (Sparse LTS) ويرمز له بالرمز (SLTS)

ان مقدر (SLTS) يتم الحصول عليه على وفق الصيغة الاتية :-

$$\hat{\beta}_{SparseLTS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{h} \left(r^2(\beta)\right) i{:}n + h\lambda \sum_{j=1}^{p} |\beta_j| \qquad (11)$$

ان مقدر Sparse LTS هو مقدر حصين ومشابه الى مقدر Lasso حيث انه يعمل على :-

1) تحسين أداء التنبؤ من خلال تقليل التباين اذا كان حجم العينة صغير نسبة إلى عدد المتغيرات التوضيحية .

2) يتضمن تفسير عالي بسبب اختيار النموذج بشكل اني .

3) يتجنب المشاكل الحسابية لطرق الانحدار التقليدية الحصينة في حالة البيانات عالية الابعاد .

اما الخوارزمية لحساب مقدر Sparse LTS فتوصف كالاتي :-

بثبات معلمة الجزاء λ ، يتم تعريف دالة الهدف ووفقاً للصيغة الاتية :-

$$Q(H,\beta) = \sum_{i\in H} (y_i - \acute{x}_i\beta)^2 + h\lambda \sum_{j=1}^{p} |\beta_j| \qquad (12)$$

والذي يكون فيه مجموع مربعات البواقي الجزائية L_1 penalized تعتمد على العينة الفرعية $H \subseteq \{1, \ldots, n\}$

كما ان حجم عل عينة جزئية هو h اي ان :-

$$|H| = h$$

ويتم حساب مقدر Lasso لكل عينة فرعية H ووفقاً للصيغة الاتية :-

$$\hat{\beta}_H = \underset{\beta}{\operatorname{argmin}}\, Q(H,\beta) \qquad (13)$$

بعد ذلك يتم حساب مقدر Sparse LTS والذي يرمز لـه بـالرمز $\hat{\beta}_{Hopt}$ وبالاعتماد على العينة الفرعية الجديدة H_{opt} ووفقاً للمعادلة الاتية :-

$$H_{opt} = \underset{H\subseteq\{1,\dots,n\}:\,|H|=h}{\operatorname{argmin}} \quad Q(H,\hat{\beta}) \tag{14}$$

عليه فأن مقدر Sparse LTS يكون مطابقاً إلى إيجاد المجموعة الجزئية $h \leq n$ من المشاهدات والتي تكون مطابقة الى مقدر Lasso والتي تقابل اصغر مجموع لمربعات البواقي الجزائية.

<u>2-4-3 مقدر MM Lasso [6][13]</u>

ان طريقة المربعات الصغرى الجزائية مع دالة الجزاء (Lasso) تكون غير حصينة بمعنى انها تتأثر بالقيم الشاذة . ولذلك فقد اقترح (Darwish & Buyuklu) عام 2015 استبدال دالة الخسارة المربعات الصغرى الجزائية مع دالة الجزاء (Lasso) بدالة خسارة (MM) الحصينة لنحصل على طريقة (MM Lasso) وهي طريقة فعالة للتعامل مع مشكلة الإبعاد والقيم الشاذة وان المقدر الناتج من هذه الطريقة يدعى بمقدر (MM Lasso) .

ان طريقة (MM Lasso) تتضمن ثلاث مراحل :-

المرحلة الأولى : يتم حساب مقدر أولي $(\hat{\beta}_{ini})$ ذو نقطة انهيار عالية ولكن ليس بالضرورة ان يكون كفوءاً.

المرحلة الثانية : يتم حساب مقدر (M-scale) الحصين للبواقي$(\hat{\sigma})$ بالاعتماد على المقدر الأولي. المرحلة الثالثة : يتم حساب مقدر$(L_1$-Penalized M) بثبات scale$\hat{\sigma}$ ، بداية التكرار من $\hat{\beta}_{ini}$ وباستعمال دالة خسارة نظمن الحصول على الكفاءة المطلوبة .

ليكن $\hat{\beta}_{ini}$ مقدر اولي . ولتكن $r = r(\hat{\beta}_{ini})$ تمثل البواقي والتي تحسب كما في الصيغة الاتية:-

$$r = r(\hat{\beta}_{ini}) = y_i - \acute{x}_i\hat{\beta}_{ini} \tag{15}$$

وليكن $\hat{\sigma}_{ini}$ هي M-scale من البواقي r :

$$\frac{1}{n-\hat{q}}\sum_{i=1}^{n}\rho_0\left(\frac{r(\beta)}{\hat{\sigma}_{ini}}\right) = \delta \tag{16}$$

حيث ان :

$\hat{q}$: تمثل عدد المعلمات المقدرة غير الصفرية في $\hat{\beta}$ والتي تعتمد على معلمة الجزاء (λ).

δ: يمثل ثابت ويحقق الصيغة الاتية :-

$$\delta = E\emptyset[\rho(z)] \tag{17}$$

حيث ان :-

$\emptyset$: تمثل دالة التوزيع الطبيعي المعياري .

عليه فأن دالة الهدف لطريقة (MM lasso) تكون وفقاً للصيغة الاتية:-

$$L(x,y,\beta) = \hat{\sigma}_{ini}{}^2\sum_{i=1}^{n}\rho\left(\frac{r(\beta)}{\hat{\sigma}_{ini}}\right) + n\lambda\sum_{j=1}^{p}|\beta_j| \tag{18}$$

وان(ρ)هي دالة اخرى محددة بحيث ان :

$$\rho \leq \rho_0$$

ليكن $\rho(r) = \rho_{PBI}\left(\frac{r}{k_1}\right)$ و $\rho_0(r) = \rho_{PBI}\left(\frac{r}{k_0}\right)$

ويفرض ان كل الدوال (ρ) تكون زوجية وموجبة ، فأن قيمة (k_0) يجب ان تختار بحيث يتم الحصول على نقطة انهيار عالية لمقدر (MM Lasso) .

ان اختيار (k_1) سوف يحدد كفاءة التقارب للتقدير من دون التأثير على نقطة الانهيار .

في الترتيب ليكن $\rho \le \rho_0$ ، يجب ان يكون لدينا $k_1 \ge k_0$ ، وان القيمة الكبيرة ل k_1 تعطي كفاءة عالية لطريقة (MM Lasso) والتي من المكن ان تتحقق تحت شرط التوزيع الطبيعي .

ليكن

$$\Psi(t) = \acute{\rho}(t) \quad , \quad W(t) = \frac{\Psi(t_i)}{t} \tag{19}$$

وان

$$t_i = \frac{r_i}{\widehat{\sigma}_{ini}} \quad , \quad w_i = \frac{W(t)}{2} \tag{20}$$

كذلك فأن :

$$w = (w_1, w_2, \ldots, w_n)' \quad , \quad W = \mathrm{diag}(w) \tag{21}$$

ومن خلال اشتقاق المعادلة (18) بالنسبة الى (β) ومساواتها للصفر نحصل على مقدر (MM Lasso) والذي يكون وفقاً للصيغة الاتية :-

$$\widehat{\beta}^{(i+1)}{}_{MM-lasso} = \left(\acute{X}W^{(i)}X + \lambda\, \Lambda^{(i)}\right)^{-1}\left(\acute{X}W^{(i)}y\right) \tag{22}$$

Λ : معكوس المصفوفة وتم تعريفها في الفقرة (4-3-1) الاتية :-

1) يتم استعمال مقدر اولي ذو نقطة انهيار عالية (تم اختيار مقدر Sparse LTS) . ويتم من خلاله تقدير البواقي وفقاً للصيغة التالية :-

$$r_i(\widehat{\beta}_{ini}) = y_i - \acute{x}_i\, \widehat{\beta}_{ini} \quad , \quad 1 \le i \le n$$

2) يتم حساب مقدر(M-scale) الحصين للبواقي($\widehat{\sigma}$) بالاعتماد على المقدر الأولي .

3) عند كل تكرار مع بقاء ($\widehat{\sigma}_{ini}$) ثابتة ، يتم حساب البواقي ${r_i}^{(j-1)}$ والوزن المقترن $w({r_i}^{(j-1)})$ وفقآ إلى دالة الوزن .

4) حل معادلة المربعات الصغرى التكرارية الموزونة (IRLS) ووفقاً للصيغة الاتية :-

$$\widehat{\beta}^{(j)}{}_{MM-lasso} = \left(\acute{X}W^{(j-1)}X + \lambda\, \Lambda^{(i)}\right)^{-1}\left(\acute{X}W^{(j-1)}y\right) \tag{23}$$

ويتم تكرار الخطوتين (3) و(4) حتى يصبح المقدار $\frac{|{r_i}^{(j)} - {r_i}^{(j-1)}|}{{r_i}^{(j-1)}}$ اقل من المسموح به.

اما بالنسبة لدوال الوزن لخوارزمية (IRLS) فقد تم استعمال دالة (Tukey's bisquare) والتي تكون وفقاً للصيغة الاتية :-

$$\rho_{BI}(t) = \begin{cases} \frac{k_{BI}}{6}\left[1-\left(1-\left(\frac{t}{k_{BI}}\right)^2\right)^3\right] & \text{if } |t| \le k_{BI} \\ \frac{k_{BI}}{6} & \text{if } |t| > k_{BI} \end{cases} \qquad (24)$$

وان

$$\Psi(t)_{BI} = \begin{cases} \left(1-\left(\frac{t}{k_{BI}}\right)^2\right)^2 & \text{if } |t| \le k_{BI} \\ 0 & \text{if } |t| > k_{BI} \end{cases} \qquad (25)$$

حيث ان

$\Psi(t) = \acute{\rho}(t)$:

k_{BI} : ثابت التناغم (Tuning Constant) وان $k_{BI} > 0$.

ان تقدير القياس ($\widehat{\sigma}_{ini}$) يتطلب تصحيحه بالنسبة للبيانات عالية الابعاد . وفقآ إلى ما اشارة الباحثين (Maronna, and. Yohai) عام (2010) . فأن ($\widehat{\sigma}_{ini}$) تصحح وكما في الصيغة الاتية :-

$$\tilde{\sigma} = \frac{\widehat{\sigma}_{ini}}{1-\left(k_1+\frac{k_2}{n}\right)\hat{q}/n} \quad , \qquad k_1 = 1.29 \ , k_1 = -6.02 \qquad (26)$$

اما فيما يتعلق بمعلمة الجزاء (λ) فيتم اختيارها من خلال الخطأ التنبوئ المقدر ل(MM Lasso) ولقيم مختلفة ل (λ) كمعيار العبور الشرعي (Cross Validation) . ومن الممكن استعمال k-fold (cross validation process) والذي يتطلب اعادة حساب تقدير الأوقات k . فعندما k=n ("leave-one- out") يمكننا استخدام التقريب لتجب اعادة التقدير . ويتم استدعاء $\hat{y}_{-i}$ المطابقة ل y_i والتي تحسب من دون استخدام المشاهدة i-th .

اي ان

$$y_{-i} = \acute{x}_i\widehat{\beta}^{(-i)}$$

حيث ان $\widehat{\beta}^{(-i)}$ يكون تقدير هو (MM lasso) والذي يحسب من دون المشاهدات i . عليه تقريب تايلور من الدرجة الاولى من التقدير يعطي الاخطاء التنبوئية التقريبية وكما في الصيغة الاتية:-

$$r_i = y_i - \hat{y}_{-i} \approx \left(1+\frac{W(t_i)h_i}{1-h_i\Psi(t_i)}\right) \qquad (27)$$

و ان قيمة (h) تحسب وفقاً للصيغة الاتية :-

$$h_i = \acute{x}_i\left(\sum_{i=1}^{n}\Psi(t_i)x_i\acute{x}_i + 2\lambda\,\Lambda^{(i)}\right)^{-1} x_i \qquad (28)$$

ومن خلال الأخطاء التنبؤية $r_{-} = (r_{-1}, r_{-2}, \dots, r_{-n})^{-1}$ ، يتم حساب متوسط مربعات الخطاء الحصين (MSE) ويرمز له بالرمز $\tau(r_{-})^2$ ، حيث ان τ قياس مع ثابت حالة ($c_\tau = 5$) ، ويتم اختيار λ التي تجعل من MSE اصغر ما يمكن .

4-الجانب التجريبي

من اجل المقارنة بين المقدرات الجزائية الحصينة الواردة في الجانب النظري تم الاعتماد على اسلوب المحاكاة (Simulation) بطريقة (Monte Carlo) بغية الحصول على عدد كبير من التجارب والتي تكون اكثر شمولية من حيث حجوم العينات والقيم الاولية للمعلمات ، وقد تم الاعتماد على المعايير الاحصائية متوسط مربعات الخطأ (MSE) ،معدل الايجابية الكاذب (*FPR*) ومعدل السلبية الكاذب(*FNR*) والتي توصف كالاتي :-

1) متوسط مربعات الخطأ *Mean Square Error (MSE)*

ان احدى اهداف تقدير النموذج المتبعثر (*Sparse Model Estimation*) هو تحسين اداء التنبؤ ، والمقدرات المختلفة يتم تقيمها من خلال متوسط مربعات الخطأ ويرمز له بالرمز (*MSE*) والذي يكون وفقاً للصيغة الاتية :-

$$\mathrm{MSE}(\hat{\beta}) = \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)^2 \tag{29}$$

ويكون المقدر الافضل هو الذي يعطي اقل (MSE) .

2) فيما يتعلق بالتبعثر ، يتم تقييم النماذج المقدرة من خلال معدل الايجابية الزائف (*False Positive Rate*) ويرمز له بالرمز (*FPR*) ومعدل السلبية الزائف (*False Negative Rate*) ويرمز له بالرمز(*FNR*)

ان معدل الايجابية الزائف هو المعاملات المساوية للصفر في النموذج الحقيقي ، ولكن تكون غير صفرية في التقدير . اما معدل السلبية الزائف هي المعاملات غير الصفرية في النموذج الحقيقي ولكن تكون صفرية في التقديري.

$$FPR(\hat{\beta}) = \frac{\left|\{j \in \{1, \dots, p\}: \hat{\beta}_j \neq 0 \cap \beta_j = 0\}\right|}{\left|\{j \in \{1, \dots, p\}: \beta_j = 0\}\right|} \tag{30}$$

$$FNR(\hat{\beta}) = \frac{\left|\{j \in \{1, \dots, p\}: \hat{\beta}_j = 0 \cap \beta_j \neq 0\}\right|}{\left|\{j \in \{1, \dots, p\}: \beta_j \neq 0\}\right|} \tag{31}$$

1-4 خطوات اجراء المحاكاة

4-1-1توليد المتغيرات التوضيحية

يتم توليد المتغيرات التوضيحية وفق التوزيع الطبيعي متعدد المتغيرات ووفق الصيغة الاتية:-

$$X \sim MN(0, \Sigma)$$

حيث ان :-

$$\Sigma_{ij} = \rho^{|i-j|} \quad , \quad \rho = 0.5$$

2-1-4 توليد الاخطاء العشوائية

يتم توليد الاخطاء العشوائية وفقاً للتوزيع الطبيعي بمتوسط (صفر) وتباين (1) اي ان :-

$$\epsilon_i \sim N(o,1) \qquad ,i = 1,2,\dots,n$$

اما نسب التلويث فهي *(0%,10%,20%)* وان عملية التلويث تمت عن طريق تلويث توزيع حد الخطأ ب $N(20,1)$ بدل $N(o,1)$

3-1-4 وصف التجارب والقيم الافتراضية

يتم وصف القيم الافتراضية للمعلمات وحجوم العينات كالاتي :-

التجربة الاولى :- *p=40 ,n=50,100*

$$\beta = (3,1.5,0,0,2,0,\dots,0)$$

التجربة الثانية :- *p=150 ,n=50,100*

$$\beta = (3,3,3,3,3,0,\dots,0)$$

جدول (1) يبين نتائج التجربة الاولى ولجميع المقدرات الجزائية الحصينة

نسب التلويث	n	Methods	MSE	FPR	FNR
0%	50	Lasso	0.00967	0.22	0
		SLTS	0.01813	0.14	0
		MM Lasso	0.00651	0.12	0
10%		Lasso	0.27384	0.17	0.21
		SLTS	0.01415	0.17	0
		MM Lasso	0.00783	0.14	0
20%		Lasso	0.33831	0.15	0.36
		SLTS	0.01117	0.19	0
		MM Lasso	0.38083	0.18	0.01
0%	100	Lasso	0.00326	0.19	0
		SLTS	0.00573	0.07	0
		MM Lasso	0.00280	0.06	0
10%		Lasso	0.11908	0.19	0.06
		SLTS	0.00419	0.10	0
		MM Lasso	0.00311	0.09	0
20%		Lasso	0.19007	0.17	0.17
		SLTS	0.00386	0.13	0
		MM Lasso	0.00489	0.13	0

جدول(2) يبين نتائج التجربة الثانية ولجميع المقدرات الجزائية الحصينة

نسب التلويث	n	Methods	MSE	FPR	FNR
0%	50	Lasso	0.00136	0.09	0
		SLTS	0.00328	0	0
		MM Lasso	0.00074	0	0
10%		Lasso	0.04928	0.09	0
		SLTS	0.00252	0	0
		MM Lasso	0.00089	0	0
20%		Lasso	0.09244	0.10	0.02
		SLTS	0.00210	0.01	0
		MM Lasso	0.00128	0.01	0
0%	100	Lasso	0.00015	0.05	0
		SLTS	0.00143	0	0
		MM Lasso	0.00011	0	0
10%		Lasso	0.02489	0.08	0
		SLTS	0.00110	0	0
		MM Lasso	0.00012	0	0
20%		Lasso	0.04350	0.08	0
		SLTS	0.00090	0	0
		MM Lasso	0.00013	0	0

5-الاستنتاجات والتوصيات

1) من خلال تحليل نتائج التجربة الاولى فأن مقدر (MM Lasso) هو الافضل كونه يعطي اقل قيمة لـ (MSE) ولمختلف حجوم العينات.

2) في حالة زيادة نسب التلويث بنسبة (20%) فأن مقدر (SLTS) هو الافضل كونه يعطي اقل قيمة لـ (MSE) .

3)من خلال تحليل نتائج التجربة الثانية فأن مقدر (MM Lasso) هو الافضل كونه يعطي اقل قيمة لـ MSE) ولمختلف حجوم العينات.

4) حقق مقدر (MM Lasso) افضلية وفي معظم التجارب من ناحية اختيار المتغيرات كونه يعطي اقل قيمة لـ (FPR) .

5) نوصي باعتماد طريقة (MM Lasso) في حالة وجود الشواذ في البيانات .

6) نوصي باستعمال دوال جزاء اخرى في عملية التقدير.

المصادر

1- علي ،عمر عبد المحسن، المهنا ، فراس احمد محمد (2010) "حول تقليص تقدير المركبات الرئيسية مع التطبيق " ، المجلة العراقية للعلوم الاحصائية ، العدد :14.

2 - حمود ، مناف يوسف ، صالح ، طارق عزيز (2015) "استعمال بعض طرائق تقدير الانموذج شبه المعلمي احادي المؤشر " ، مجلة العلوم الاحصائية ، العدد :7

3- Alfons, A. Croux, C . Gelper, S (2013) "Sparse least trimmed squares regression for analyzing high dimensional large data sets," The Annals of Applied Statistics, vol. 7, no. 1, pp. 226–248, 2013.

4-Buhlmann, Peter. Geer, Sara (2013) " for High Dimensional Data Methods, Theory and Applications". Springer

5- Coster, Jamie De (2003) "Notes on Applied Linear Regression". Department of Social Psychology Free University Amsterdam .

6- Darwish. Kamal. Buyuklu, Hakan (2015)"Robust Linear Regression Using L1-Penalized MM-Estimation for High Dimensional Data".American Journal of Theoretical and Applied Statistics.4(3),PP 78-84.

7- Fan,J. Li,, (2001)"Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," Journal of the American Statistical Association, vol. 96, no. 456, pp. 1348–1360 .

8- Fu, Wenjiang J (1998), "Penalized Regressions: The BridgeVersus the Lasso", Journal of Computational and Graphical Statistics, Volume 7, Number 3, Pages 397–416

9- James, Gareth. Witten, Daniela . Hastie , Trevor. Tibshirani, Robert. (2013) " An Introduction to Statistical Learning with Applications in R". Springer .

10- Kutner, Michael H. Nachtsheim Christopher J . Neter , John. Li, William. (2007) " Applied Linear Statistical Models" Fifth edition.

11- Muller , Patric . (2000), "L_1Regularization for Non-linear Models," Thesis Submitted To The Council University.

12- P.J.Rousseeuw, K. Van Driessen, Computing LTS regression for large data sets, Technical Report, University of Antwerp, 1998.

13- Susant, Yuliana . Pratiw, Hasih . H, Sri .Sulistijowati , Lian. Twenty .(2014)" M Estimation, S Estimation, AND MM Estimation Robust Regression "International Journal of Pure and Applied Mathematics, Volume 91 No. 3,PP. 349-360 .

14- Tibshirani, R. (1996)"Regression shrinkage and selection via the lasso" J. Royal. Statist. Soc B., vol. 58, no. 1, pp. 267–288

comparison between some of the robust penalized estimators using simulation

Abstract

The penalized least square method is a popular method to deal with high dimensional data ,where the number of explanatory variables is large than the sample size . The properties of penalized least square method are given high prediction accuracy and making estimation and variables selection

At once. The penalized least square method gives a sparse model ,that meaning a model with small variables so that can be interpreted easily .The penalized least square is not robust ,that means very sensitive to the presence of outlying observation , to deal with this problem, we can used a robust loss function to get the robust penalized least square method ,and get robust penalized estimator and it can deal problems of dimensions and outliers .In this paper a compression had been made Sparse LTS estimator and MM Lasso by using simulation and the simulation results show that the MM Lasso is best for every experiments, Depending on the criteria for the Mean Square Error, False Positive Rate and False negative Rate .

Key word: strong criminal capabilities، Using simulation .