ISSN (1681- 6870)

مقارنة بين بعض المقدرات الجزائية في الانحدار التقسيمي ذو الابعاد العالية

م.د.علي حميد يوسف[1]
[1] جامعة واسط ، كلية الادارة والاقتصاد،قسم الاحصاء

م.م.مروة جمعة طعمة[2]
[2] جامعة سومر ، كلية الادارة والاقتصاد،قسم ادارة الاعمال

المستخلص

في هذا البحث تم المقارنة بين عدد من المقدرات الجزائية في نموذج الانحدار التقسيمي ذو الابعاد العالية وخاصاً التي يكون فيها عدد المتغيرات كبير ، وهذه المقدرات هي (elastic net - lasso - ridge) وتمت عملية المقارنة بين هذه المقدرات عن طريق المحاكاة وبالاعتماد على المقاييس الاحصائية جذر متوسط مربعات الخطأ التنبؤي، معدل الايجابية الزائف (FPR) ومعدل السلبية الزائف (FNR) . وتم التوصل الى افضلية مقدر elastic net على باقي المقدرات ولمختلف حجوم العينات .

الكلمات المفتاحية: الانحدار التقسيمي، Lasso، Ridge،Elastic-Net

Comparison Between Some of Penalized Estimators for High Dimensional Quantile Regression

Dr. Ali Hameed Yousif
Wasit University / College of Administration and Economics / Department of Statistics / Iraq.
ahameed@uowasit.edu.iq

Marwa Juma Tohme
Sumer University / College of Administration and Economics / Department of Business Administration / Iraq.
marwajuma1989@gmail.com

Abstract:

In this research, many of penalized estimator have been compared in the quantile regression model with high dimensions, and these estimator are (ridge - lasso - elastic net). The process of comparison among of those estimators were done by simulation and based on statistical measures, root mean predictive error, false positive rate (FPR) and false negative rate (FNR).Simulation results show that an elastic net estimator is the best of other estimators.

Keywords: **Quantile regression model, false positive rate, false negative rate.**

ISSN (1681- 6870)

1. المقدمة [4,6,7] Introduction

الانحدار التقسيمي هو نوع من تحليل الانحدار المستخدم في الإحصاء والقياس الاقتصادي. وكما هو معلوم فأنه من خلال طريقة المربعات الصغرى فأنه يتم تقدير المتوسط الشرطي لمتغير الاستجابة عند قيم محددة للمتغيرات التوضيحية ، بينما الانحدار التقسيمي يهدف الى تقدير التوزيع الشرطي لمتغير الاستجابة عند نقاط مختلفة . وبشكل أساسي ، فأن الانحدار التقسيمي هو امتداد الانحدار الخطي ويتم استخدامه عندما تكون شروط الانحدار الخطي غير قابلة للتطبيق.

في السنوات الاخيرة زاد الاهتمام بموضوع الانحدار التقسيمي حيث يوفر معلومات اكثر من انحدار المتوسط الكلاسيكي .

ان الانحدار التقسيمي يكون مرغوب إذا كانت الدوال التقسيمية الشرطية ذات أهمية. احدى ميزات الانحدار التقسيمي بالنسبة الى انحدار المربعات الصغرى العادية ، هو أن تقديرات الانحدار التقسيمية تكون أكثر حصانة ضد القيم الشاذة في متغير الاستجابة. ومع ذلك ، فإن عامل الجذب الرئيسي للانحدار التقسيمي يتجاوز ذلك. فقد تكون المقاييس المختلفة للاتجاه المركزي والتشتت الإحصائي مفيدة للحصول على تحليل أكثر شمولية للعلاقة بين المتغيرات .

ان الانحدار التقسيمي قد اقترح من قبل Koenker and Bassett عام 1978 حيث قدم كورقة سمنار . فلو فرضنا ان لدينا عينة عشوائية $(y_1, x_1) \dots (y_n, x_n)$ ، فأن نموذج الانحدار الخطي التقسيمي للتقسيم θ^{th} و$(0 < \theta < 1)$ يكون كالاتي :-

$$y_i = x_i'\beta + u_i \qquad , i = 1,2, \dots, n \qquad (1)$$

حيث ان :-

β :- يمثل موجه المعلمات وان $\beta = (\beta_1, \beta_2, \dots, \beta_p) \in R^p$

ISSN (1681- 6870)

u_i :- متغير عشوائي مستقل بتقسيم θ^{th} مساوي للصفر

فمن الممكن ان نبين ان المعلمات β يمكن تقديرها من خلال المعادلة الاتية :-

$$min_\beta \sum_{i=1}^{n} \rho_\theta(y_i - x_i'\beta) \qquad (2)$$

اذ ان :-

$$\rho_t(t) = \begin{cases} \theta t & if \quad t \geq 0 \\ -(1-\theta)t & if \quad t < 0 \end{cases} \qquad (3)$$

2.هدف البحث

يهدف هذا البحث المقارنة بين عدد من المقدرات الجزائية في الانحدار التقسيمي وهذه المقدرات هي (ridge – elastic net – lasso) وتمت عملية المقارنة عن طريق استخدام المحاكاة بطريقة مونت كارلو وبالاعتماد على عدد المقاييس الاحصائية وبمختلف حجوم العينات .

3. الانحدار الجزائي التقسيمي [6,10]

Regularized quantile regression

ان الانحدار التقسيمي يكون غير فعال في التعامل مع البيانات ذات الابعاد العالية كما لا يمكن من خلاله القيام بعملية اختيار المتغيرات(Variable Selection) والتي تعد احدى المسائل المهمة في الاحصاء . وللتغلب على تلك المشاكل فأنه يتم اضافة حد الجزاء الى دالة الخسارة التقسيمية لنحصل على طريقة الانحدار التقسيمي الجزائي والذي يكون بالصيغة الاتية :-

ISSN (1681- 6870)

$$Q_\theta(\beta) = \sum_{i=1}^{n} \rho_\theta(y_i - x_i'\beta) + \lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|) \qquad (4)$$

حيث ان :-

$p_\lambda(.)$:- تمثل دالة الجزاء

λ :- تمثل معلمة الجزاء وان $\lambda \geq 0$ وهي تعمل على الموائمة بين التحيز والتباين .

وتوجد العديد من المقدرات الجزائية في الانحدار التقسيمي ومنها .

1.3 مقدر Ridge quantile regression [5,8]

ان انحدار الحرف هو تقنية لتحليل البيانات التي تعاني من مشكلة التعدد الخطي شبه التام ، حيث انه عندما تظهر مشكلة التعدد الخطي فأن تقديرات المربعات الصغرى سيكون لها تباين عالي . وقد اقترح Hoeral and Kennard عام 1970 اضافة درجة من التحيز الى تقديرات الانحدار كما ان انحدار الحرف سيقلل من الاخطاء القياسية. ان دالة الجزاء Ridge تكون وفقاً للصيغة الاتية :-

$$\sum_{j=1}^{p} \beta_j^2 \qquad (5)$$

وبإضافة دالة الجزاء Ridge الى دالة الخسارة نحصل على دالة الانحدار الجزائية التقسيمية بالاعتماد على دالة الجزاء Ridge Ridge يتم الحصول عليه من خلال الصيغة الاتية :-

$$\hat{\beta}_{RQR} = argmin_\beta \left\{ \sum_{i=1}^{n} \rho_\theta(y_i - x_i'\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \qquad (6)$$

2.3 مقدر Lasso quantile regression [1,2,9]

ان انحدار Lasso هي مختصر ل (Least Absolute Shrinkage and Selection Operator) والذي قد اقترح من قبل Tibshiraniعام 1996 . وتتلخص فكرة Lasso انها تعمل على مبدأ تصغير مجموع المربعات بالنسبة للخطأ وفق قيد معين يتم فرضه على المعلمات والذي يمثل المجموع المطلق للمعلمات . ومن خلال طبيعة القيد فأن مقدر Lasso يعمل على جعل عدد من المعاملات مساوية للصفر، وتقليص الأخرى بمقدار معين وبالتالي فأنه يعمل على اختيار المتغيرات المهمة في النموذج وبمعنى اخر فأنه يقوم بعملية التقدير واختيار المتغيرات في ان واحد . ان دالة الجزاء Lasso تكون وفقاً للصيغة الاتية :-

$$\sum_{j=1}^{p}|\beta_j| \qquad (7)$$

وبإضافة دالة الجزاء Lasso الى دالة الخسارة نحصل على دالة الانحدار الجزائية التقسيمية بالاعتماد على دالة الجزاء Lasso يتم الحصول عليه من خلال الصيغة الاتية :-

$$\hat{\beta}_{LQR} = \operatorname{argmin}_{\beta}\left\{\sum_{i=1}^{n}\rho_{\theta}(y_i - x_i'\beta) + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \qquad (8)$$

3.3 مقدر Elastic-Net quantile regression [10, 11]

ان انحدار elastic-net مزيج ما بين انحدار ridge وانحدار lasso والذي يعمل بشكل جيد عندما يكون هناك الكثير من المتغيرات عديمة الفائدة التي يجب إزالتها من النموذج وله اداء جيد في معالجة المتغيرات المرتبطة. ان الباحثان (Zou& Hastie) عام 2005 قد اقترحا دالة الجزاء elastic- net والتي تكون كما في الصيغة الاتية :-

$$P_{\lambda\alpha}(\beta) = \lambda\sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right], \quad 0 < \alpha < 1 \qquad (9)$$

اذ إن :-

ISSN (1681- 6870)

$P_{\lambda\alpha}(\beta)$: تمثل دالة الجزاء (Penalty Function) .

λ :- تمثل معلمة الجزاء(Penalty Parameter) وان $\lambda > 0$

وبإضافة دالة الجزاء elastic- net الى دالة الخسارة نحصل على دالة الانحدار الجزائية التقسيمية بالاعتماد على دالة الجزاء elastic- net يتم الحصول عليه من خلال الصيغة الاتية :-

$$\hat{\beta}_{ENQR} = argmin_{\beta}\left\{\sum_{i=1}^{n}\rho_{\theta}(y_i - x_i'\beta) + P_{\lambda\alpha}(\beta)\right\} \qquad (10)$$

ولحساب مقدر elastic-net quantile يتم الاعتماد على الخوارزمية المقترحة من قبل (Congrui and Jian) عام (2016) والتي تدعى بخوارزمية (semismooth Newton coordinate descent) ويرمز لها بشكل مختصر ب (SNCD) .

4. المحاكاة Simulation [3]

من اجل المقارنة بين عدد المقدرات يتم تنفيذ تجارب المحاكاة بطريقة Monte Carlo وبالاعتماد على برنامج R وبتكرار مساوي الى (200) ، اما عملية توليد البيانات فتكون من خلال نموذج الانحدار الاتي :-

$$y = \acute{x}\beta + \sigma u \qquad (11)$$

حيث ان :-

y :متجه المتغير المعتمد من الدرجة nx1 .

x : مصفوفة المتغيرات التوضيحية من الدرجة nxp

β :متجه المعلمات من الدرجة px1 .

ISSN (1681- 6870)

u : متجه الخطأ العشوائي من الدرجة nx1 .

σ :- تأخذ القيمتين (0,1)

ان تجارب المحاكاة يكون كالاتي :-

التجربة الاولى n=50, p=15

$$\beta = (1,1,1,1,1,1,1,1,1,1,0,\dots,0)$$

التجربة الثانية n=100, p=25

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

اما x يتم توليدها من خلال التوزيع الطبيعي متعدد المتغيرات وكما في الصيغة الاتية :-

$$x \sim MN(0,\sum)$$

وان

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

وبالنسبة بالنسبة الى القيم الافتراضية ل (quantile) فهي $\theta = (0.25,0.50,0.75)$

ويتم الاعتماد على جذر متوسط مربعات الخطأ التبؤي للمقارنة بين المقدرات والذي يكون كما في الصيغة الاتية :-

$$RMSPE(\hat{\beta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^* - x_i^{*\prime}\hat{\beta}\right)^2} \qquad (12)$$

حيث ان

y_i^* , $x_i^{*\prime}$ يمثلان المتغيرات التوضيحية والمعتمدة على التوالي بالاعتماد على بيانات الاختبار.

ISSN (1681- 6870)

وفيما يتعلق بعملية اختيار المتغيرات، فيتم الاعتماد على المقياسين معدل الايجابية الزائف (False Positive Rate) والذي يرمز له بالرمز (FPR) وكذلك معدل السلبية الزائف (False Negative Rate) والذي له بالرمز (FNR) وكما في الصيغتين الاتيتين :-

$$\mathrm{FPR}(\hat{\beta}) = \frac{\left|\{j \in \{1,\dots,p\}: \hat{\beta}_j \neq 0 \cap \beta_j = 0\}\right|}{\left|\{j \in \{1,\dots,p\}: \beta_j = 0\}\right|} \qquad (13)$$

$$\mathrm{FNR}(\hat{\beta}) = \frac{\left|\{j \in \{1,\dots,p\}: \hat{\beta}_j = 0 \cap \beta_j \neq 0\}\right|}{\left|\{j \in \{1,\dots,p\}: \beta_j \neq 0\}\right|} \qquad (14)$$

جدول (1): يبين نتائج المحاكاة للتجربة الاولى ولجميع المقدرات عندما P=15,n=50

θ	σ	estimators	MAPE	FPR	FNR
0.25	$\sigma = 1$	Ridge	1.5115	1	0
		Lasso	1.2747	0.62	0
		Elastic Net	1.2742	0.63	0
	$\sigma =2$	Ridge	2.5640	1	0
		Lasso	2.5375	0.55	0.05
		Elastic Net	2.4985	0.61	0.03
0.5	$\sigma = 1$	Ridge	1.4126	1	0
		Lasso	1.2557	0.54	0
		Elastic Net	1.2565	0.55	0
		Ridge	2.4929	1	0

	σ =2	Lasso	2.4776	0.51	0.03
		Elastic Net	2.4636	0.54	0.02
0.75	σ = 1	Ridge	1.4771	1	0
		Lasso	1.3382	0.59	0
		Elastic Net	1.3372	0.56	0
	σ =2	Ridge	2.5797	1	0
		Lasso	2.6117	0.50	0.07
		Elastic Net	2.5919	0.51	0.05

جدول (2): يبين نتائج المحاكاة للتجربة الثانية ولجميع المقدرات عندما$p=25,n=100$

θ	σ	estimators	MAPE	FPR	FNR
0.25	**σ = 1**	**Ridge**	**1.3992**	**1**	**0**
		Lasso	**1.1352**	**0.413**	**0**
		Elastic Net	**1.1379**	**0.432**	**0**
		Ridge	**2.4136**	**1**	**0**

ISSN (1681- 6870)

	$\sigma = 2$	Lasso	2.2636	0.3975	0.03
		Elastic Net	2.2654	0.415	0.026
0.5	$\sigma = 1$	Ridge	1.3702	1	0
		Lasso	1.099	0.388	0.002
		Elastic Net	1.1052	0.425	0
	$\sigma = 2$	Ridge	2.3709	1	0
		Lasso	2.1974	0.3925	0.034
		Elastic Net	2.1939	0.417	0.03
0.75	$\sigma = 1$	Ridge	1.4032	1	0
		Lasso	1.1279	0.378	0.002
		Elastic Net	1.1357	0.389	0
	$\sigma = 2$	Ridge	2.4076	1	0
		Lasso	2.2436	0.367	0.04
		Elastic Net	2.2523	0.384	0.034

ISSN (1681- 6870)

5. الاستنتاجات والتوصيات

من خلال تجارب المحاكاة تم التوصل الى الاتي:

1- افضليه مقدر Elastic Net بشكل عام في معظم تجارب المحاكاة من ناحية اقل قيمة لجدر متوسط مربعات الخطأ التبؤي.

2- حصل مقدر Lasso على اقل قيمية بالنسبة ب (FPR) و(FNR) .

3- افضل قيمة لانحدار التقسيمي هي عندما $(\theta = 0.5)$ وبحي نتائج المحاكاة .

4 – اعتماد مقدر Elastic Net في عملية التقديرفي حالة كون عدد المتغيرات التوضيحية كبير .

المصادر :-

اولاً :- المصادر العربية

1- عبودي ، عماد حازم ويوسف، علي حميد (2017) "مقارنة بين بعض المقدرات الجزائية الحصينة باستخدام المحاكاة ".مجلة العلوم الاقتصادية والادارية . العدد 100 ، المجلد 23 .

2-عبودي ، عماد حازم ويوسف، علي حميد (2017) "مقارنة بين مقدري Huber Lasso و Huber Elastic Net باستخدام المحاكاة ".مجلة الكوت للعلوم الاقتصادية والادارية. العدد 28 ، الجزء الاول .

ثانياً :- المصار الاجنبية

[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, 7(1), 226-248.

[4] Bühlmann, P., & Van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

ISSN (1681- 6870)

[5] Hoerl, A.E., and Kennard, R.W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems, "Technometrics, 12, 55–67.

[6] Koenker, R. (2005). Quantile Regression. Cambridge Books. Cambridge University Press.

[7]Koenker, R. and G. J. Bassett (1978). Regression quantiles. Econometrica 46, 33{50.

[8] Maronna, R (2011) "Robust ridge regression for high–dimensional data", Technimetrics, vol. 53, pp. 44–53

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

[10] Yi ,Congrui . Huang , Jian (2016)" Semismooth Newton Coordinate Descent lgorithm for Elastic–Net Penalized Huber Loss Regression and Quantile Regression . Journal of Computational and Graphical Statistics, Volume 26, 2017 - Issue 3

[11] Zou, Hui.& Hastie, Trevor. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B, 67, 301–320.