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

استعمال طريقة انحدار لاسو البيانية الممكنة FGLasso لتقدير معالم نموذج الانحدار الخطّي العام في ظلّ وجود بيانات عالية الأبعاد مع تطبيق

Using Feasible Graphical Lasso Regression Method to Estimate the Parameters of General Linear Regression Model Under High Dimensional Data with Application

محمد جاسم فرحان[1]
Mohammed Jasim Farhan

أ.م. د. أحمد مهدي صالح[2]
Asst. Pro. Ahmed Mahdi Salih

كلّية الإدارة والاقتصاد/ جامعة واسط

المستخلص

تُعدّ مشكلة الأبعاد الكبيرة من المشاكل المهمّة التي تعاني منها البيانات، فبوجود بيانات عالية الأبعاد نموذج الانحدار الخطّي العام تزداد صعوبة استعمال طرائق التقدير الكلاسيكية كطريقة الإمكان الأعظم أو طريقة العزوم، إذْ تقود هذه الطرائق إلى تقديرات متحيّزة وغير كفؤة لتقدير معالم النموذج. وهنا دعت الحاجة إلى استعمال طرائق أخرى لتقدير معالم نموذج الانحدار الخطّي العام. وفي بحثنا هذا تَمَّ استعمال طريقتين هما: طريقة SICA **Smooth Integration of Counting Absolute Deviation** ، وكذلك تَمّ استعمال طريقة لاسو البيانية الممكنة FGLasso ، وتمّ استعمال بيانات محاكاة، وكذلك بيانات حقيقية تُمثّل مستوى التلوّث في نهر دجلة. تَمَّتْ مقارنة النتائج باستعمال متوسّط مربّعات الخطأ MSE، و أظهرت النتائج أفضلية مقدّر FGLasso.

الكلمات المفتاحية :- لاسو البيانية الممكنة ، الأبعاد الكبيرة ، التلوّث.

Abstract

The high dimensions problem is one of the important problems that affect the data, and with company of high dimensional data in the general linear regression model, it becomes very difficult to use classical estimation methods such as maximum likelihood method and moments

method because these methods lead to biased and inefficient estimates for parameters of the model. The necessity to use new methods to estimate the parameters of the general linear regression model. In this research, two methods were used, the ***Smooth Integration of Counting Absolute Deviation*** **SICA** method, as well as the ***Feasible Graphical Lasso*** method **FGLasso**. Simulation data were used, as well as real data representing the level of pollution in the Tigris River. The results compared using the **Mean Square Error MSE** and the results showed that the **FGLasso** estimator is the best.

Keywords:- Feasible Graphical Lasso, High Dimensions, Pollution

1- المقدّمة

تُعدّ مشكلة الأبعاد الكبيرة للعيّنة تحت الدراسة، عندما يكون عدد المتغيّرات ρ أكبر من حجم العيّنة n إحدى المشاكل التي تواجه أيّ باحث إحصائي هذه المشكلة من التحدّيات على المستوى النظري، وكذلك العملي لكثير من الدراسات والبحوث. ويعتمد الباحثون إلى اختصار أو استبعاد مجموعة معيّنة من التغيّرات ذات التأشير القليل أو غير المؤثّر إحصائيا. في حين يرى الآخرون ضرورة أخذ التغيّرات كُلّها تحت الدراسة، واتّباع بعض الأساليب النظرية لحلّ المشاكل الرئيسة التي تجلبها الزيادة الكبيرة في عدد المتغيّرات نسبة إلى حجم العيّنة. وفي دراستنا هذا سنعتمد انموذج الانحدار الخطّي العام [1] .

$$\boldsymbol{Y = XB + \varepsilon} \qquad \textbf{...(1)}$$

إذْ إنّ:

$\mathbf{Y}$ هومتّجه ($\boldsymbol{(n \times 1)}$) يمثّل المتغيّر المعتمد.

$\boldsymbol{X}$ تمثّل مصفوفة($\boldsymbol{n \times \rho}$) من المتغيّرات التوضيحية، والتي تحتوي على كمّية كبيرة من المتغيّرات.

$\boldsymbol{\varepsilon}$ يمثّل متّجه($n\times1$)، ويمثّل الخطأ العشوائي، وفي حين B، وهو متّجه ($\rho \times 1$)من المعالم المراد تقديرها، إذّ إنّ تقدير هذا المعالم عادة يكون باستعمال طرائق التقدير الكلاسيكية مثل طريقة الإمكان الأعظم، ولكن لوجود مشكلة الأبعاد الكبيرة تلجأ إلى استعمال طرائق أكثر تطوّرا التغلّب على المشاكل التي تجلبها الأبعاد الكبير [3] .

2- أهمّية وهدف البحث:

تبرز أهمّية البحث من خلال أهمّية ظاهرة تلوّث مياه الأنهار، وما لها من تأثير كبير في حياة الفرد، و كذلك مستقبل الأجيال اللاحقة. وتتنوّع أسباب التلوّث في العراق، و كذلك تتعدّد أنواع الملوّثات منها: العضوية (نتاج مخلّفات الإنسان)، و منها الصناعية (مخلّفات المصانع)، وكذلك هناك ملوّثات طبيعية بيئية. و تُعدّ هذه الظاهرة جديرة بالدراسة والتحليل، واستعمال الأدوات، و الأساليب الإحصائية؛ لتحليل بياناتها، ودراسة سلوكياتها الحالية، والتنبُّؤ بسلوك الظاهرة وحجمها مستقبلا.

يهدف بحثنا هذا إلى استعمال نموذج الانحدار الخطّي العام؛ لتحليل بيانات تلوّث نهر دجلة، إذْ إنّ العيّنة أساسا تعاني من مشكلة الأبعاد الكبيرة، أيّ إنّ عدد المتغيّرات تحت الدراسة أكبر من حجم العيّنة. وهنا دعت الحاجة إلى استعمال طرائق جديدة لتقدير معالم نموذج الانحدار الخطّي عالي الأبعاد، في ظلّ وجود هذه المشكلة والمقارنة بين هذه الطرائق؛ لاختيار الطريقة الأفضل في التقدير.

3- طريقة SICA : *Smooth Integration of Counting Absolute Deviation*

بشكل عام تأخذ طرائق الانحدار الجزائية أشكال عدّة، والهدف النهائي هو تقليل الأخطاء التجريبية، إذْ تُمّ استعمال تقنيات عدّة لتقليص متّجه معالم الانحدار. وهنا قسّم $Lv\ \delta\ Fan$ في عام 2009 طريقة جديدة تُدعى طريقة $SICA$ ***Smooth Integration of Counting Absolute Deviation*** *، وتعتمد على تحوير مَعْلَمَة الجزاء* λ ؛ لكي تكون نسبية تحت شروط الأبعاد الكبيرة. وهذا العمل أساسا تَمَّ تقديمه من قبل [7] Nikolova ، والتي اقترحت أنْ تكون دالّة الجزاء هي نسبية بين المجموع الطبيعي $Lp-Norms$ من الدرجة الأولى كما يأتي:

$$f\ (\lambda,\beta)=\lambda\frac{(a+1)\|\beta\|_1}{a+\|\alpha\|_1} \quad \text{... (2)}$$

إذْ إنّ a هو ثابت أكبر من الصفر $a>0$ ، وتمثّل α مقدّرات مسبقة لمعالم الانحدار. قام [2] $Fan\ \delta\ lv$ باقتراح تعديل على مقام المعادلة (2) كما يأتي:

$$\boldsymbol{f(\lambda,\beta)=\lambda\frac{(a+1)\|\beta\|_1}{(a+\|\alpha\|_1)^2}} \quad \text{... (3)}$$

وهنا كأنّما أصبحت دالّة الجزاء هي نسبية بين مجموع الطبيعي من الدرجة الأولى والدرجة الثانية، واقترح **Lv δ Fan** استعمال تقديرات المربعات الصغرى الآتية [6] :

$$\boldsymbol{\alpha = (\acute{X}X)^{+}\acute{X}Y} \quad \textbf{...(4)}$$

أذْ إنّ (+) إشارة الى معكوس $\textbf{Moore} - \textbf{Penrose}$

وحيث إنّ [4]

$$\hat{\boldsymbol{\beta}_{SKA}} = \textbf{arg}_{\boldsymbol{\beta}}\, \boldsymbol{min}\, (\boldsymbol{n^{-1}\acute{\varepsilon}\varepsilon + \lambda w\|\beta\|_1}) \quad \textbf{...(5)}$$

إذْ إنّ $\mathbf{W}$

$$\mathbf{W} = \boldsymbol{\lambda}\frac{\boldsymbol{a+1}}{\boldsymbol{(a+\|\alpha\|_1)^2}} \quad \textbf{...(6)}$$

و بحلّ المتباينة في معادلة (5)

$$\boldsymbol{L(\beta)} = (\mathbf{Y} - \mathbf{X\beta})'\,(\boldsymbol{Y - X\beta}) + \boldsymbol{\lambda w\acute{c}\beta} \quad \textbf{...(7)}$$

إذْ إنّ $\boldsymbol{C}$ متّجه أحادي بأبعاد $(1 \times \rho)$ولتبسيط المعادلة [10]

$$\boldsymbol{L(\beta)} = (\boldsymbol{Y'Y - 2\beta X'Y + \beta X'X\beta}) + \boldsymbol{\lambda wc'\beta} \quad \textbf{...(8)}$$

وباشتقاق المعادلة (8) بالنسبة إلى $\boldsymbol{\beta}$ ومساواته بالصفر

$$\mathbf{0} = -\boldsymbol{X'Y} + \boldsymbol{X'X\beta} + \boldsymbol{\lambda wc} \qquad \textbf{...(9)}$$

$$\boldsymbol{X'Y} - \boldsymbol{\lambda wc} = \boldsymbol{X'X\beta} \qquad \textbf{...(10)}$$

وباستعمال معكوس $Moore - penrose$ اقترح $Lv\ \&\ Fan$ استعمال الصيغة الآتية في حالة الأبعاد الكبيرة كما يأتي [9]:

$$\boldsymbol{\hat{\beta}^{SLCA}} = (\boldsymbol{X'X})^{+}(\boldsymbol{X'Y} - \boldsymbol{\lambda wc}) \qquad \textbf{...(11)}$$

إذْ إنّ $\mathbf{0 < \lambda < 1}$ ويَتُمّ اقتراح قيمها من قبل الباحث [3].

4- طريقة انحدار Lasso البيانية الممكنة ***Feasible Graphical Lasso***

تَـمّ اقتـراح الطريقـة مـن قبـل TAN وآخـرون [7] فـي عـام 2018. ويعتمـد علـى تطـوير مقـدّر المربعـات الصـغرى العامّـة GLS، الـذي تَـمّ تقديمـه أساسـا مـن قبـل Zellner [8] فبـالعودة إلـى النمـوذج (1) يكون مقدّر GLS بالصيغة الآتية [5] :

$$\boldsymbol{\hat{\beta}_{FGLS}} = (\boldsymbol{X'\Omega X})^{-1}\boldsymbol{X'\Omega Y} \qquad \textbf{...(12)}$$

إذْ إنّ مصفوفة Ω تَمّ تعريفها من قبل Zellner بالصورة الآتية [12]

$$\Omega = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{n-3} \\ \rho^3 & \rho^2 & \rho & 1 & \dots & \rho^{n-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \dots & 1 \end{bmatrix}$$

يمثّل ρ معامل الارتباط الخطّي العام. تَمّ اقتراح تطوير للمقدّر الموجود للمعادلة (12)، كما يأتي: وأُطلق عليه تسمية مقدّر المربعات الصغرى الممكن بالصيغة الآتية:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{FGLS}} = (\boldsymbol{X'\Sigma^{-1}X})^{-1}\boldsymbol{X'\Sigma^{-1}Y} \qquad \textbf{... (13)}$$

إذْ إنّ

$$\boldsymbol{\Sigma} = \frac{\mathbf{1}}{\boldsymbol{n}}\,\boldsymbol{\acute{\varepsilon}\,\varepsilon}$$

إذْ إنّ ε يمثّل المتّجه الخطأ العشوائي

ويكون المقدّر في المعادلة (13) متحيّزا وغير مستقرّ في حالة الأبعاد الكبيرة. وهنا اقترح TAN

وآخرون تعديلا على المقدّر في المعادلة (13) وكما يأتي [11] :

$$\boldsymbol{\beta_{FGLasso}} = \left(\boldsymbol{X'\widehat{\Omega}\Sigma X}\right)^{-1}\boldsymbol{X'\widehat{\Omega}\Sigma Y} \qquad \textbf{... (14)}$$

إذْ إنّ

$$\boldsymbol{\widehat{\Omega} = tr(\Omega\Sigma) - \log\det(\Omega) + \lambda\, tr(\Omega\acute{\Omega})} \qquad \textbf{... (15)}$$

وأطلق على المقدّر في المعادلة (14) بمقدّر Lasso البياني الممكن، ويكون مناسبا للعمل تحت شروط البيانات الكبيرة، عندما يكون عدد المتغيّرات تحت الدراسة أكبر من حجم العيّنة، إذْ إنّ $\boldsymbol{0 < \lambda < 1}$ ، ويَتُمّ اقتراح قيمها من قبل الباحث.

5- المحاكاة:

تَمّ هناك توليد بيانات بمختلف أحجام العينات n=10, 20, 30, 40, 50, 60, 70, 80. و كذلك تَمّ اختيار ثلاث قيم لمَعْلَمَة الجزاء $\lambda = 0.3\,, 0.6\,, 0.9$. تَمَّ توليد البيانات بالاعتماد على مقاييس بيانات حقيقية بطريقة التحويل المعكوس حول

التلوّث، إذْ يمثّل Y نسبة التلوّث العامّة، و المصفوفة X تمثّل مجموعة متغيّرات بعدد 80 تمثّل الملوّثات بأنواعها، وتَمَّت المقارنة بين الطريقتين بالاعتماد على متوسّط مربعات الخطأ MSE بالصيغة الآتية[11] .

$$MSE = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2/n \qquad \dots (16)$$

إذْ كانت نتائج المحاكاة كالآتي:

الجدول رقم (1)

متوسّط مربّعات الخطأ للطريقتين P=80

	$\lambda = 0.3$		$\lambda = 0.6$		$\lambda = 0.9$	
n	**SICA**	**FGLasso**	**SICA**	**FGLasso**	**SICA**	**FGLasso**
10	**27.45**	**33.24**	**23.45**	**30.11**	**23.34**	**29.43**
20	**22.01**	**21.90**	**21.90**	**21.88**	**21.33**	**21.55**
30	**20.13**	**18.92**	**20.00**	**18.65**	**19.30**	**18.33**
40	**19.10**	**17.81**	**18.80**	**17.23**	**18.39**	**17.11**
50	**18.02**	**17.93**	**17.77**	**16.22**	**17.50**	**16.01**
60	**17.76**	**15.00**	**17.04**	**14.89**	**16.88**	**14.66**
70	**17.05**	**14.11**	**16.80**	**14.00**	**16.13**	**13.55**
80	**16.51**	**12.83**	**16.03**	**12.54**	**15.92**	**12.23**

من ملاحظة الجدول رقم (1) عندما يكون حجم العيّنة 10 (صغير بالنسبة إلى عدد المتغيّرات) نلاحظ أنّ طريقة SICA كانت الأفضل، واستمرّت أفضليتها حتى حجم عيّنة مساوٍ إلى 20 ، عندما $\lambda = 0.3$. وعندما يصل حجم العيّنة إلى 30 نلاحظ تفوّق طفيف لمقدّر FGLasso . ويستمرّ هذا التفوّق البسيط إلى حجم عيّنة 50. وعند الوصول إلى حجم عيّنة 60 نلاحظ أفضلية واضحة لمقدّر FGLasso إلى حين يساوي حجم العيّنة عدد المتغيّرات، وكانت الأفضلية للمقدّرات عندما تكون مَعْلَمَة الجزاء أقرب إلى الواحد.

5- البيانات الحقيقية:

تَمَّ الحصول على عيّنة بحجم 40 تمثّل كُلّ مشاهدة فيها 10 سنتمتر مكعب من مياه نهر دجلة، تَمّ فحصها في مختبرات دائرة مياه واسط. وتَمَّ تسجيل 80 ملوّثا تمثّل الأنواع المختلفة من الملوّثات، كما هي مفصلة في الملحق وكانت النتائج كالآتي:

الجدول رقم (2)

قيمة متوسّط مربّعات الخطأ للبيانات الحقيقية n=40 , P=80

	SICA	**FGLasso**
$\boldsymbol{\lambda = 0.3}$	**37.67**	**29.56**
$\boldsymbol{\lambda = 0.6}$	**33.11**	**23.87**
$\boldsymbol{\lambda = 0.9}$	**28.32**	**19.56**

من الجدول رقم (2) يتبيّن لدينا أفضلية مقدّر **FGLasso** على مقدّر **SICA** بالنسبة للبيانات الحقيقية، و بالذات عندما تكون مَعْلَمَة الجزء أكبر ما يمكن.

6- التوصيات:

من خلال ما تَمَّ عرضه من نتائج المحاكاة والبيانات الحقيقية نوصي باعتماد مقدّر **FGLasso** لتقدير معالم نموذج الانحدار العام عالي الأبعاد بالذات، عندما يكون حجم العيّنة صغير نسبيا أو قريب إلى حجم العيّنة. كذلك نوصي باعتماد مقدّر **SICA** عندما يكون حجم العيّنة صغير جدًا نسبة إلى عدد المتغيّرات.

References.

[1]. Buhlmann. P, Van De Geer. S, " *Statistics for High-Dimensional Data: Methods, Theory and Applications*" Springer, 2011.

[2]. Fan. Y, Lv. J, "Sure Independence Screening for Ultra-High Dimensional Feature Space" Royal Statistical Society, Vol. 70, Issue. 5, pp. 849-911, 2008.

[3]. Hoerl. A, Kennard. R, "Ridge Regression: Biased Estimation for Nonorthogonal Problems" *Technometrics*, Vol. 12, No. 1, pp. 55-67, 1970.

[4]. James. W, Stein. C, "Estimation with Quadratic loss" Proceedings 4th Berkeley Symposium, pp. 361-379, 1961.

[5]. Khorshed. E, Abood, S, "Comparison between the Methods of Ridge Regression and Liu Type to Estimate the Parameters of the Negative Binomial Regression Model Under Multicollinearity Problem by Using Simulation" *Journal of Economics and administrative Science,* Vol. 24, No. 109, pp. 515-534, 2018.

[6]. Mohammed. L, Khadhm. S, "Estimate Kernel Ridge Regression Function in Multiple Regression" *Journal of Economics and administrative Science,* Vol. 24, No. 103, pp. 411-419, 2018.

[7]. Nikolova. M, "Local Strong Homogeneity of Regularized Estimator" SIAM Journal of Applied Mathematics, Vol. 61, No. 1, pp. 633-658, 2000.

[8]. Pesaran. M, Smith. R, "Signs of Impact Effects in Time Series Regression Models" Economics Letters, Vol. 122, pp. 150-153, 2014.

[9]. Salih. A, Hmood. M, "Analyzing big data sets by using different panelized regression methods with

application: surveys of multidimensional poverty in Iraq" Periodicals of Engineering and Natural Sciences.

Vol. 8, No. 2, pp. 991-999, 2020.

[10]. Salih. A, Hmood. M, "Big Data Analysis by Using One Covariate at a Time Multiple Testing (OCMT) Method: Early School Dropout in Iraq" Int. J. Nonlinear Appl. Issue. 12, No. 2, pp. 931-938, 2021.

[11]. Van Der Vaart. A, "*Asymptotic Statistics*" Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics, 1998.

[12]. Xu. K, "A new nonparametric test for high-dimensional regression Coefficients' Journal of Statistical Computation and Simulation, Vol. 5, pp. 855-867, 2017.

Appendix

Variables

no	Variable	no	Variable	No	Variable
Y	**CPR**	**29**	**DO**	**57**	**^{137}Cs**
1	**SO_3**	**30**	**CODMn**	**58**	**^{90}Sr**
2	**SO_2**	**31**	**COD**	**59**	**Coliforms**
3	**Ph**	**31**	**BOD**	**60**	**Streptococci bacteria**
4	**Turbidity** (NTU)	**32**	**NH4-N**	**61**	**Phytoplankton**
5	**Electrical conductivity**	**33**	**TP**	**62**	**Zooplankton,**
6	**Alkalinity ($CaCO_3$)**	**34**	**TN**	**63**	**Zoobenthos,**
7	**Hardness ($CaCO_3$)**	**35**	**TCu**	**64**	**Macrophytes**
8	**Ammonia ($N\text{-}NH_3$)**	**36**	**TZn**	**65**	**PbH**
9	**Nitrate ($N\text{-}NO_3^-$)**	**37**	**F−**	**66**	**H_2S_2**
10	**Apparent colors (HU)**	**38**	**TSe**	**67**	**AgS_2**
11	**Nitrite ($N\text{-}NO_2^-$)**	**39**	**TAs**	**68**	**SO_2**
12	**Phosphorus (P)**	**40**	**THg**	**69**	**SO_4**
13	**Fluoride (F^-)**	**41**	**TCd**	**70**	**ESP**
14	**Iron (Fe)**	**42**	**Cr6+**	**71**	**SAR**

15	**Escherichia *coli***	**43**	**TPb**	**72**	**Hco3**
16	**PbO**	**44**	**TCN**	**73**	**Cl**
17	**CdO_3**	**45**	**V-ArOH**	**74**	**EC**
18	**HgO_2**	**46**	**Petroleum**	**75**	**P**
19	**AgO**	**47**	**Anionic surfactant**	**76**	**CIS**
20	**MgO**	**48**	**$S2^-$**	**77**	**CO**
21	**KO**	**49**	**Sulphate**	**78**	**CO_3**
22	**Dissolved Metals**	**50**	**Mercury**	**79**	**KCl**
23	**Particulate Metals**	**51**	**Copper**	**80**	**MgCl**
24	**Exchangeable Metals**	**52**	**Zinc**		
25	**Residual Metals**	**53**	**PCB**	**-**	
26	**Essential Metals**	**54**	**HCH**	**-**	
27	**Dissolved oxygen**	**55**	**PAH**	**-**	
28	**Biochemical Oxygen**	**56**	**MgS**	**-**	