Using Some Robust Methods For Handling the Problem of Multicollinearity

استعمال بعض الطرائق الحصينة في معالجة مشكلة التعدد الخطي

أ.م. غفران اسماعيل كمال / كلية الادارة والاقتصاد / جامعة بغداد ghufran62@gmail.com
الباحث / سيف الامام سعدي خزعل / كلية الادارة والاقتصاد / جامعة بغداد
saif.alemams@gmail.com

OPEN ACCESS
P - ISSN 2518 - 5764
E - ISSN 2227 - 703X

Received:27/11/2018
Accepted :17/12/2018

المستخلص

يعد أنموذج الانحدار الخطي المتعدد من نماذج الانحدار المهمة التي اجتذبت العديد من الباحثين في مجالات مختلفة منها الرياضيات التطبيقية والاعمال والطب والعلوم الاجتماعية ، ان نماذج الانحدار الخطية التي تتضمن عدد كبير من المتغيرات التوضيحية تكون ذات اداء ضعيف بسبب كبر التباين فضلا عن ذلك تؤدي الى استنتاجات غير دقيقة ، ان احدى المشاكل المهمة في تحليل الانحدار مشكلة تعدد العلاقة الخطية حيث تعتبر واحده من اهم المشاكل التي اصبحت معروفة لدى العديد من الباحثين وكذلك تأثيراتها على أنموذج الانحدار الخطي المتعدد الى جانب تعدد العلاقة الخطية مشكلة القيم الشاذة في البيانات التي تعتبر احدى الصعوبات في بناء أنموذج الانحدار ، مما يؤدي الى تغيرات عكسية عند اتخاذ الانحدار الخطي كأساس لأجراء اختبارات الفروض .

نستعرض في هذا البحث بعض الطرائق الحصينة لتقدير معلمات أنموذج الانحدار الخطي المتعدد وهي طريقة انحدار الحرف بالاعتماد على مقدر المربعات الصغرى المشذبة (Ridge-LTS) وطريقة (Liu) بالاعتماد على مقدر المربعات الصغرى المشذبة (Liu-LTS) , ومن خلال استخدام المحاكاة تمت اجراء المقارنة بين هاتين الطريقتين وفق معيار المقارنة متوسط مربعات الخطأ (MSE) ، واتضح من خلال المقارنة ان طريقة (Liu-LTS) هي الافضل في تقدير معلمات أنموذج الانحدار الخطي المتعدد .

المصطلحات الرئيسة للبحث / الانحدار الخطي المتعدد ، التعدد الخطي ، القيم الشاذة ، مقدر **LTS** ، مقدر **Liu** ، انحدار الحرف .

Journal of Economics and Administrative Sciences
2019; Vol. 25, No.112
Pages:500- 514

* بحث مستل من رسالة ماجستير .

1-المقدمة Introduction

يعد الانحدار الخطي من الاساليب الاحصائية المتقدمة التي تضمن دقة الاستدلال من اجل تحسين نتائج البحث عن طريق الاستخدام الامثل للبيانات ، ان انعدام الاستقلالية بين المتغيرات التوضيحية تؤدي لعلاقة خطية متداخلة (بين متغيرين) او تعدد لعلاقة خطية (أكثر من متغيرين) فأن تطبيق طريقة المربعات الصغرى الاعتيادية (Ordinary Least Squares) (OLS) تؤدي الى تضخم في تباينات لمعاملات الانحدار المقدرة وبالتالي تظهر النتائج غير دقيقة أي لا تحقق خاصية أفضل تقدير خطي غير متحيز (Linear Unbiased Best Estimator) (BLUE) ، في الواقع التطبيقي تحصل مشكلة تعدد العلاقة الخطية (Multicollinearity Problem) عندما يرتبط اثنان او اكثر من المتغيرات التوضيحية بعلاقة قوية بحيث يصعب فصل اثر كل متغير على المتغير المعتمد (Mohammed, 2016:PP 48-49) [8].

بشكل عام تنقسم مشكلة تعدد العلاقة الخطية الى نوعين الاول يعرف بالتعدد الخطي التام (Perfect Multicollinearity) وفي مثل هذه الحالة يكون محدد مصفوفة المعلومات او مصفوفة فيشر (Fisher Matrix) مساوياً للصفر بعبارة اخرى ($|X'X| = 0$) ويترتب على ذلك استحالة ايجاد مقدرات معالم الأنموذج الخطي العام وابرز طراق المعالجة في مثل هذه الحالة هو استخدام اسلوب المركبات الرئيسية (Principle Components) أما النوع الثاني وهو ما يثير اهتمامنا في هذا البحث فيسمى بالتعدد الخطي شبة التام (Semi Perfect Multicollinearity) وفيه تكون محدد مصفوفة المعلومات صغيرة جداً وعندها تكون المعالم المقدرة ذات تباين كبير جداً وابرز طرق معالجته هو باستخدام مقدر انحدار الحرف (Ridge Regression) (RR) الذي اقترحه كل من (Horal and Kennard) عام (1970 م) [III] (كاظم ومسلم ،2002 : ص190) ، كذلك مقدر (Liu) الذي اقترح من قبل الباحث (Kejian.L) في عام (1993 م) وغيرها ، الى جانب تعدد الخطية تعتبر القيم الشاذة مشكلة شائعة اخرى في تحليل الانحدار ، في الاحصاء تعرف القيم الشاذة بأنها مجموعة من المشاهدات الغير اعتيادية التي تختلف اختلافاً جوهرياً عن باقي البيانات ، ايضاً هنالك عدد كبير من التعاريف لما تعنيه القيم الشاذة نذكر منها للباحثان (Rousseeum & Leroy) في عام (1987) (بأنها مشاهدات في الانحدار الخطي التي تنحرف عن الجزء الاكبر من البيانات) ، للتغلب على تلك المشاكل مجتمعة يتم استعمال طرائق بديلة عن الطرائق التقليدية تكون اكثر كفاءة في التقدير وتدعى بـــ (طرائق التقدير الحصينة) (Robust estimation methods) حيث تتصف انها قليلة الحساسية تجاه الشواذ اذ يتم الحصول عليها من خلال مقدرات حصينة تمتلك كفاءة عالية . (Hekimoglu & Erenoglu , 2013:PP419-421) [4]

2-مشكلة البحث The problem of the research

ان مشكلة البحث تكمن في تقدير معلمات انموذج الانحدار الخطي المتعدد (MLR) في ظل وجود مشكلة تعدد العلاقة الخطية والقيم الشاذة في البيانات ، حيث ان وجودها يؤثر بشكل سلبي على دقة نتائج التحليل ، لذا يلجأ العديد من الباحثين الى استعمال الطرائق الحصينة لمعالجة تلك المشاكل وبالتالي الحصول على مقدرات أكثر كفاءة واقل حساسية في التقدير .

3- هدف البحث The Aim of the Research

يهدف البحث الى المقارنة بين الطرائق الحصينة لأنموذج الانحدار الخطي المتعدد(MLR) في ظل وجود مشاكل تعدد العلاقة الخطية والقيم الشاذة في البيانات من خلال المقارنة بين المقدرات الاخرى وبالتالي الحصول على افضل تقدير .

4- الجانب النظري

يستند أنموذج الخطي العام على افتراض وجود علاقة خطية ما بين المتغير المعتمد (Y_i) وعدد من المتغيرات المستقلة ($X's$) .

$X_1, X_2, X_3, \dots, X_p$

وحد عشوائي (ε_i) ويعبر عن العلاقة بالنسبة لــ (n) من المشاهدات ، و (P) من المتغيرات وفقًا للمعادلة :

$$y_i = \beta_0+\beta_1 X_{iP}+\beta_2 X_{iP} + \cdots + \beta_P X_{iP} + \varepsilon_i \qquad \dots (1)$$

i =1 , 2 , … , n , j =1 , 2 , … ,P

وبدلالة المصفوفات يمكن صياغة أنموذج الخطي العام كما في الشكل الاتي :

$$y = XB + \varepsilon \qquad \dots (2)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1P} \\ 1 & X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2P} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{i1} & X_{i2} & \dots & X_{ij} & \cdots & X_{iP} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nj} & \dots & X_{nP} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_P \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Y : موجه مشاهدات المتغير المعتمد من الدرجة ($n * 1$) .

X : مصفوفة من الدرجة (($n * (P+1)$) وتمثل مشاهدات المتغيرات المستقلة علماً بأن العمود الاول من هذه المصفوفة يمثل الحد الثابت .

β : موجه المعالم المطلوب تقديرها من الدرجة $(P+1) * 1$.

P : عدد المتغيرات المستقلة .

n : عدد المشاهدات .

ε : موجه الاخطاء العشوائية من الدرجة ($n * 1$) .

[3] (كاظم ومسلم ، 2002 : ص 50)

سيتم التطرق في هذا البحث لآهم المقدرات الحصينة شيوعًا لتقدير معلمات أنموذج الانحدار الخطي المتعدد(MLR) وهي كالاتي :

5-مقدر المربعات الصغرى المشذبة (LTS)

Least Trimmed squares estimator

طريقة إحصائية تستخدم في تقدير معلمات أنموذج الانحدار الخطي المتعدد (MLR)، حيث اقترحت من قبل الباحث (Rousseew) عام (1984 م) ، ان مقدر (LTS) يلعب دوراً مهماً في حساب طرائق التقدير الحصينة كونه لا يتأثر بوجود القيم الشاذة وايضاً يحقق نقطة انهيار عالية (BP = 0,5) ، ان المقدر الناتج من هذه الطريقة يدعى بمقدر المربعات الصغرى المشذبة ويرمز له بالرمز (LTS) ، ويتم حسابه وفقاً للصيغة الاتية :-

$$\hat{\beta}_{LTS} = \underset{B}{\arg\min}\ Q_{LTS}(\beta) \qquad \dots (3)$$

حيث ان :

$$Q_{LTS}(B) = \sum_{i=1}^{h} e_{(i)}^{2}$$

h : ثابت وشروطه ($\frac{n}{2} < h < n$) .

$$e^2_{(1)} \leq e^2_{(2)} \leq \cdots \leq e^2_{(n)}$$

تمثل مربعات البواقي المرتبة من اقل قيمة الى اعلى قيمة .

[2] (Alma,2011:P413) [

$$e_i^2 = \frac{(y_i - \hat{B}' X_i)^2}{1 + \hat{B}'\hat{B}} \quad , \quad i = 1,2,\ldots,n \qquad \ldots(4)$$

[5] (Jung, 1978:P332)

ان h هو عدد النقاط البيانات الجيدة (الغير مشذبه) والتي تكون مستبعدة في المجموع ، حيث أن المقدر يعطي تقدير حصين من خلال تعريف $n - h$ من النقاط والتي يكون لها أكبر البواقي كالشواذ ، مما يسمح باستبعاد هذه النقاط من مجموعة البيانات بشكل كامل اعتماد على قيمة h التي تكون قريبة جدًا من نقاط البيانات الجيدة وذلك لان أعلى عدد من النقاط الجيدة يستعمل في التقدير، وفي هذه الحالة فأن مقدر LTS سوف يعطي افضل تقدير ممكن .

أن هذا المقدر يكون مكافئاً حسابياً لطريقة المربعات الصغرى الاعتيادية (OLS) اذا تم استبعاد نقاط البيانات الغير طبيعية بصورة دقيقة ومع ذلك اذا كان هناك نقاط بيانات اكثر من التي تم استبعادها او تشذيبها فأنها غير كفوءة .

[2] (Alma,2011:P413)

لحساب مقدر LTS هنالك العديد من الخوارزميات وأحدى أهم الخوارزميات تدعى (C-steps) التي اقترحت من قبل الباحثان(Rousseeuw and Van Driessen) عام (2006 م) ، من أجل تطبيق فكرة الخوارزمية فيما يلي عدة خطوات :-

1- يتم اختيار مجموعه جزئية أولية بحجم $\mathrm{H_{old}}$

2- يتم حساب مقدر المربعات الصغرى الاعتيادية$\hat{\mathrm{B}}_{\mathrm{old}}$ وبالاعتماد على المجموعة الجزئية $\mathrm{H_{old}}$

3- يتم حساب البواقي $e_{old}(i)$ لكل i=1,2…,n

4- ترتيب القيم المطلقة تصاعدياً من اقل قيمة الى اعلى قيمة بحيث تعطى توافيق π ل

$$|e_{old}(\pi(1))| \leq |e_{old}(\pi(2))| \leq \cdots \leq |e_{old}(\pi(n))|$$

5- يتم ايجاد مجموعة جديدة $\mathrm{H_{new}} = \{\pi(1), \pi(2), \ldots, \pi(h)\}$.

6- يتم ايجاد مقدر (LTS) $\hat{\beta}_{new}$ والتي يكون مساوياً الى مقدر المربعات الصغرى وبالاعتماد على المجموعة الجزئية$\mathrm{H_{new}}$.

[9](Rousseeuw & Driessen, 2006:P32)

6-طرائق التقدير:

لتقدير معلمات أنموذج الانحدار الخطي المتعدد سيتم الاعتماد في هذا البحث على بعض الطرائق الحصينة وكالاتي :-

6-1- طريقة انحدار الحرف بالأعتماد على مقدر المربعات الصغرى المشذبة

Robust Ridge method for based on the LTS estimator

هي أحدى الطرائق الحصينة لمقاومة مشاكل تعدد العلاقة الخطية والقيم الشاذة بين المتغيرات التوضيحية ، أن اسلوب انحدار الحرف (Ridge Regression) (RR) يعتبر أحد بدائل طرائق التقدير عندما توجد هناك مشكلة التعدد الخطي شبه التام ، حيث قدمت لأول مرة عام (1970 م) من قبل الباحثان (Horal & Kennard) حيث تم استعمالها لتقدير معلمات أنموذج الانحدار الخطي المتعدد من خلال معالجة التأثيرات الغير مرغوب بها باستخدام اسلوب انحدار الحرف (RR) بدلا من طريقة (OLS) .

[3](El-Dereny & Rashwan,2011:P589)

لإزالة تأثير تعدد العلاقة الخطية تتم أضافة كمية صغيرة موجبة k إلى العناصر القطرية لمصفوفة المعلومات ($X'X$) أي ($k > 0$) ، وقد أفترض في هذا المقدر وجود مصفوفة γ ذات درجة ($p \times p$) وأعمدتها تمثل المتجهات المميزة ($q_1 q_2, q_3 \dots, q_p$) للمصفوفة ($X'X$) ومن هنا فأن :

$$\gamma'(X'X)\gamma = \Lambda = diag\ (\lambda_1, \lambda_1, \dots \lambda_p)$$

$$Z = X\gamma$$

$$\alpha = \gamma'\beta$$

Z : مصفوفة (n*P) ثابتة .

α : موجه (P*1) المعلمات .

γ : تمثل مصفوفة متعامدة ، أعمدتها تمثل المتجهات المميزة المقابلة للجذور المميزة لمصفوفة المعلومات ($X'X$) .

$\gamma'\gamma = \gamma'\gamma = I$ **Where**

وعليه فأن أنموذج الانحدار الخطي العام يمكن اعادة كتابته بالشكل التالي :

$$Y = Z\alpha + u$$

ان تقدير المربعات الصغر الاعتيادية لـ α يعطي كالاتي :

$$\hat{\alpha}_{LS} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y \qquad \dots (5)$$

علماً إن معاملات الانحدار الأصلية لــ (OLS) هي :

$$\hat{\beta}_{LS} = \gamma\hat{\alpha}_{LS}$$

وأن مقدرات الحرف الاعتيادية (Ordinary Ridge Regression) (ORR) لـ α تعطى بالشكل :-

(Kan, et.al, 2013:PP646-645) [6]

$$\hat{\alpha}_{ORR} = (Z'Z + kI)^{-1}Z'y \quad , \qquad k > 0\ , \quad k_1 = k_2 = \dots = k_p = k$$

$$= (\Lambda + kI)^{-1}Z'y \qquad \dots (6)$$

$$= B^{-1}Z'y \quad , \qquad B = (\Lambda + kI)$$

$$= (I - kB^{-1})\,\hat{\alpha}_{LS}$$

$$\hat{B}_{ORR} = \gamma\hat{\alpha}_{ORR}$$

[1] **(Alguraibawi, et.al, 2015:P308)**

اذ ان :

$$k = P\sigma^2/\hat{\alpha}'_{LS}\,\hat{\alpha}_{LS} \qquad \dots (7)$$

P : عدد المتغيرات التوضيحية .

σ^2 : متوسط مربعات الفروق بين القيم الحقيقة والتقديرية للمتغير التابع .

(Kan, et.al, 2013:PP646-645) [6]

وفي حالة كون $k = 0$ فانه مقدرات الحرف الاعتيادي (ORR) تتحول إلى مقدرات المربعات الصغرى الاعتيادية (OLS) .

$$\hat{\alpha}_{ORR}(k) = \frac{\lambda_i}{\lambda_i + k}\hat{\alpha}_{LS}$$
$$\hat{\alpha}_{ORR}(0) = \hat{\alpha}_{LS}$$

[7] (Kejian, 1993:P395)

وأن مقدر الحرف الاعتيادي (ORR) هو متحيز للمعلمة الاصلية ، وأن مقدار التحيز هو :

$$Bias\,(\hat{\alpha}_{ORR}(k)) = kB^{-1}\,\hat{\alpha} \qquad \ldots (8)$$

ومصفوفة التباين لمقدر (ORR) بالشكل الاتي :

$$\mathrm{Var}\,(\hat{\alpha}_{ORR}(k)) = \sigma^2(\mathrm{I} - kB^{-1})\,\Lambda^{-1}(\mathrm{I} - kB^{-1})' \qquad \ldots (9)$$

ومصفوفة متوسط مربعات الخطأ لمقدر (ORR) بالشكل الاتي :

$$\mathrm{MSE}\,(\hat{\alpha}_{ORR}(k)) = \mathrm{Var}\,(\hat{\alpha}_{ORR}(k)) + [\mathrm{Bias}(\hat{\alpha}_{ORR}(k))][\mathrm{Bias}\,(\hat{\alpha}_{ORR}(k))]'$$
$$= \sigma^2(\mathrm{I} - kB^{-1})\,\Lambda^{-1}(\mathrm{I} - kB^{-1})' + k^2B^{-1}\,\hat{\alpha}_{OLS}\hat{\alpha}'_{OLS}B^{-1} \qquad (10)$$

[1] (Alguraibawi, et.al, 2015:P308)

بالرغم من أن مقدر ($\hat{\alpha}_{ORR}$) غالباً ما يستخدم لمعالجة مشكلة تعدد العلاقة الخطية الا انه لا يمتلك حصانة لمقاومة القيم الشاذة بين المتغيرات التوضيحية ، للتغلب على هذه المشكلة أقترح عام(2012 م) كل من (Kan ,et.al) مقدر انحدار الحرف الحصين بالاعتماد على طريقة المربعات الصغرى المشذبة في تقدير معلمات الأنموذج الانحدار الخطي (Generalized Linear Model) (GLM) والذي يرمز له بالرمز ($\hat{\alpha}_{LTS-RIDGE}$) وصيغته كالأتي :

$$\hat{\alpha}_{LTS} \qquad \ldots (11) = \left(\mathrm{Z}'\mathrm{Z} + k_{LTS}\right)^{-1}\mathrm{Z}'\mathrm{Z}\,\hat{\alpha}_{LTS-RIDGE}$$

حيث أن :

$$k_{LTS} = p\sigma^2/\alpha'_{LTS}\alpha_{LTS} \qquad \ldots (12)$$

p : عدد المتغيرات التوضيحية .

σ^2 ، α_{LTS} : يتم الحصول عليهما من حل طريقة المربعات الصغرى المشذبة (LTS) .

[6] (Kan, et.al, 2013: P647)

6-2 **طريقة** (Liu) **الحصين بالاعتماد على طريقة المربعات الصغرى المشذبة** :

Robust Liu estimator for based on the method LTS

لقد اقترحت عدة أساليب أو طرائق إحصائية لتقدير معلمات الأنموذج الاحصائي بحيث أصبحت طرائق التقدير كثيرة جداً، وأن الهدف من عملية التقدير هو الحصول على افضل وأكفأ النتائج في محاولة لتمثيل المجتمع تمثيلاً جيداً ، ان الطريقة المستخدمة تعد من الطرائق الحصينة لمعالجة مشكلة تعدد العلاقة الخطية والقيم الشاذة بين المتغيرات التوضيحية ، قام الباحث (Liu) في عام (1993 م) باقتراح مقدر جديد لمعالجة مشكلة تعدد العلاقة الخطية إذ قام بدمج مميزات مقدر انحدار الحرف الاعتيادي (ORR) ومقدر (Stein) عام (1956 م) بمقدر خاص أطلق عليه مقدر (LE) ، حيث يمتاز مقدر (ORR) كونه مؤثر في التطبيق العملي ولكنه داله معقدة في بـ K ، وايضا بالنسبة لمقدر (Stein) فمن مميزاته يكون دالة خطية في C ولكن التقليص (Shrinkage) هو نفسه لكل عنصر من عناصر α_S .

$$\hat{\alpha}_S = c\hat{\alpha}_{LS} \quad , \quad 0 < c < 1$$

لذلك تكون الافضلية لمقدر (LE) يتغلب بها على مقدر (ORR) حيث ان LE دالة خطية بمعلمة التحيز d لذلك يكون من السهل حسابها أكثر من معلمة الحرف k لمقدر (ORR) ، ويرمز لمقدر (LE) بالرمز ($\hat{\alpha}_{LE}$) .

$$\hat{\alpha}_{LE} = (Z'Z_{+}\mathrm{I})^{-1}(Z'Y + d\hat{\alpha}_{LS}) \quad , \ 0 < d < 1 \qquad \dots (13)$$

$$d_1 = d_2 = \cdots = d_p = d$$

ويمكن كتابتها كما يلي :

$$= (\Lambda_{+}\mathrm{I})^{-1}(\Lambda + d\mathrm{I})\hat{\alpha}_{LS}$$

Λ : مصفوفة متعامدة ، اعمدتها تمثل المتجهات المميزة المقابلة للجذور المميزة لمصفوفة المعلومات .

(Kejian, 1993:P395) [7]

وان :

$$d = 1 - \sigma^2 \left[\frac{\sum_{i=1}^{p} 1/\lambda_i (1 + \lambda_i)}{\sum_{i=1}^{p} \beta_i^2 / (\lambda_i + 1)^2} \right] \qquad ..(14)$$

(Kan, et.al, 2013: P646) [6]

والتوقع لمقدر $\hat{\alpha}_{LE}$ بالشكل الأتي :

$$\mathrm{E}\hat{\alpha}_{LE} = (\Lambda_{+}\mathrm{I})^{-1}(\Lambda + d\mathrm{I})E\,\hat{\alpha}_{LS} \qquad \dots (15)$$

$$= (\Lambda_{+}\mathrm{I})^{-1}(\Lambda + d\mathrm{I})\,\alpha$$

وأن مقدر(LE) متحيز بالنسبة للمعلمة α ومقدار التحيز هو :

$$Bias\,(\hat{\alpha}_{LE}) = E(\hat{\alpha}_{LE} - \alpha) \qquad \dots (16)$$

$$= -(\Lambda_{+}\mathrm{I})^{-1}(\Lambda + d\mathrm{I})\,\alpha$$

ومصفوفة التباين لمقدر (LE) بالشكل الاتي :

$$\mathrm{Var}(\hat{\alpha}_{LE}) = \sigma^2 (\Lambda_{+}\mathrm{I})^{-1}(\Lambda + d\mathrm{I})\,\Lambda^{-1}(\Lambda_{+}\mathrm{I})^{-1}(\Lambda + d\mathrm{I})'$$

$$= \sigma^2 (1 - \mathrm{M})\Lambda^{-1}(1 - \mathrm{M})' \qquad \dots (17)$$

اذ ان :

$$\mathrm{M} = (\Lambda_{+}\mathrm{I})^{-1}(\Lambda + d\mathrm{I})$$

ومصفوفة متوسط مربعات الخطأ لمقدر (LE) بالصيغة الاتية :

$$MSE\,(\hat{\alpha}_{LE}) = Var(\hat{\alpha}_{LE}) + \left(Bias(\hat{\alpha}_{LE})\right)^2$$

$$= \sigma^2 (1 - \mathrm{M})\Lambda^{-1}(1 - \mathrm{M})' + \mathrm{M}\,\alpha\alpha'\,\mathrm{M}' \qquad \dots (18)$$

[7] **(Kejian, 1993:P395)**

بالاعتماد على طريقة المربعات الصغرى المشذبة (LTS) التي تعتبر من أكثر الطرائق الحصينة شيوعا في تقدير معلمات أنموذج الانحدار الخطي (GLM) ، وبالتالي تكون الصيغة النهائية لمقدر ($\hat{\alpha}_{LTS-LIU}$) وفقا للشكل الاتي :-

$$\hat{\alpha}_{LTS-LIU} = (\Lambda_{+}\mathrm{I})^{-1}(\Lambda + d_{LTS}\mathrm{I})\hat{\alpha}_{LTS} \qquad \dots (19)$$

اذ ان :

$$d_{LTS} = 1 - \sigma^2\left[\frac{\sum_{i=1}^{p}\frac{1}{\lambda_i(1+\lambda_i)}}{\sum_{i=1}^{p}\beta_i^2/(\lambda_i+1)^2}\right] \qquad \dots (20)$$

β , σ^2 يتم الحصول عليهما من حل طريقة المربعات الصغرى المشذبة (LTS) .
[6] (Kan, et.al, 2013:P648)

7- الجانب التجريبي

7-1-مفهوم المحاكاة The Concept of Simulation

يعد اسلوب المحاكاة من الاساليب العلمية الرصينة التي تقوم على اعطاء صورة لظاهرة حقيقية طبق الاصل من أجل محاكاة أكبر قدر من الحالات ليتسنى الاستفادة في دراسة خواص تلك الظاهرة المدروسة [1](الجشعمي ،2007 :ص34) .

غالباً ما يتم اللجوء لأسلوب المحاكاة للتأكد من تحقق نظام حقيقي موجود أصلاً ، أو لصعوبة الحصول على بيانات اللازمة لدراسة ظاهرة معينة ، اي عندما يصعب إثبات البرهان الرياضي بشكل نظري لبيان أفضلية طرائق تقدير معينة على حساب أخرى . (النداوي ،2008 :ص39) [2]

7-2- خطوات اجراء المحاكاة :

لقد تضمنت تجارب المحاكاة لهذا البحث كتابة عدد من البرامج بلغة (R) في توليد البيانات ، حيث يتم تحديد حجوم العينات (n=20, n=50, n=100) ونسب تلوث مختلفة $(\tau = 5\%, 15\%)$ ، وقد تم استخدام الصيغة الاتية في توليد المتغيرات التوضيحية :-

$$X_{ij} = p\, v_{ik} + (1-p^2)^{1/2} v_{ij} \quad , \qquad i = 1,2,\dots,n \;, \qquad j = 1,2,\dots,P$$

v_{ij} : الإعداد العشوائية المولدة والتي تتبع التوزيع الطبيعي القياسي .

v_{ip} : يمثل قيم العمود الاخير من اعمدة المتغيرات المولدة .

P : يمثل عدد المتغيرات المرتبطة .

i : يمثل عدد المشاهدات .

ρ : يمثل قيمة الارتباط بين المتغيرات التوضيحية في الأنموذج المدروس ، الذي أخذا القيم $(\rho = 0.90, 0.95, 0.99)$.

[1] (Alguraibawi, et.al, 2015:P313)

ويكون الأنموذج كالاتي :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i \qquad ..(21)$$

$$i = 1,2,..,n$$

يتم توليد الاخطاء العشوائية وفقاً للتوزيع الطبيعي :

$$\varepsilon_i \sim (0, \sigma^2) \quad , \quad i = 1,2,\dots,n$$

وبالنسبة لقيم الانحراف المعياري فهي ($\sigma = 0.5, 1, 1.5$) .

يتم تحديد القيم الافتراضية للمعلمات وبافتراض أن عدد المعالم هي $P = 6$ وفقاً للقيم التالية :-

$$\beta_1 = 5 \,,\ \beta_2 = 3 \,, \beta_3 = \sqrt{6} \,,\ \beta_4 = 0 \,,\ \beta_5 = 0 \quad,\ \beta_0 = 0$$

[6] (Kan, et.al, 2013:P648)

يتم تقدير معلمات انموذج الانحدار الخطي المتعدد(MLR) وفق طرائق التقدير التي تم عرضها في الجانب النظري من البحث كما يلي :-

1- طريقة انحدار الحرف بالاعتماد على طريقة المربعات الصغرى المشذبة (Ridge-LTS) .

2-طريقة (Liu) بالاعتماد على طريقة المربعات الصغرى المشذبة (Liu-LTS) .

وقد تم الاعتماد على المقياس الاحصائي متوسط مربعات الخطأ(MSE) للأنموذج لمعرفة اي النماذج افضل في تمثيل البيانات وكان عدد مرات تكرار التجربة R=1000 مرة .

7-3-**تحليل نتائج تجربة المحاكاة** :

Analyzing the results of the simulation experiment

جدول(1) تقدير المعلمات و(MSE) للطريقتين ، ولكافة حجوم العينات عندما تكون نسبة التلوث (τ=5%) وقيمة الانحراف المعياري ($\sigma = 1$) .

Coffe	N	Parameters / Methods	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	MSE
$\rho = 0.90$	n=20	Ridge-LTS	0.171744	0.15654	0.14101	0.09151	0.11265	0.01924
		Liu-LTS	0.38649	0.06354	0.02101	0.03967	0.02701	0.02787
	n=50	Ridge-LTS	0.17032	0.13374	0.06402	0.07954	0.07720	0.01114
		Liu-LTS	0.60059	0.17597	-0.0256	-0.05556	-0.03511	0.00969
	n=100	Ridge-LTS	0.025497	0.02321	0.02412	0.02175	0.02164	0.00910
		Liu-LTS	0.24547	-0.0937	-0.1146	-0.00079	0.00901	0.00785
$\rho = 0.95$	n=20	Ridge-LTS	0.21446	0.19820	0.17278	0.09139	0.10694	0.01462
		Liu-LTS	0.49096	0.10576	0.02269	-0.09326	-0.00525	0.02548
	n=50	Ridge-LTS	0.21712	0.16816	0.07106	0.09578	0.08983	0.00924
		Liu-LTS	0.75419	0.32914	-0.1009	-0.14183	-0.09834	0.00674
	n=100	Ridge-LTS	0.03784	0.03538	0.03666	0.03412	0.03407	0.00828
		Liu-LTS	0.26307	-0.1134	-0.1358	-0.00575	0.00947	0.01010
$\rho = 0.99$	n=20	Ridge-LTS	0.24174	0.42204	0.05478	0.03519	0.11265	0.01095
		Liu-LTS	0.55670	0.40624	-0.0740	-0.28193	0.03933	0.01601
	n=50	Ridge-LTS	0.46171	0.30961	-0.0641	0.04163	0.02030	0.00753
		Liu-LTS	1.33403	0.95422	-0.5933	-0.5412	-0.35795	0.00619
	n=100	Ridge-LTS	0.09716	0.09059	0.09592	0.08885	0.08892	0.00572
		Liu-LTS	0.33797	-0.1472	-0.0665	-0.06264	-0.02737	0.00996

جدول(2) تقدير المعلمات و(MSE) للطريقتين، ولكافة حجوم العينات عندما تكون نسبة التلوث (τ=5%) وقيمة الانحراف المعياري ($\sigma = 1.5$) .

Coffe	N	Parameters / Methods	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	MSE
$\rho = 0.90$	n=20	Ridge-LTS	0.14226	0.13212	0.11414	0.06583	0.09786	0.02255
		Liu-LTS	0.341217	-0.0009	-0.0475	-0.04032	0.05295	0.03674
	n=50	Ridge-LTS	0.04339	0.03734	0.03677	0.03251	0.03309	0.01884
		Liu-LTS	0.27863	-0.0757	-0.0594	-0.01806	-0.03051	0.01941
	n=100	Ridge-LTS	0.01224	0.01090	0.01169	0.01038	0.01031	0.00994
		Liu-LTS	0.19116	-0.0758	-0.1162	0.00281	0.01487	0.01028
$\rho = 0.95$	n=20	Ridge-LTS	0.21178	0.19531	0.16001	0.05345	0.08428	0.01759
		Liu-LTS	0.48660	0.07370	-0.0334	-0.13394	0.00394	0.02904
	n=50	Ridge-LTS	0.06618	0.05571	0.05591	0.05067	0.05229	0.01554
		Liu-LTS	0.32515	-0.1026	-0.0498	-0.05399	-0.04779	0.01953
	n=100	Ridge-LTS	0.02342	0.02159	0.02283	0.02101	0.02097	0.00924
		Liu-LTS	0.20786	-0.0925	-0.1396	-0.00300	0.01823	0.01085
$\rho = 0.99$	n=20	Ridge-LTS	0.22664	0.52210	-0.0218	-0.03110	0.10061	0.02063
		Liu-LTS	0.55747	0.51841	-0.1392	-0.40584	0.03578	0.02749
	n=50	Ridge-LTS	0.56447	0.36463	-0.2035	-0.03906	-0.06379	0.01163
		Liu-LTS	1.53288	1.14791	-0.8471	-0.70370	-0.46124	0.00992
	n=100	Ridge-LTS	0.07895	0.07164	0.07853	0.07090	0.07074	0.00741
		Liu-LTS	0.28488	-0.1234	-0.0473	-0.04733	-0.01061	0.01019

جدول (3) تقدير المعلمات و(MSE) للطريقتين ، ولكافة حجوم العينات عندما تكون نسبة التلوث (τ=15%) وقيمة الانحراف المعياري ($\sigma = 1$) .

Coffe	N	Parameters / Methods	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	MSE
$\rho = 0.90$	n=20	Ridge-LTS	0.17174	0.15654	0.14101	0.09151	0.11265	0.01924
		Liu-LTS	0.38649	0.06354	0.02101	-0.03967	0.02701	0.02787
	n=50	Ridge-LTS	0.17032	0.13374	0.06402	0.07954	0.07720	0.01114
		Liu-LTS	0.60059	0.17597	-0.0256	-0.05556	-0.03511	0.00785
	n=100	Ridge-LTS	0.02549	0.02321	0.02412	0.02175	0.02164	0.00910
		Liu-LTS	0.24547	-0.0937	-0.1146	-0.00079	0.00901	0.00969
$\rho = 0.95$	n=20	Ridge-LTS	0.21446	0.19820	0.17278	0.09139	0.10694	0.01322
		Liu-LTS	0.49096	0.10576	0.02269	-0.09326	-0.00525	0.02250
	n=50	Ridge-LTS	0.21712	0.16816	0.07106	0.09578	0.08983	0.00924
		Liu-LTS	0.75419	0.32914	-0.1009	-0.14183	-0.09834	0.00674
	n=100	Ridge-LTS	0.03784	0.03538	0.03666	0.03412	0.03407	0.00828
		Liu-LTS	0.26307	-0.1134	-0.1358	-0.00575	0.00947	0.01010
$\rho = 0.99$	n=20	Ridge-LTS	0.24174	0.42204	0.05478	0.03519	0.11265	0.01095
		Liu-LTS	0.55670	0.40624	-0.0740	-0.28193	0.03933	0.01601
	n=50	Ridge-LTS	0.46171	0.30961	-0.0641	0.04163	0.02030	0.00753
		Liu-LTS	1.33403	0.95422	-0.5933	-0.54120	-0.35795	0.00619
	n=100	Ridge-LTS	0.09716	0.09059	0.09592	0.08885	0.08892	0.00572
		Liu-LTS	0.33797	-0.1472	-0.0665	-0.06264	-0.02737	0.00996

جدول (4) تقدير المعلمات و(MSE) للطريقتين ، ولكافة حجوم العينات عندما تكون نسبة التلوث (τ=15%) وقيمة الانحراف المعياري ($\sigma = 1.5$) .

Coffe	N	Parameters / Methods	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	MSE
$\rho = 0.90$	n=20	Ridge-LTS	0.14268	0.13149	0.11440	0.06647	0.09780	0.02940
		Liu-LTS	0.34068	-0.0025	-0.0487	-0.04049	0.05350	0.04170
	n=50	Ridge-LTS	0.13169	0.10211	0.03596	0.05435	0.05118	0.01537
		Liu-LTS	0.57760	0.11852	-0.0336	-0.03556	-0.01713	0.01170
	n=100	Ridge-LTS	0.01236	0.01105	0.01176	0.01044	0.01039	0.00997
		Liu-LTS	0.19073	-0.0776	-0.1126	0.00127	0.01422	0.01013
$\rho = 0.95$	n=20	Ridge-LTS	0.21197	0.19441	0.15914	0.05486	0.08428	0.02191
		Liu-LTS	0.48487	0.07215	-0.0345	-0.13269	0.00521	0.03395
	n=50	Ridge-LTS	0.19560	0.14653	0.03463	0.06750	0.06008	0.01353
		Liu-LTS	0.78004	0.32345	-0.1541	-0.15910	-0.11002	0.01056
	n=100	Ridge-LTS	0.02350	0.02167	0.02287	0.02102	0.02099	0.00939
		Liu-LTS	0.20821	-0.0927	-0.1380	-0.00318	0.01781	0.01044
$\rho = 0.99$	n=20	Ridge-LTS	0.23735	0.51883	-0.0239	-0.03421	0.09815	0.01982
		Liu-LTS	0.56470	0.51859	-0.1455	-0.40554	0.03421	0.02538
	n=50	Ridge-LTS	0.55861	0.35997	-0.1984	-0.03585	-0.06138	0.01165
		Liu-LTS	1.52058	1.12589	-0.8362	-0.6909	-0.45747	0.00993
	n=100	Ridge-LTS	0.07895	0.07164	0.07853	0.07090	0.07074	0.00741
		Liu-LTS	0.28488	-0.1234	-0.0473	-0.07972	-0.01061	0.01019

4-7 - **مناقشة تجارب المحاكاة في حالة وجود تلوث** (τ = 5% ، 15%)

من خلال النتائج المبينة بالجداول (1) (2) (3) (4) نلاحظ ما يلي :-

1- يتضح لنا ان قيمة متوسط مربعات الخطأ (MSE) للأنموذج تتناقص كلما زاد حجم العينة ولجميع طرائق التقدير المدروسة .

2- في حالة في حالة وجود تلوث (τ =5%) وتوزيع الاخطاء التوزيع الطبيعي بمتوسط (0) وانحراف معياري ($\sigma = 1$) نلاحظ من خلال مقياس المقارنة (MSE) للجدول (1) ان طريقة (Ridge-LTS) افضل طريقة عندما يكون حجم العينة (n=20) وقيمة معامل الارتباط ($p = 0.90$) ، بينما تحقق طريقة (Liu-LTS) اقل قيمة عندما يكون حجم العينة (n=50 , n=100) ، وعندما تكون حجوم العينات (n=50 , n=100) و قيمة معامل الارتباط ($p = 0.95$) ، تحقق طريقة (Ridge-LTS) الأفضلية من خلال معيار المقارنة (MSE) للأنموذج ،بينما تحقق طريقة (Liu-LTS) عندما يكون حجم العينة (n=50) و قيمة معامل الارتباط ($p = 0.95$) اقل قيمة ل (MSE) ، ولكن عند زيادة معامل الارتباط الى($p = 0.99$) تحقق طريقة (Liu-LTS) اقل قيمة ل (MSE) عند حجوم العينات (n=20 , n=50) وتحقق طريقة (Ridge-LTS) اقل قيمة عندما يكون حجم العينة (n=100)

3- وفي حالة وجود التلوث (τ =5%) وتوزيع الاخطاء بالتوزيع الطبيعي بمتوسط (0) وانحراف معياري ($\sigma = 1.5$) نلاحظ من خلال مقياس المقارنة (MSE) للجدول (2) ان طريقة (Ridge-LTS) افضل طريقة عند احجام العينات (n=20, n=50, n=100) عندما تكون قيمة معامل الارتباط ($p = 0.95, 0.90$) وذلك لا نها حققت اقل قيمة ل (MSE) للأنموذج ، ايضاً حققت طريقة (Ridge-LTS) الأفضلية لا قل قيمة لمعيار المقارنة(MSE) عند احجام العينات (n=20 , n=100) عند قيمة معامل الارتباط ($p = 0.99$) ، بينما حققت طريقة (Liu-LTS) اقل قيمة عند حجم العينة (n=50) كونها حققت اقل قيمة لمتوسط مربعات الخطأ (MSE) .

4- وفي حالة في حالة وجود تلوث (τ =15%) وتوزيع الاخطاء التوزيع الطبيعي بمتوسط (0) وانحراف معياري ($\sigma = 1$) نلاحظ من خلال مقياس المقارنة (MSE) للجدول (3) ان طريقة (Ridge-LTS) افضل طريقة عندما يكون حجم العينة (n=20, , n=100) عندما تكون قيمة معامل الارتباط ($p = 0.90, 0.95, 0.99$) ، بينما تحقق طريقة (Liu-LTS) اقل قيمة ل (MSE) عندما يكون حجم العينة (n=50) .

5-وفي حالة في حالة وجود تلوث (τ =15%) وتوزيع الاخطاء التوزيع الطبيعي بمتوسط (0) وانحراف معياري ($\sigma = 1.5$) نلاحظ من خلال مقياس المقارنة (MSE) للجدول (4) ان طريقة (Ridge-LTS) افضل طريقة عندما يكون حجم العينة (n=20, , n=100) عندما تكون قيمة معامل الارتباط ($p = 0.90, 0.95, 0.99$) ، بينما تحقق طريقة (Liu-LTS) اقل قيمة ل (MSE) عندما يكون حجم العينة (n=50) .

8-الاستنتاجات والتوصيات

8-1- الاستنتاجات :

1- اظهرت النتائج لهذا البحث ان طريقة (Ridge-LTS) هي الافضل في معظم تجارب المحاكاة ، حيث تمتلك اقل قيمة لمتوسط مربعات الخطأ (MSE) مقارنة مع الطريقة الاخرى .
2- ان طريقة (Liu-LTS) تحقق اقل قيمة ل (MSE) عندما تكون نسبة التلوث قليلة .
3- اثبتت طريقة (Ridge-LTS) انها اكثر كفاءة في تقدير المعلمات عندما تكون حجوم العينات كبيرة ونسبة التلوث عالية .
4- تتناقص قيمة ال (MSE) عند زيادة حجوم العينات ونسب التلوث.

8-2-التوصيات :

1- استخدام طريقة (Ridge-LTS) في تقدير معلمات أنموذج الانحدار الخطي المتعدد (MLR) وباختلاف احجام العينات لما تبديه من كفاءة ومرونة في التطبيق .
2- اعتماد طريقة (Ridge-LTS) في تقدير معلمات أنموذج الانحدار الخطي المتعدد (MLR) في حالة احجام العينات الكبيرة ونسب التلوث العالية .
3- اعتماد طريقة (Liu-LTS) عند حجوم العينات الصغيرة ونسب تلوث معينة .

9- المصادر العربية:-

1. الجشعمي، حسين علي عبد الله 2007 م ، " مقارنة لبعض المقدرات الحصينة لمعالم النماذج اللاخطية " اطروحة دكتوراه في الاحصاء، كلية الادارة والاقتصاد، الجامعة المستنصرية .
2. النداوي ، سرى صباح كتيب 2008م ، " مقارنة بعض المقدرات الحصينة في الدوال التميزية مع تطبيق عملي" رسالة ماجستير في الاحصاء ، كلية الادارة والاقتصاد ، جامعة بغداد .
3. كاظم ، أموري هادي و مسلم ، باسم شليبه 2002م ، " القياس الاقتصادي المتقدم النظرية والتطبيق"، مكتبة دنيا الامل ، بغداد.

10- **المصادر الاجنبية** :-

1. Alguraibawi , M., Midi , H., & Rana , L. S. (2015) “Robust Jackknife Ridge Regression to Combat Multicollinearity and High Leverage Points in Multiple Linear Regressions” Economic Computation and Economic Cybernetics Studies and Research, NO. 4 ,PP 305-322.

2. Alma, Ö. G. (2011) “Comparison of Robust Regression Methods in Linear Regression” Int. J. Contemp. Math. Sciences,NO. 6(9), PP 409-421.

3. El-Dereny , M., and Rashwan , N. I . (2011) “Solving Multicollinearity Problem Using Ridge Regression Models” Int. J. Contemp. Math. Sciences, NO.6(12) , PP 585 – 600.

4. Hekimoglu, S., and Erenoglu,R. Z. (2013) “A new GM-estimate with high breakdown point ” Acta Geod Geophys,NO. 48 ,PP 419-437.

5. Jung, K. M. (1978) “Least Trimmed Squares Estimator in the Errors-in-Variables Model ” The American Statistician, NO. 34(3),PP 331-338.

6. Kan , B., Alpu, Ö., & Yazıcı B. (2013) “Robust Ridge and Robust Liu Estimator for Regression Based on the Lts Estimator” Journal of Applied Statistics, NO. 40 (3),PP 644-655.

7. Kejian, L.(1993) “A new class of blased estimate in linear regression” Statistical Papers, NO . 22(2), PP 393-402.

8. Mohammed, M.A. (2016) Robust Techniques for Linear Regression with Multicollinearity and Outliers. Thesis Submitted to the School of Graduated Studies, in Fulfillment of the Requirements for the Degree of "Doctor of Philosophy of statistics, Universiti Putra Malaysia .

9. Rousseeuw, P. J., and Driessen , K. V. (2006) “Computing Lts Regression for Large Data Sets” Data Mining and Knowledge Discovery, NO. 12, PP 29-45.

Using Some Robust Methods For Handling the Problem of Multicollinearity

Abstract

The multiple linear regression model is an important regression model that has attracted many researchers in different fields including applied mathematics, business, medicine, and social sciences , Linear regression models involving a large number of independent variables are poorly performing due to large variation and lead to inaccurate conclusions , One of the most important problems in the regression analysis is the multicollinearity Problem, which is considered one of the most important problems that has become known to many researchers , As well as their effects on the multiple linear regression model, In addition to multicollinearity, the problem of outliers in data is one of the difficulties in constructing the regression model , Leading to adverse changes when taking linear regression as a basis for hypothesis testing .

In this paper, we present some robust methods for estimating the parameters of the multiple linear regression model, a ridge regression method for based on the LTS estimator and Liu method for based on the LTS estimator, Using the simulation, these two methods were compared according to the mean squares error (MSE) , The comparison showed that the Liu-LTS method is the best in estimating the parameters of the multiple linear regression model.

Keywords : Multiple Linear Regression , Multicollinearity, outliers, ridge regression, LTS-estimator, Liu-estimator.