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

المقارنة بين طريقتي MوS لتقدير انحدار الحرف الحصين في ظل وجود القيم الشاذة باستعمال أسلوب المحاكاة

منتهى داخل هندول(1)   أ. م. د. طارق عزيز صالح(2)

جامعة واسط – كُلّية الإدارة والاقتصاد

https://doi.org/10.29124/kjeas.1549.29

المستخلص

في الغالب تعتمد الطرائق الإحصائية على افتراض أن التوزيع الاحتمالي لبيانات أنموذج الانحدار يكون طبيعياً ولكن في الحالة التطبيقية غالباً ما تكون هذه البيانات تتوزع توزيعات أخرى أو قد تكون ملتوية والسبب يعود إلى وجود بعض القيم المتطرفة .

في هذه البحث استعملت طريقتان لتقدير معالم أنموذج الانحدار الحصين، إذ استعمل انحدار الحرف لمعالجة مشكلة التعدد الخطي باستعمال طرائق حصينة وهي (M , S) لبيان أثر وجود القيم الشاذة في قيم المشاهدات إذ استعمل أسلوب المحاكاة بالاعتماد على أنموذج يتضمن متغيرين مستقلين ، بهدف إجراء المقارنة بين تلك الطرائق وبيان الطريقة الأفضل في التقدير باستعمال مقياس مقارنة معدل متوسط مربعات الخطأ AMSE لقياس كفاءة الأنموذج .

الكلمات المفتاحية : القيم الشاذة ، مشكلة التعدد الخطي ، المقدرات الحصينة ، طريقة -M ، طريقة –S ،انحدار الحرف

ABSTRACT

Mostly, statistical methods are based on the assumption that the probability distribution of regression model data is normal, but in the applied case, these data often have other distributions or may be skewed, due to the presence of some outliers.

In this research, two methods were used to estimate the parameters of the hippocampal regression model. If literal regression is used to address the problem of multicollinearity using powerful methods, namely (M, S) to show the effect of the presence of outliers in observation values, if a simulation method is used based on a model that includes two independent variables. In order to make a comparison between these methods and indicate the best estimation method using AMSE mean square error comparison measure to measure model efficiency.

Keywerds : Outlievs , Multicollinearity problem , Robust estimators , M-methed , S-methed ,Ridge Regression .

1-المقدمة:

1-1-تمهيد :

تعدُ إحدى أكثر موضوعات البحث استخداماً وتطبيقاً خلال السبعينيات والثمانينيات، طريقة مقترحة حديثًا، تسمى انحدار الحرف (Ridge Regression) تستعمل لتقدير معاملات الانحدار الخطي المتعددة. على الرغم من وجود مقدمات سابقة لهذه المنهجية في الأدبيات من قبل (Crone, 1972, 134) و (Hoerl, 1962, 57) و (Draper, 1963, 471) ، إلا أن شعبيتها ارتفعت بشكل كبير مع نشر مقال (Hoerl and Kennard, 1975, 58).

يعود تطور وأهمية استعمال هذا الأسلوب إلى ظهور مشكلة شائعة بشدة عند تقدير نموذج الانحدار الخطي المتعدد، وهي مشكلة التعدد الخطي. إذ يسمح انحدار الحرف من خلال خصائص مقدراته إلى إمكانية التخفيف من هذه المشكلة وتقدير النموذج المستهدف بوجودها مع تقليل أخطاء النموذج إلى أصغر ما يمكن. كما تظهر مع هذه المشكلة وعند إجراء التحليل الوصفي للبيانات وجود القيم المتطرفة، تؤدي هذه المشكلة إلى عدم توزيع البيانات وفق التوزيع الطبيعي ثَّمّ مخالفتها لافتراضات نموذج الانحدار وعدم صلاحيته لتقدير البيانات، لحل هذه المشكلة تم دمج الأساليب الحصينة Robust مع نموذج الحرف للتوصل إلى نموذج انحدار الحرف الحصين. نهدف من خلال هذه الدراسة إلى دراسة نظرية وتوصيف لأهم خصائص نموذج الحرف الحصين ومقدراته وشرح المشاكل التي يمكن حلها، باستعمال نموذج الانحدار الخطي ونموذج انحدار الحرف .

2-1- مشكلة البحث:

تعد مشاكل التعدد الخطي والقيم المتطرفة من أكثر المشاكل التي تعترض الباحثين عند تقدير النماذج الإحصائية والتي تؤدي إلى عدم الوصول إلى نتائج موثوقة، إذ استعمل نموذج انحدار الحرف الحصين، إذ يسمح انحدار الحرف من خلال خصائص مقدراته إلى إمكانية التخفيف من هذه المشكلة وتقدير النموذج المستهدف بوجودها مع تقليل أخطاء النموذج إلى أصغر ما يمكن. كما تظهر مع هذه المشكلة وعند إجراء التحليل الوصفي للبيانات وجود القيم المتطرفة و تؤدي إلى عدم توزيع البيانات وفق التوزيع الطبيعي مما يعني مخالفتها لافتراضات نموذج الانحدار وعدم صلاحيته لتقدير البيانات، لحل هذه المشكلة تم دمج الأساليب الحصينة Robust مع نموذج الحرف للتوصل إلى نموذج انحدار الحرف الحصين.

3-1- هدف البحث :

تقدم المربعات الصغرى العادية (OLS) تقديرات جيدة في الانحدار إذا تم استيفاء الافتراضات جميعها. ومع ذلك، إذا لم يتم الالتزام بالافتراضات بسبب وجود القيم المتطرفة والتعدد الخطي، فقد تتشوه تقديرات المعلمات بشدة. وبذلك يهدف هذا البحث إلى:

1- شرح مفهوم وخصائص انحدار الحرف الحصين.

2- توصيف مشاكل الانحدار التي يمكن حلها من خلال نموذج انحدار الحرف الحصين.

3- بيان مقدرات انحدار الحرف الحصين وطرائقه .

4- مقارنة نتائج تقدير نموذج الانحدار الخطي وانحدار الحرف الحصين في حال وجود مشكلة تعدد خطي وقيم متطرفة لمجموعة من المتغيرات التوضيحية .

2-الجانب النظري:

2- 1-مفاهيم اساس:

2-1-1مشكلة التعدد الخطي: [7]

تحدث العلاقة الخطية المتعددة عندما ترتبط المتغيرات المستقلة في نموذج الانحدار. يعد هذا

الارتباط مشكلة لأن المتغيرات المستقلة يجب أن تكون غير مرتبطة. إذا كانت درجة الارتباط بين المتغيرات عالية بدرجة كافية، فقد تتسبب في حدوث مشكلات عند ملاءمة النموذج وتفسير النتائج. الهدف الرئيس لتحليل الانحدار هو عزل العلاقة بين كل متغير مستقل والمتغير التابع. تفسير معامل الانحدار هو أنه يمثل متوسط التغيير في المتغير التابع لكل وحدة تغيير واحدة فالمتغير مستقل عندما مع ثبات المتغيرات المستقلة الأخرى جميعها. هذا الجزء الأخير مهم بشكل كبير لفهم نشوء مشكلة التعدد الخطي.

الفكرة هي إمكانية تغيير قيمة أحد المتغيرات المستقلة وليس المتغيرات الأخرى. ومع ذلك، عندما ترتبط المتغيرات المستقلة فإنها تشير إلى أن التغييرات في متغير واحد مرتبطة بالتحولات في متغير آخر. كلما كانت العلاقة أقوى، زادت صعوبة تغيير متغير دون تغيير آخر. يصبح من الصعب على النموذج تقدير العلاقة بين كل متغير مستقل والمتغير التابع بشكل مستقل لأن المتغيرات المستقلة تميل إلى التغيير في انسجام وهو توضيح لفكرة التعدد الخطي Multicollinearity.

2-1-2 القيم المتطرفةً:(Outliers Values) [3]

يُمثَّل للقيم المتطرفة بنقاط بيانات بعيدة كل البعد عن نقاط البيانات الأخرى. بعبارة أخرى، هي قيم غير عادية في مجموعة البيانات. تعد القيم المتطرفة مشكلة بالنسبة للعديد من التحليلات الإحصائية لأنها يمكن أن تتسبب في تضليل الاختبارات لنتائج مهمة أو تشويه النتائج الحقيقية. لا توجد قواعد إحصائية ثابتة لتحديد القيم المتطرفة بشكل نهائي. يعتمد العثور على القيم المتطرفة على معرفة مجال الموضوع وفهم عملية جمع البيانات. على الرغم من عدم وجود تعريف رياضي قوي، لكن هناك إرشادات واختبارات إحصائية يمكن استعمالها للعثور على القيم المتطرفة.

2-1-3-مفهوم انحدار الحرف (Ridge Regression**):** [6]

انحدار الحرف هو طريقة شائعة لتقدير المعلمات تستعمل لمعالجة مشكلة تعدد العلاقة الخطية التي تنشأ بشكل متكرر في الانحدار الخطي المتعدد. انتشر استعمال هذه النماذج بصورة واسعة وكانت الدولة الأكثر استعمالاً لهذه الطريقة هي الصين.

تُعزى شعبية هذا الموضوع إلى أهمية المشكلة التي يعالجها – مشكلة التعدد الخطي في سياق الانحدار الخطي المتعدد - ومدى ملاءمة المنهجية ليتم تنفيذها بسهولة في الممارسة بناءً على فحص ما يسمى بتتبع الانحراف، وهو جزء من معاملات الانحدار المقدرة كدالة لمعلمة الحرف. ينتج عن منهجية انحدار الحرف فئة من المقدرات المتحيزة مفهرسة بمعامل قياسي غير سالب. يتمثل التحدي في تحديد المقدر ضمن هذه الفئة لاستعماله في سياق مشكلة معينة، أي تحديد

أفضل خيار لمعامل الحرف. هذا المسعى لبناء أفضل خيار هو الدافع الرئيس للعدد الكبير من المنشورات البحثية حول هذا الموضوع خلال العقود القليلة التي أعقبت المنشورات الأساس. تضمنت العديد من هذه المهام دراسات محاكاة واسعة النطاق للمقدرات العشوائية لمعامل الانحراف (أي المقدرات التي هي دوال للمتغير التابع).

إن التأثير السلبي لتعدد العلاقة الخطية على مقدر المربعات الصغرى (LS) في سياق الانحدار معروف جيدًا. تم تطوير مناهج للتخفيف من هذا التأثير وتمحور الكثير حول الحذف المتغير، أي إزالة واحد أو أكثر من المتغيرات المستقلة لتحسين تكييف مصفوفة الارتباط الناتجة عن المتغيرات المستقلة المتبقية. من ناحية أخرى، يوفر انحدار الحرف وسيلة لمعالجة مشكلة تعدد العلاقة الخطية دون إزالة المتغيرات من المجموعة الأصلية للمتغيرات المستقلة. أثبت هذا أنه ميزة جذابة للغاية في بعض التطبيقات كما هو موضح بواسطة، ومع ذلك، يمكن استعمال تتبع الحرف في اختيار متغير إذا رغب الباحث في ذلك. تم استعمال هذا النهج المخصص في دراسة (McDonald and Schwing471,1973,) المتعلقة بمؤشرات تلوث الهواء بمعدلات الوفيات والنتائج الخاصة بالمشكلة اللاحقة مقارنة بشكل إيجابي مع تلك التي تم الحصول عليها بواسطة الاختيار المتغير باستعمال إحصاء Mallows's Cp مع الانحدارات الممكنة جميعها.

لصياغة انحدار الحرف، تمّ استعمال الوصف الذي قدمه (McDonald and Schwing،1973)، إذ لعبت تقنيات الانحدار الخطي المتعددة دورًا بارزًا في دراسات الارتباط بين تلوث الهواء ومعدلات الوفيات، قد يوفر نهج LS أساسًا مناسبًا للتنبؤ الكلي، ولكن عندما تكون المتغيرات التوضيحية غير متعامدة non-orthogonal، فإنها تفشل في كثير من الأحيان في إعطاء الوزن المناسب للمتغيرات التفسيرية الفردية المستعملة كمتنبئات و في العديد من المشكلات التي لا يتم فيها الحصول على البيانات من تجربة مصممة جيدًا أو مضبوطة، كما هو الحال في دراسات تلوث الهواء التي تنطوي على متغيرات اجتماعية واقتصادية وطقس ومتغيرات أخرى غير خاضعة للرقابة، يتطلب عدم التعامد هذا بأن يتم التعامل مع التأثيرات بتقنيات أخرى غير حلول LS العادية. تقدم الأوراق الأساس، وصفًا ممتازًا للنظرية والتطبيقات لما يُطلق عليه الآن "انحدار الحرف".

تميل تقديرات معاملات الانحدار إلى أن تصبح كبيرة جدًا في القيم المطلقة، ومن الممكن أن يكون لدى البعض إشارة خاطئة. تزداد فرص مواجهة مثل هذه الصعوبات كلما انحرفت اشعة التنبؤ عن التعامد. وانطلاقا من النموذج القياسي للانحدار الخطي المتعدد:

$$y = \beta x + \varepsilon$$

إذ $E(\epsilon) = 0$ و $E(\varepsilon\acute{\varepsilon}) = \sigma^2 I_n$ و x هي $n \times p$ من رتبة المتغيرات الكاملة. المصفوفة I_n هي مصفوفة الوحدة في البعد n. يُفترض أن تكون المتغيرات معيارية بحيث تكون $x\acute{x}$ بشكل مصفوفة ارتباط. وأنّ $\gamma = \acute{x}y$ هو متجه معاملات الارتباط لمتغيرالاستجابة مع المتغيرات المستقلة.

يتم التوحيد القياسي بطرح متوسط المتغير ثم من خلال القسمة على الانحراف المعياري للمتغير $(n-1)^{1/2}$. لتكن:

$$\hat{\beta} = (\acute{x}x)^{-1}\acute{x}y$$

تقدر β باستعمال المربعات الصغرى. الصعوبات في هذا التقدير القياسي هي نتيجة مباشرة لمتوسط المسافة بين $\hat{\beta}$ و β. ولاسيما إذا كانت L^2 المسافة المربعة بينهما عندئذ:

$$L^2 = (\hat{\beta} - \beta)(\beta - \hat{\beta})$$

$$E(L^2) = \sigma^2 track(\acute{x}x)^{-1}$$

$$E(\hat{\beta}\beta) = \beta\beta + \sigma^2 track(\acute{x}x)^{-1}$$

من حيث القيم الكامنة $(\lambda_1 \geq \lambda_2 \ldots . \geq \lambda_p > 0$ eigenvalues) ل $\acute{x}x$ ومن ثَمَّ يمكننا:

$$E(L^2) = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} > \sigma^2 \lambda_p^{-1}$$

$$E(\hat{\beta}\beta) = \beta\beta + \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} > \beta\beta + \sigma^2 \lambda_p^{-1}$$

وعندما يوَزّعُ الخطأ بشكل طبيعي:

$$Var(L^2) = 2\sigma^4 \sum_{i=1}^{p} \lambda_i^{-2} > 2\sigma^4 \lambda_p^{-2}$$

ولأن متجهات x تنحرف أكثر عن التعامد، تصبح λ_p أصغر، ويمكن توقع ان تكون $\hat{\beta}$ ابعد من متجه المعلمة الحقيقي β. ومن ثَمَّ يعتمد تقدير انحدار الحرف على:

$$\hat{\beta}(k) = (\acute{x}x + kI)^{-1}\acute{x}y, k \geq 0$$

للمعادلة جانبان، الأول هو تتبع الحرف وهو مخطط ثنائي الابعاد ل $\hat{\beta}(k)$ ومجموع مربعات البواقي (RSS)، ومن اجل قيم k في المجال [1،0]، من الممكن ان تتجاوز k القيمة 1، عادة ما يُكتَفى برسم المعاملات ل $0 \leq k \leq 1$ لتحديد قيمة k التي يكون بعدها أثر الحرف مستقر تماماً. يعمل التتبع على تصوير العلاقات المتبادلة المعقدة الموجودة بين متجهات التنبؤ غير المتعامدة وتأثير هذه العلاقات المتبادلة على تقدير β. الجانب الثاني هو تحديد قيمةk التي تعطي تقديرًا أفضل لـ β عن طريق تقليل تأثير L^2. وتجدر الإشارة إلى أنّ المقدر $\hat{\beta}(k)$ يصبح متحيزاً عندما تكون $k > 0$، وعندما تكون $k = 0$ تقل هذه المقدرات إلى المربعات الصغرى العادية غير المتحيزة.

4-1-2- انحدار الحرف الحصين:

هو مزيج من الحرف والانحدار الحصين للتعامل مع مشكلة الخطية المتعددة والقيم المتطرفة في وقت واحد. سيؤدي ذلك إلى تخفيف آثار كلتا المشكلتين في نموذج الانحدار الخطي الكلاسيكي. لحساب RobustRidgeEstimator ، فإن الصيغة المستعملة هي:

$$\widehat{\beta_{RR}} = (XX + K_R I)^{-1} XY$$

حيث K_R تسمى معامل الحرف الحصين، ويتم الحصول عليها باستعمال انحدار الحرف الحصين بدلاً من استعمال تقدير OLS. ومن هنا وفقاً لهذه الطريقة يتم استبدال المقدرات العادية:

$$mean, \sigma^2, \alpha^2$$

بالمقدرات الحصينة:

$$median, \sigma^2_{Robust}, \alpha^2_{Robust}$$

2-1-5- طرائق التقدير الحصين:

هناك مجموعة من الطرائق التي يبنى عليها تقدير النماذج الحصينة وتتمثل وفق الآتي:

1-تقديرات M [1, 5]

ناقش الباحث Fax عام (2002) الطريقة العامة الأكثر شيوعا للانحدار الحصين وهي طريقة مقدرات M– والتي قدمها الباحث Huber عام (1964) وهي تعمم تقدير احتمالي في سياق نماذج الانحدار وهي تكاد أن تكون كفاءة طريقة المربعات الصغرى الاعتيادية نفسها (OLS) فأن مبدأ طريقة تقدير – M هو تقليل دالة الأخطاء وهي طريقة لحل المعادلات غير الخطية لان الأوزان تعتمد على المعامل المجهول M وβ فلا يمكننا حساب المتوسط المرجح بشكل واضح لكن تمثيل المعادلة الموزونة لمقدرات – M يؤدي إلى خوارزمية تكرارية بسيطة لحساب مقدر – M وسوف يتم توضح الخوارزمية لاحقا ويعتمد اختيار الدالة (ψ) على تفضيل مقدر الوزن (W)للتعيين القيم المتطرفة . لذا يكون مناسبا استعمال طريقة المربعات الصغرى الاعتيادية لحل المشكلات غير الخطية كما إن افتراض الحالة الطبيعية لا يمكن عقده في طريقة المربعات الصغرى الاعتيادية لان قيمة الخطأ ستزيد من تكبير المجاميع الأربع .

يعد تقدير – M واحدة من طرائق تقدير الانحدار الحصين حيث يشير الحرف (M) الى ان تقدير (M) هو تقدير اقصى احتمال إذا كان المقدر

$$\hat{\beta} = \beta_n(X_1, X_2, \cdots, X_n)$$

$$\beta = E[\beta_n(X_1, X_2, \cdots, X_n)] \quad \cdots (1-1)$$

توضح المعادلة (1-1) إن مقدر $\hat{\beta} = \beta_n(X_1, X_2, \cdots, X_n)$ هو تقدير غير متحيز وله ادنى تباين لذلك فمقدر – M لديه اصغر مقدر للتباين مقارنة بتقديرات التباين الأخرى إذ إن :

$$\mathrm{Var}\,(\hat{\beta}) \geq \frac{[\hat{\beta}]^2}{nE\left[\frac{d}{d\beta} \mathrm{Ln}\, f(x_i;\beta)\right]^2}$$

إذ إن $\hat{\beta}$هومقدر خطي وغير متحيز لـ β

إن تقدير- M هو امتداد لأسلوب الكثافة الاحتمالية القصوى وهي من التقديرات الحصينة وفي هذه الطريقة يمكن استبعاد بعض البيانات الشاذة والتي في بعض الأحيان ليس من المناسب دائماً القيام بها لاسيما إذا حذفت من البيانات المهمة

والصيغة التقديرية لـ $\hat{\beta}_M$ تحسب وفق الصيغة الآتية :

$$\hat{\beta}_M = \min \sum_{i=1}^{n} \rho(y_i - \sum_{j=0}^{k} X_{ij}\beta_j) \quad \dots (3)$$

$$\min \sum_{i=1}^{n} \rho(u_i) = \min \sum_{i=1}^{n} \rho\left(\frac{e_i}{\sigma}\right) = \min \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{\sigma}\right) \quad \cdots (4)$$

وللحصول على مقدر σوفق المعادلة الآتية :

$$\hat{\sigma} = \frac{MAD}{0.6745} = \frac{\text{median}\,|e_i - \text{median}(e_i)|}{0.6745}$$

تستعمل دالة الهدف (Tukey)ثنائية المربع لـ ρ وفق المعادلة الآتية

$$\rho(u_i) = \begin{cases} \frac{{u_i}^2}{2} - \frac{{u_i}^4}{2\,C^2} + \frac{{u_i}^6}{2C^4} & |u_i| \le C \\ \frac{C^2}{6} & |u_i| > C \end{cases}$$

والصيغة التقديرية لـ $\hat{\beta}_M$ تكون وفق الصيغة الآتية

$$\hat{\beta}_M = \sum_{i=1}^{n} \rho'\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{\sigma}\right) = 0 \qquad j = 0,1,\dots k$$

والاشتقاق الجزئي الأول من $\hat{\beta}_M$إلى β يكون بالشكل الآتية :

$$\sum_{i=1}^{n} X_{ij}\,\psi\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}}\right) = 0 \qquad j = 0,1,\dots k \qquad \cdots (5)$$

إذ إن

$\psi = \acute{\rho}$ وان x_{ij} هي المشاهدات i-th للمتغير $x_{i0} = 1$

وتعطي المعادلة (4) حل للمعادلة (3) من خلال تحديد الدالة الآتية :

$$W(e_i)=\frac{\psi\left(\frac{y_i-\sum_{j=0}^{k}x_{ij}\beta}{\hat{\sigma}}\right)}{\left(\frac{y_i-\sum_{j=0}^{k}x_{ij}\beta}{\hat{\sigma}}\right)} \qquad \cdots (6)$$

إذ إن

$u_i=\frac{e_i}{\hat{\sigma}}$ ويمكننا اعادة كتابة المعادلة (6) وفق الصيغة الآتية :

$$w_i=(w(u_i)=\frac{\psi(u_i)}{u_i}=\begin{cases}\dfrac{u_i\left(1-\left(\frac{u_i}{c}\right)^2\right)^2}{u_i} & , \quad |u_i|\le C\\ 0 & , \quad |u_i|>C\end{cases}$$

إذ يمكن كتابة المعادلة بالشكل الآتي

$$w_i=\begin{cases}\left[1-\left(\frac{u_i}{c}\right)^2\right]^2 & , \quad |u_i|\le C\\ 0 & , \quad |u_i|>C\end{cases}$$

إذ إن قيمة (c=4.685) لدالة (Tukey) الموزونة ثنائية التربيع وتصبح المعادلة رقم (5) بالشكل الآتي

$$\sum_{i=1}^{n}x_{ij}w_i(y_i-\sum_{j=0}^{k}X_{ij}\beta_j) \quad j=0,1,\dots k \ \dots(7)$$

ويمكن حل المعادلة 7بطريقة المربعات الصغرى الموزونة بشكل متكرر (IRLS) وفي هذه الطريقة سوف نفترض أن هنالك تقديراً أوليا لـ $\hat{\beta}^{\circ}$ وهو تقدير مقاس بـ $\hat{\sigma}_i$و لـ j من المعلمات فأن

$$\sum_{i=1}^{n}x_{ij}w_i^{\circ}\left(y_i-\sum_{j=0}^{k}x_{ij}\beta^{\circ}\right)=0 \qquad ,j=0,1,\dots k \ \dots(8)$$

ويمكن إعادة كتابة المعادلة (8) بالشكل الآتي

$$\acute{x}w_i x\beta=\acute{x}w_i y \qquad \cdots(9)$$

إذ إن w_i مصفوفة قطرية من الدرجة $n\times n$

و تعرف المعادلة (9) بمعادلة المربعات الصغرى الموزونة (WLS)ويعطى حل هذه المعادلة مقدراً لـ β وبالشكل الآتي :

$$\hat{\beta} = (\acute{x} w_i x)^{-1} \acute{x} w_i y$$

حيث يتم تقديم وصف تفصيلي لتقدير- M باستعمال دالة (Tukey) وفق الخوارزمية الآتية :

Step(1) : تقدير معاملات الانحدار على البيانات باستعمال طريقة المربعات الصغرى الاعتيادية (OLS)

Step(2) : اختيار افتراضات نموج الانحدار الخطي العام .

Step(3) : الكشف عن وجود قيم متطرفة في البيانات باستعمال طريقة

Step(4) :نحسب تقديرات $\hat{\beta}^{\circ}$بطريقة المربعات الصغرى الاعتيادية (OLS)

Step(5) :نحسب القيمة المتبقية (الأخطاء) وفق$e_i = y_i - \hat{y}_i$

Step(6) : نحسب قيمة الانحراف المعياري $\hat{\sigma}_i$ وفق الصيغة الآتية

$$\hat{\sigma}_i = \frac{\mathrm{MAD}}{0.6745} = \frac{\mathrm{median}|e_i| - \mathrm{median}(e_i)}{0.6745}$$

Step(7) : نحسب قيمة الخطأ العشوائي بالشكل الآتي

$$u_i = \frac{e_i}{\hat{\sigma}_i}$$

Step(8) : نحسب القيمة المرجحة (w_i)وبالشكل الاتي

$$w_i = \begin{cases} \left[1 - \left(\frac{u_i}{4.685}\right)^2\right]^2 & , \quad |u_i| \le 4.685 \\ 0 & , \quad |u_i| > 4.685 \end{cases}$$

Step(9) : نحسب $\hat{\beta}_M$ باستعمال طريقة (WLS).

Step(10) : نكرر الخطوات من (5) الى (9) للحصول على قيمة مقارنة لـ $\hat{\beta}_M$.

Step(11) :الاختيار لتحديد ما اذا كانت المتغيرات المستقلة لها تأثير معنوي عالٍ على المتغير التابع (المعتمد) .

2- تقديرات–S [1,5,8,9] S-Estimates

اقترح الباحثان Rovsseeuw and yohai في عام 1984تقدير-S الذي يعتمد على المقياس المتبقي لتقدير- M . وهو من التقديرات الضعيفة لعدم مراعاة توزيعات البيانات بسبب استعمال الوسيط فقط كقيمة مرجحة .وتستعمل هذه الطريقة للانحراف المعياري للأخطاء المتغلب على نقاط ضعف الوسيط ويمكن تحديد مقدر – S من خلال المعادلة الإاتية :

$$\hat{\beta}_s = \min_{\beta} \hat{\sigma}_s(e_1, e_2, \cdots, e_n)$$

وتحديد الحد الادنى لمقدر المقياس الحصين $\hat{\sigma}_s$ من خلال الأتي

$$\mathrm{Min} \sum_{i=1}^{n} \rho \left(\frac{y_i - \sum_{j=0}^{k} x_{ij} \beta}{\hat{\sigma}_s} \right)$$

$$\text{And } \hat{\sigma}_s \sqrt{\frac{1}{nk} \sum_{i=1}^{n} w_i \, e_i^2} \qquad ; \quad \text{iteration } >1$$

$$K=0.199 \quad , \quad w_i = w\,\sigma(U_i) = \frac{\rho(U_i)}{U_i}$$

والتقدير الأولي يكون بالشكل الآتي

$$\hat{\sigma}_s \quad \frac{\mathrm{median}|e_i - \mathrm{median}(e_i)|}{0.6745} \quad , \text{iteration } = 1$$

إذ إن

ويمكن الحصول على الحل عن طريق التفاضل لـ β

$$\sum_{i=1}^{n} \acute{\rho} \left(\frac{y_i - \sum_{j=0}^{k} x_{ij} \beta}{\hat{\sigma}_s} \right) = 0 \qquad , \quad j = 0,1, \ldots k$$

$$\sum_{i=1}^{n} x_{ij} \psi \left(\frac{y_i - \sum_{j=0}^{k} x_{ij} \beta}{\hat{\sigma}_s} \right) = 0 \qquad , \quad j = 0,1, \ldots k \qquad (1)$$

إذ إن

Ψ :تمثل دالة مشتقة من ρ $(\acute{\rho} = \psi)$ وهي دالة موزونة IRLS(IterativelReweighted Least Squares)

$$\psi(u_i) = \rho(u_i) = \begin{cases} u_i\left[1-\left(\frac{u_i}{c}\right)^2\right]^2 & , \quad |u_i| \le C \\ 0 & , \quad |u_i| > C \end{cases}$$

$$W_i(u_i) = \frac{\psi(u_i)}{u_i} = \begin{cases} \dfrac{u_i\left[1-\left(\frac{u_i}{c}\right)^2\right]^2}{u_i} & , \quad |u_i| \le C \\ 0 & , \quad |u_i| > C \end{cases}$$

$$W_i(u_i) = \frac{\psi(u_i)}{u_i} = \begin{cases} u_i\left[1-\left(\frac{u_i}{c}\right)^2\right]^2 & , \quad |u_i| \le C \\ 0 & , \quad |u_i| > C \end{cases}$$

$u_i = \frac{e_i}{\hat{\sigma}_s}$ and c=1.547 (Constant)

ويمكن حل المعادلة (1) باستعمال طريقة IRLS(المربعات الصغرى الموزونة التكرارية) إذ إن التقديرات –S تمثل الحل الذي يجد اصغر تشتتاً بدلاً من تقليل التباين قيم الأخطاء ($\min\hat{\sigma}(e_1(\hat{\beta}), e_2(\hat{\beta}), \cdots, e_n(\hat{\beta}))$

ويقلل تقدير- Sالحصين وفق المعادلة الآتية :

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}_s}\right) b$$

b: يمثل ثابت معرف على انه $b = E\phi[\rho(e_i)]$

ϕ : يمثل التوزيع الطبيعي القياس

وحل النتائج يكون وفق الصيغة الآتية

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{e_i}{\hat{\sigma}_s}\right)$$

يمكن استبدالها بدالة وزن مناسبة كما هو الحال مع معظم إجراءات تقدير – Mوعادة ما تُستعمل إحدى دوال الوزن كدالة وزن Huber .

وتُعد مقدرات –Sأكثر قوة من مقدرات – M لان مقدرات – Sلها تحيز مقارب أصغر وتباين مقارب أصغر في البيانات الملوثة .ويمكن أن تصل مقدرات – Sإلى نقطة انهيار عالية $\beta_p = 0.5$ ولها كفاءة مقاربة تبلغ 0.29في ظل افتراض الأخطاء الموزعة بشكل طبيعي .للباحث Maranna و أخرين في عام (2006)

ويمكن توضيح خوارزمية تقديرات – S بالخطوات الآتية

Step(1) : تقدير معاملات الانحدار على البيانات باستعمال طريقة المربعات الصغرى الاعتيادية (OLS)

Step(2) :اختيار افتراضات نموذج الانحدار الخطي العام

Step(3) : الكشف عن وجود قيم متطرفة في البيانات باستعمال طريقة

Step(4) : نحسب تقديرات $\hat{\beta}^{\circ}$بطريقة المربعات الصغرى الاعتيادية (OLS)

Step(5) :نحسب القيمة المتبقية (الأخطاء) وفق $e_i = y_i - \hat{y}_i$

Step(6) : نحسب قيمة الانحراف المعياري وفق الاتي

$$\hat{\sigma}_s \begin{cases} \dfrac{\text{median}|e_i - \text{median}(e_i)|}{0.6745} & , \quad \text{iteration } = 1 \\ \sqrt{\dfrac{1}{nk}\sum_{i=1}^{n} w_i\, e_i^2} & , \quad \text{iteration } > 1 \end{cases}$$

Step(7) : نحسب قيمة u_i بالشكل الاتي

$$(u_i) = \begin{cases} \begin{cases} \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2 & , \quad |u_i| \leq 1.547 \quad , \text{iteration} = 1 \\ 0 & , \quad |u_i| > 1.547 \end{cases} \\ \frac{\rho(u)}{u^2} & \quad \text{iteration} > 1.547 \end{cases}$$

$$u_i = \frac{e_i}{\hat{\sigma}_s}$$

Step(8):نحسب القيمة المرجحة (w_i) بالشكل الاتي :

$$W_i \begin{cases} \begin{cases} \left[1-\left(\frac{u_i}{c}\right)^2\right]^2 & , \quad |u_i| \le 1.547 \quad , \text{iteration} = 1 \\ 0 & , \quad |u_i| > 1.547 \end{cases} \\ \frac{\rho(u)}{u^2} & \quad \text{iteration} > 1.547 \end{cases}$$

Step(9) : نحسب $\hat{\beta}_s$ باستعمال طريقة WLS مع وزن w_i

$$\hat{\beta} = (\acute{x} w_i x)^{-1} \acute{x} w_i y$$

Step(10) : نكرر الخطوات من (5) إلى (9) للحصول على قيمة مقاربة لـ $\hat{\beta}_s$.

Step(11) :الاختيار لتحديد ما إذا كانت المتغيرات المستقلة لها تأثير معنوي عالٍ على المتغير التابع (المعتمد) .

6-1-2-معيار متوسط مربعات الخطأ:

يقيس متوسط مربعات الخطأ (MSE) مقدار الخطأ في النموذج الإحصائي إذّ يعمل على تقييم متوسط الفرق التربيعي بين القيم الفعلية والمتوقعة. عندما لا يوجد خطأ في النموذج، فإن MSE يساوي صفرًا. مع زيادة خطأ النموذج، تزداد قيمته. يُعرف متوسط مربع الخطأ أيضًا باسم متوسط مربع الانحراف (MSD). في الانحدار، يمثل متوسط مربع الخطأ متوسط مربعات البواقي .يمكن حساب متوسط مربع الخطأ من خلال المعادلة الآتية **(Willmott, and Matsuura, 2005, 81)**:

$$\text{AMSE} = \frac{\sum (yi - \hat{y}i)}{n}$$

حيث y هي البيانات الفعلية، $\hat{y}$ البيانات المقدرة باستعمال النموذج، n: حجم العينة أو عدد المشاهدات. تشبه حسابات متوسط مربع الخطأ التباين. خطوات تحديد MSE ، تتم من خلال أخذ القيمة الفعلية، وطرح القيمة المتوقعة ثم بتربيع هذا الفرق ويتم تكرار ذلك المشاهدات جميعها. بعد ذلك، يتم جمع هذه القيم التربيعية كلها وقسمتها على عدد المشاهدات. لاحظ أن البسط هو مجموع الأخطاء التربيعية (*SSE*) ، إذ يقلل الانحدار الخطي منها. يقسم MSE ببساطة SSE على حجم العينة.

3-**الجانب التطبيقي**

1-3المحاكاة :

1-1-3 خطة المحاكاة :

يمكن تنفيذ برنامج المحاكاة باستعمال لغة R .إذ تضمن البرنامج أربع مراحل لتقدير النموذج الانحدار .

أولا: تحديد القيم الافتراضية

1- توليد البيانات بالاعتماد على نموذج الانحدار الخطي الآتي :

$$Y_i=\beta_\circ+\beta_1 x_{1i}+\beta_2 x_{2i}+\sigma\varepsilon_i$$

i=1,2,3,…,n

بالافتراض قيم المعلمات التالية

$\beta_\circ = 0.5$, $\beta_1=1$, $\beta_2=1.5$

2-على حد الخطأ العشوائي باستعمال التوزيع الطبيعي القياسي نحصل على

N=(0,1)

3- استعملت مجموع عينات مختلفة وهي

N= 25 , 50 , 100

والمتغيرات التوضيحية هي

K = 2

4-تم افتراض العينات تحتوي على قيم شاذة وهي

$$\varepsilon = 0.0\ , 0.10\ , 0.30$$

وكل تجربة بحثت عند مستويين الانحراف المعياري وهما

$$\sigma = 1.5$$

وقيم معامل الارتباط بأخذ القيم الآتية

$$\rho = 0.0\ , 0.90$$

ثانياً: مرحلة توليد البيانات

تم توليد المتغيرات التوضيحية وفقاً للصيغة الآتية

$$X_{ij} = (1 - p^2)Z_{ij} + z\rho$$

i=1,2,…,n , j=1,2,…,p

إذ أنَّZijهي توليد من التوزيع الطبيعي القياسي وثمَّ توليد الخطأ العشوائي من التوزيع الطبيعي بمتوسط صفري وتباين بالنسبة للأخطاء التي لا تعاني من وجود القيم الشاذة وبمتوسط 1 وتباين .... بالنسبة للخطأ التي تعاني من القيم الشاذة وكما يأتي :

$$\varepsilon_1 \sim N(\,0\,,\sigma^2)$$

$\varepsilon_2 \sim \mathrm{N}(\,1\,,\sigma^2)$

$\varepsilon_{\mathrm{N}} = (\varepsilon_1\,,\varepsilon_2)$

ويتم توليد المتغير المعتمد بالاعتماد على معادلة الانحدار التالية:

$Y_i = X^T\beta ++\varepsilon_i$

$$i = 1,2,\dots,n$$

ثالثاً: (مرحلة التقدير):

في هذه المرحلة تُجرى عملية التقدير لمعلمات نموذج الانحدار باستعمال طرائق التقدير الآتية:

1- طريقة انحدار الحرف مع مقدر (RR-M) M
2- طريقة انحدار الحرف مع مقدر(RR-S) S

رابعاً: مرحلة المقارنة بين الطرائق

لغرض إجراء المقارنة بين طرائق التقدير بالاعتماد على المعيار معدل متوسط مربعات الخطأ (AMSE)

u) مقارنة بالطرائق الأخرى بالاعتماد على معدل متوسط مربعات الخطأ (AMSE).

وكما مبين في الصيغة الآتية:

$MSE(\boldsymbol{\hat{\beta}}) = \frac{1}{R}\sum_{\boldsymbol{I=1}}^{\boldsymbol{R}}(\boldsymbol{\hat{\beta}_I} - \boldsymbol{\beta})'(\boldsymbol{\hat{\beta}_I} - \boldsymbol{\beta})$

إذ إن R تمثل عدد مرات تكرار التجربة وهي تساوي (R=1000)

3-1-2-نتائج المحاكاة :

تم إجراء نتائج المحاكاة وكانت النتائجكما يأتي :

الجدول رقم (1) يمثل القيم التقديرية لمتجه المعلمات وقيم معدل متوسط مربعات الخطأ عندما $\boldsymbol{\rho = 0.0}$, $\boldsymbol{\sigma = 1.5}$, $\boldsymbol{p = 2 \varepsilon = 0.0\%}$

Sample Size	Values	Ridge Regression (RR)	
		M	S
25	β_0	0.470023	0.461518
	β_1	0.925338	0.909566
	β_2	1.397855	1.375411
	AMSE	0.318589	0.34627
50	β_0	0.488136	0.485562
	β_1	0.977596	0.971937
	β_2	1.459172	1.450682
	AMSE	0.148811	0.152399
100	β_0	0.497752	0.497065
	β_1	0.987914	0.986435
	β_2	1.478493	1.476304
	AMSE	0.071328	0.071513

نلاحظ من خلال الجدول رقم (1) أن طريقة M هي أفضل طريقة تقدير مقارنة بالطرائق الأخرى لمختلف حجوم العينات لانحدار الحرف بالاعتماد على معدل متوسط مربعات الخطأ AMSE.

الجدول (2): يمثل القيم التقديرية لمتجه المعلمات وقيم معدل متوسط مربعات الخطأ AMSE للنموذج عندما P=2 و

σ =1.5 و ρ=0.0 و ϵ=10%

Sample Size	Values	Ridge Regression (RR)	
		M	S
25	β_0	0.48346	0.480203
	β_1	0.967309	0.961743
	β_2	1.434613	1.424847
	MSE	0.350821	0.354823
50	β_0	0.489694	0.488458
	β_1	0.982342	0.980802
	β_2	1.460358	1.457278
	MSE	0.163784	0.164264
100	β_0	0.488502	0.487931
	β_1	0.998869	0.998531
	β_2	1.481741	1.480473
	MSE	0.075014	0.074918

نلاحظ من خلال الجدول رقم (2) أن طريقة M هي أفضل طريقة تقدير مقارنة بالطرائق الأخرى لمختلف حجوم العينات باستثناء حجم العينة (n=100) فكانت الأفضلية لطريقة S لانحدار الحرف بالاعتماد على معدل متوسط مربعات الخطأ (AMSE).

الجدول (3): يمثل القيم التقديرية لمتجه المعلمات وقيم معدل متوسط مربعات الخطأ AMSE للنموذج عندما P=2 و

σ =1.5 و ρ=0.0 و ϵ=30%

Sample Size	Values	Ridge Regression (RR)	
		M	S
25	β_0	1.009454	0.993114
	β_1	0.90277	0.881815
	β_2	1.221094	1.190963
	MSE	1.1379	1.106277
50	β_0	1.296695	1.284942
	β_1	0.92938	0.9225
	β_2	1.345379	1.328828
	MSE	1.100241	1.075025
100	β_0	1.391751	1.388337
	β_1	0.956038	0.954858
	β_2	1.417355	1.412397
	MSE	1.088498	1.069347

نلاحظ من خلال الجدول رقم (3) أن طريقة S هي أفضل طريقة تقدير مقارنة بالطرائق الأخرى لمختلف حجوم العينات مقارنة بالطرائق الأخرى بالاعتماد على معدل متوسط مربعات الخطأ (AMSE).

الجدول (4): يمثل القيم التقديرية لمتجه المعلمات وقيم معدل متوسط مربعات الخطأ AMSE للنموذج عندما

P=2 و σ =1.5 و ρ=0.9 و ϵ=0.0%

Sample Size	Values	Ridge Regression (RR)	
		M	S

25	β_0	2.830303	2.144956
	β_1	0.944118	0.839314
	β_2	1.199012	0.914169
	MSE	10.18526	6.348629
50	β_0	2.963445	2.534999
	β_1	0.951212	0.956248
	β_2	1.327561	1.105691
	MSE	8.842751	4.866428
100	β_0	3.149595	2.917858
	β_1	1.057123	1.059782
	β_2	1.319289	1.20935
	MSE	8.304954	4.129971

نلاحظ من خلال الجدول رقم (4) أن طريقةSهي أفضل طريقة تقدير مقارنة بالطرائق الأخرى لمختلف حجوم العينات بالاعتماد على معدل متوسط مربعات الخطأ (AMSE).

الجدول (5): يمثل القيم التقديرية لمتجه المعلمات وقيم معدل متوسط مربعات الخطأ AMSE للنموذج عندما

P=2 و σ =1.5 و ρ=0.9 و ϵ=10%

Sample Size	Values	Ridge Regression (RR)	
		M	S
25	β_0	0.473592	0.47554
	β_1	1.099437	1.096333
	β_2	1.356712	1.363292

	MSE	3.362552	4.630114
50	**β_0**	0.487751	0.489161
	β_1	1.024383	1.05001
	β_2	1.452636	1.430447
	MSE	1.908511	2.267596
100	**β_0**	0.492308	0.493444
	β_1	1.05393	1.049565
	β_2	1.424233	1.43171
	MSE	1.045161	1.185067

نلاحظ من خلال الجدول رقم (5) أن طريقةM هي أفضل طريقة تقدير مقارنة بالطرائق الأخرى لمختلف حجوم العينات لانحدار الحرف بالاعتماد على معدل متوسط مربعات الخطأ (AMSE).

الجدول (6): يمثل القيم التقديرية لمتجه المعلمات وقيم معدل متوسط مربعات الخطأ AMSE للنموذج عندما

P=2 و σ =1.5 و ρ=0.09 و ϵ=30%

Sample Size	**Values**	**Ridge Regression (RR)**	
		M	**S**
25	**β_0**	1.36334	1.381035
	β_1	1.282709	1.330651
	β_2	1.479775	1.447286
	MSE	2.16055	2.53776
50	**β_0**	1.435327	1.438414
	β_1	1.125387	1.162548

	β_2	1.382285	1.344776
	MSE	2.087716	2.240925
100	β_0	1.433842	1.443073
	β_1	1.12564	1.139315
	β_2	1.3153	1.309418
	MSE	1.719493	2.115933

نلاحظ من خلال الجدول رقم (6) أن طريقة Mهي أفضل طريقة تقدير مقارنة بالطرائق الأخرى لمختلف حجوم العينات بالطرائق الأخرى بالاعتماد على معدل متوسط مربعات الخطأ (AMSE).

4- الاستنتاجات والتوصيات :-

4-1- الاستنتاجات :

بالاعتماد على نتائج التقديرات للطرائق المستعملة في الجانب التطبيقي تم التوصل الى بعض من الاستنتاجات وكالآتي

1- نستنتج أن أفضل طريقة تقدير هي M بناء على معيار المقارنة AMSE .

2- أظهرت طريقة M تفوق في النتائج لأنها تمتلك أقل AMSE .

3- سجلت طريقة S الحصينة كأسوأ طريقة تقدير وذلك لأن معيار AMSE كان أعلى من نظيراته في الطرائق الأخرى .

4-2-التوصيات :

وخلاصة ما تقدم توصي الباحثة اعتماد دراسة خاصة لطرائق التقدير الحصينة جميعها المستعملة في هذه البحث في حالة وجود مشكلة تعدد الإبعاد لبيان إي من هذه الطرائق هي الأكفأ و استعمال حجوم عينات وعدد متغيرات توضيحية كبيرة لغرض بيان مدى فاعلية طرائق التقدير المدروسة لانحدار الحرف الحصين .

المصادر :

1-AdewaleLukman, OlatunjiArowolo and KayodeAyinde, Some Robust Ridge Regression for handling Multicollinearity and Outlier ,lnternational Journal of Scienes: Basic and Applied Research (IJSBAR).

2- CHUN YU and WEIXIN YAO, Robust Linear Regression : A Review and Comparison ,Chun Yu is Instructor, School of Statistics,Jiangxi University of Finance and Economics,Nanchang , 330013.

3-Chiang, L. H., Pell, R. J., &Seasholtz, M. B. (2003). Exploring process data with the use of robust outlier detection algorithms. Journal of Process Control, 13(5), 437-449.

4-Crone L.The singular value decomposition of matrices and cheap numerical filtering of systems of linear equations. J Franklin Inst 1972, 294:133–136.

5-Ehab M.Almetwally And Hisham M.Almongy , 2018, Comparison Between M-estimation, S-estimation , And MM Estimation Methods of Robust Estimation with Application and Simulation.

6-McDonald GC, Schwing, RC. Instabilities of regression estimates relating air pollution to mortality. Technometrics. 1973, 15:463–48.

7-Shrestha, N. (2020). Detecting multicollinearity in regression analysis.American Journal of Applied Mathematics and Statistics, 8(2), 39-42

8- Soner Cankaya , Samet Hasan Abacl , (2015) , A Comparatiue Study of Some Estimation Methods in Simple Linear Regression Model for Different Sample Sizes in Presence of Outliers , Turkish Journal of Agriculture – Food Science and Technology , 3(6) : 380-389 .

9-Yuliana Susanti , Hasih Pratiwi , Srib Sulistijowati H. , Twenty Liana , Sebelas Maret University , 2013 , International Jornal of Pure and Applied Mathematics , M Estimation , S Estimation , AND MM Estimation IN Robust Regression .