مقارنة بعض المقدرات الحصينة لتقدير معلمات انموذج الانحدار الخطي المتعدد بوجود نقاط الانعطاف العالية باستعمال المحاكاة

م.م. ارشد حميد حسن
جامعة ديالى / كلية الإدارة والاقتصاد
قسم الإحصاء
arshadeco@uodiyala.edu.iq

م.م شيماء محمد أحمد
كلية بلاد الرافدين الجامعة
قسم إدارة الأعمال
shimaamohammed@bauc14.edu.iq

ISSN 2709-6475 DOI: https://dx.doi.org/10.37940/BEJAR.2021.S.6

تأريخ استلام البحث 2021/5/16 **تأريخ قبول النشر 2021/7/26**

المستخلص

يعد نموذج الانحدار الخطي المتعدد أحد النماذج الخطية المستخدمة على نطاق واسع لتحليل العديد من البيانات البحثية في المجالات الاقتصادية والطبية والاجتماعية، ويهدف البحث إلى الحصول على مقدرات عالية الكفاءة. إن وجود مشادات شاذة تؤثر على هذه الكفاءة ولابد من الكشف عن وجود المشاهدات الشاذة (نقاط الانعطاف العالية) في المتغيرات التوضيحية ومعالجتها عن طريق تقدير المعلمات باستعمال بعض طرق التقدير الحصينة الموضحة من خلال البحث، مقدرM، ومقدر MM، ومقدر S ومقدر GM_2، ومن أجل معرفة أفضلية المقدرات تم استعمال أسلوب المحاكاة بين طرائق التقدير باختلاف حجوم العينات (n = 100,50,20) وافتراض نسب تلوث مختلفة (τ=10%,20%,30%) وبالاعتماد على معيار المقارنة متوسط مربعات الخطأ (MSE) للأنموذج للوصول إلى الطريقة الأفضل. أظهرت النتائج إن مقدر (MM) حقق كفاءة عالية في تقدير المعلمات مقارنة مع باقي المقدرات.

أما فيما يتعلق بالجانب التطبيقي من هذه الدراسة فقد تم توظيف بيانات حقيقية مأخوذة من الجهاز المركزي للإحصاء خاصة بنتائج استمارة استبيان مسح تقييم الأمن الغذائي والهشاشة للأسرة في العراق لسنة 2016، من خلال وصف البيانات المتمثلة بمعدل الانفاق الشهري التقديري لرب الأسرة على السلع والخدمات غير الغذائية، إذ نلاحظ وجود أربع مشاهدات ذات نقاط انعطاف عالية في البيانات وهذه المشاهدات هي (23,46,53,94)، إذ تم الكشف عنها بواسطة عناصر القطر الرئيس المصفوفة القبعة، وكذلك من نتائج الجانب التطبيقي نلاحظ معنوية متغير الانفاق الإجمالي لرب الأسرة على التعليم ومتغير الانفاق الإجمالي المسكن والكهرباء ومحروقات أخرى.

الكلمات المفتاحية: المشاهدات الشاذة، نقاط الانعطاف العالية، مقدر M، مقدر MM، مقدر S، مقدر GM_2.

مجلة اقتصاديات الأعمال
العدد (خاص – ج1) أيلول / 2021
الصفحات: 89-103

Comparison of some robust estimators for multiple linear regression model parameter estimation with the presence of high leverage points by simulation

Abstract

The multiple linear regression model is one of the linear models widely used to analyze many research data in the economic, medical and social fields, and the research aims to obtain high-efficiency capabilities that the presence of outliers affects this efficiency and the presence of outliers observations (high leverage points must be detected), In the illustrative variables and treating them by estimating the parameters using some of the invulnerable estimation methods illustrated through the research, the M estimator, the MM estimator, the S estimator and the GM2 estimator. In order to know the best of the estimators, a simulation method was used between estimation methods with different sample sizes ($n = 100, 50, 20$) and assuming different Contamination ratios ($\tau = 10\%, 20\%, 30\%$) and based on the comparison criterion average squares of error (MSE) For the model to reach the best method, the results showed that the (MM) estimator achieved high efficiency in estimating the parameters compared with the rest of the estimators.

As for the application side of this study, real data taken from the Central Bureau of Statistics was employed regarding the results of the survey questionnaire for the Food Security and Vulnerability Assessment of Households in Iraq for the year 2016, by describing the data represented by the estimated average monthly expenditure of the head of the household on non-food goods and services. We notice the presence of four observations with high leverage points in the data, and these observations are (23,46,53,94) as they were revealed by the main country components of the Hat Matrix, as well as from the results of the application, we note the significance of the variable of the head of the household's total expenditure on education and the variable of total housing spending Electricity and other fuels.

Key words: Outliers, High Leverage Points, M estimator, M estimator, S estimator, GM_2 estimator.

المقدمة:

تكمن فلسفة الإحصاء من حيث آلية التطبيق في محاولة نمذجة الظواهر المختلفة بنماذج أقرب ما يمكن إلى الواقع الفعلي، وإن هذه النماذج تقاس درجة قوتها بحسب درجة تقاربها مع الخواص الإحصائية وهي على أشكال وأنواع مختلفة فمنها الاحتمالي والتي تعتمد في صياغتها على الاحتمالات الصرفة (نماذج السلاسل الزمنية، نماذج سلاسل ماركوف، نماذج المعقولية) ومنها النماذج السببية، والتي تقوم صياغة نماذجها على ما يعرف بالسبب ونتيجة السبب وتأتي في مقدمة هذه النماذج ما تسمى بنماذج الانحدار، إذ تقوم نماذج الانحدار باستكشاف العلاقة ما بين السبب والذي يعرف إحصائياً بالمتغيرات التوضيحية (التفسيرية) وبين ما هو نتيجة السبب أو ما يعرف بمتغير الاستجابة (المتغير المعتمد).

تعزى طريقة المربعات الصغرى الاعتيادية ((Ordinary Least Squares (OLS) إلى عالم الرياضيات الألماني (Carl Friedrich Gauss)، إن مقدرات المربعات الصغرى الاعتيادية (OLS) تحت الافتراضات الأساسية هي أفضل تقدير خطي غير متحيز (Linear Unbiased Estimator Best)(BLUE) نظراً لتحقق جميع الافتراضات الخاصة بها، فضلاً عن ذلك تكون الأخطاء العشوائية مستقلة وموزعة بشكل متماثل، بعبارة أخرى تكون لمقدرات المربعات الصغرى الاعتيادية (OLS) ذات التباين الأقل بين جميع المقدرات الخطية، إن وجود مشاهدات شاذة في البيانات يؤدي إلى عدم تحقق بعض الافتراضات الخاصة بتقدير طريقة المربعات الصغرى الاعتيادية، إذ أن في أغلب الأحيان تؤدي المشاهدات الشاذة إلى مشكلتين الأولى هي عدم توزيع الأخطاء الخاصة بنموذج الانحدار طبيعاً والثانية تؤدي إلى ظهور مشكلة التعدد الخطي شبه التام.

وإن وجود المشاهدات الشاذة (Outliers) في البيانات التي تعد إحدى الصعوبات في بناء أنموذج الانحدار والتي تصنف ثلاثة أنواع، قيم شاذة في المحور Y (قيم شاذة عامودية)، قيم شاذة في كلا الاتجاهين (X Y)، قيم شاذة في المحور X (نقاط الانعطاف العالية) (High Leverage Points)(HLPs)، ولكن ما يثير اهتمامنا هو تحديد نقاط الانعطاف العالية (HLPs) بسبب تأثيرها الكبير على المقدرات المختلفة، التي تعرف على أنها (تلك المشاهدات التي تقع في نطاق المحور الاحداثي X وتكون على نوعين (جيدة، رديئة).

ومن البديهي أن نقول إن سلامة البيانات هي مسألة ضرورية لسلامة النتائج بغض النظر عن التحليل المعتمد، عليه وباعتبار ان تحليل الانحدار هو تحليل واسع الانتشار والتطبيق فتصبح مهمة تنقية بياناته من المشاهدات الشاذة مهمة ضرورية لسلامة النتائج والاستنتاجات المبنية عليها.

أولاً: منهجية البحث:

1. مشكلة البحث:

إن مشكلة البحث تكمن في تقدير معلمات أنموذج الانحدار الخطي المتعدد في ظل وجود المشاهدات الشاذة التي تنحرف بشكل ملحوظ عن المشاهدات الأخرى التي تقع في نطاق المحور الاحداثي X والتي تدعى نقاط الانعطاف العالية مما نحصل على تأثيرات مقنعة (غير ظاهرة) وذلك يؤثر سلباً على دقة النتائج وغالباً ما تؤدي الى استنتاجات مظللة.

2. هدف البحث:

يهدف هذا البحث إلى المقارنة بين المقدرات الحصينة لأنموذج الانحدار الخطي المتعدد في ظل وجود المشاهدات الشاذة ذات الانعطاف العالية وسوف يتم المقارنة باستعمال معيار متوسط مربعات الخطأ (Mean Square Error) من خلال اسلوب المحاكاة بطريقة مونت كارلو (Monte-Carlo)، فضلاً عن تطبيقها على بيانات حقيقية مأخوذة من الجهاز المركزي للإحصاء والخاصة بـ(إنفاق الاسرة في جوانب اجتماعية عدة) وبالتالي الحصول على أفضل تقدير.

ثانياً: الجانب النظري:

في هذا الجانب يتم التطرق إلى أنموذج الانحدار الخطي المتعدد وطريقة المربعات الصغرى الاعتيادية (OLS) كما يتضمن عرض بعض الطرائق الحصينة القادرة على التعامل مع مشكلة وجود نقاط الانعطاف العالية ليكون مدخلاً لما يليه من الجانب التطبيقي.

1. انموذج الانحدار الخطي المتعدد Multiple Linear Regression Model:

إن تحليل الانحدار الخطي المتعدد يهدف إلى دراسة وتحليل أثر عدة متغيرات توضيحية على متغير الاستجابة، إذ يستخدم أنموذج الانحدار الخطي المتعدد للتنبؤ للقيم المستقبلية عن طريق تقدير معلمات النموذج التي تعتمد في النموذج التقديري للتنبؤ، وأن العلاقة الخطية بين عدة متغيرات توضيحية ومتغير الاستجابة يطلق عليها بـ(الانحدار الخطي المتعدد)، الذي يأخذ شكلاً رياضياً خطياً صيغته كالآتي: (كاظم ومسلم، 2002: 50)

$$Y_i = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + e_i \quad , \quad i = 1,2,\ldots\ldots,n \, , \quad j = 1,2,\ldots\ldots,k \ldots\ldots\ldots(1)$$

ويمكن التعبير عن المعادلة (1) بأسلوب المصفوفات وكما يأتي:

$$\mathrm{Y} = \mathrm{X}'\beta + U \ldots\ldots\ldots(2)$$

إذ أن:

Y: يمثل متجه مشاهدات متغير الاستجابة من الدرجة $(n \times 1)$.

X: يمثل مصفوفة مشاهدات المتغيرات المستقلة من الدرجة $((n \times (k+1))$.

β: يمثل متجه المعلمات المجهولة من الدرجة $((k+1) \times 1)$.

k : عدد المتغيرات التوضيحية.

n : عدد المشاهدات.

U: يمثل موجه الأخطاء العشوائية من الدرجة $(n \times 1)$.

غالباً ما تستعمل طريقة المربعات الصغرى الاعتيادية (OLS) لتقدير معلمات الانحدار لأنموذج الانحدار الخطي المتعدد حيث تأخذ الصيغة الآتية:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \ldots\ldots\ldots(3)$$

إن أنموذج الانحدار بني على مجموعة من الفروض للحصول على أفضل تقدير لمعلمات الأنموذج باستعمال طريقة المربعات الصغرى الاعتيادية (OLS) وعند تحقق جميع هذه الفروض نحصل على أفضل التقديرات لمعلمات الأنموذج باستعمال طريقة المربعات الصغرى الاعتيادية الـ(OLS) وتتميز مقدراتها بخاصية أفضل تقدير خطي غير متحيز (BLUE)، أما في حالة عدم

تحقق أو غياب أحد هذه الفروض فإنها ستؤدي إلى مشاكل كثيرة في التقديرات، ومن هذه المشاكل مشكلة عدم توزيع متجه الأخطاء للتوزيع الطبيعي، وهذا يعود إلى وجود مشاهدات شاذة في البيانات قيد الدراسة.

2. المشاهدات الشاذة Outliers:

عرف Bross المشاهدات الشاذة التي تظهر منحرفة بشكل كبير عن سائر مكونات العينة التي وجدت فيها تلك العينة. أما Freeman فقد عرف المشاهدة الشاذة بأنها أية مشاهدة لم تتولد بالطريقة العامة التي ولدت الأغلبية العظمى من مشاهدات البيانات. ويمكن تعريفها إحصائياً بأنها المشاهدة المتأتية من مجتمع مختلف عن المجتمع قيد البحث. ويمكن تعريف بعض أنواع المشاهدات الشاذة على النحو الآتي:

- **المشاهدة المتطرفة Extreme Observation:**

هي المشاهدات المتأتية من نفس توزيع المتغير (المتغيرات) أي من نفس المجتمع ولكنها تتميز عن بقية المشاهدات إما بكبرها أو صغرها.

- **المشاهدات الشاذة المتطرفة Extreme Outlier:**

هي المشاهدات التي تقع في آخر ذيل البيانات بعد ترتيبها تصاعدياً أو تنازلياً أي إنها تكون أكبر من (3σ) أو أصغر من (-3σ) في حالة رسم البيانات تحت منحنى التوزيع الطبيعي القياسي.

- **نقاط النفوذ العالية High Leverage Points:**

هي النقاط التي يكون المتجه الصفي (xi) من المصفوفة (X) بعيدة عن بقية متجهات البيانات الأخرى، بمعنى آخر فإن المشاهدات الواقعة في قطر المصفوفة $(H = X(X'X)^{-1}X')$ تكون كبيرة قياساً بنظيرتها من المشاهدات الأخرى في فضاء المتغيرات التوضيحية (X) ولا تخص متغير الاستجابة (Y).

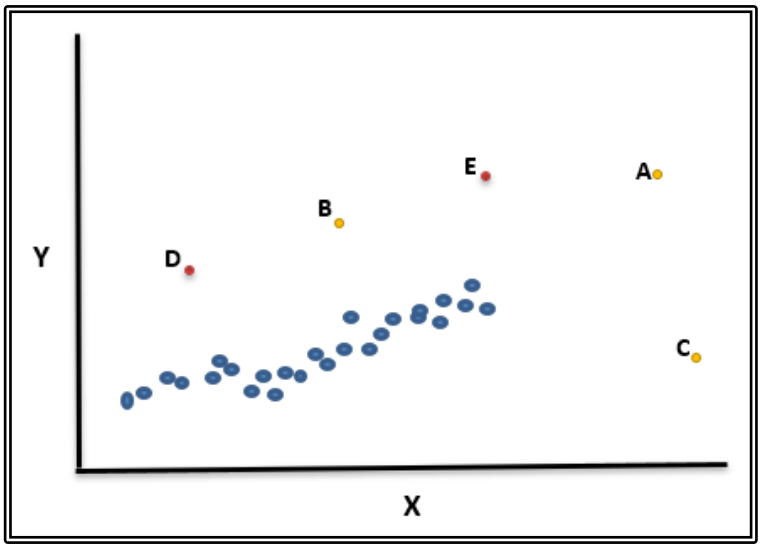
- **المشاهدة المؤثرة Influential Observation:**

هي تلك المشاهدات التي تؤثر بشكل مفرط، انفرادياً أو مجتمعة على توفيق معادلة الانحدار مقارنة بنظريتها من المشاهدات الأخرى لمجموعة البيانات (ناسي والجبوري، 2001: 3).

ومما تقدم يمكن ملاحظة ما يأتي:

1. المشاهدات الشاذة ليست بالضرورة مشاهدات مؤثرة.
2. نقاط النفوذ العالية ليست بالضرورة نقاطاً مؤثرة.
3. المشاهدات التي لها بواقي كبيرة تكون مشاهدات غير مرغوب فيها لأن توفيق المربعات الصغرى يتأثر بالبواقي الكبيرة والمشاهدات التي لها باقي صغير لا تعني بالضرورة إنها مشاهدات طبيعية لأن نقاط النفوذ العالية لها بواقي صغيرة وتؤثر على توفيق الانموذج.

ويمكن توضيح المشاهدات الشاذة من الشكل الانتشاري الآتي:



الشكل (1) أنواع المشاهدات الشاذة

إذا أضفنا كلاً من النقاط A, B, C بشكل انفرادي إلى نقاط الانموذج نستنتج أن:

فيما يتعلق بالنقطة A لها باقي صغير (لأن قيمة (Y) قريبة من المستقيم المار ببقية النقاط)، نقطة نفوذ عال لأنها شاذة بقيمة (X) وليس لها تأثير على توفيق معادلة الانحدار أي إنها نقطة نفوذ عال ولكنها غير مؤثرة.

أما فيما يتعلق بالنقطة B ليست نقطة نفوذ عال (لأنها تقع في مركز (X) لكنها مشاهدة شاذة (لها باقي كبير) ونقطة تأثير (دخولها لا يغير الميل ولكنه يغير نقطة التقاطع مع المحور(Y)).

أما فيما يتعلق بالنقطة C فهي مشاهدة شاذة (لها باقي كبير) نقطة نفوذ عال (لأنها نقطة متطرفة في فضاء (X) ومشاهدة مؤثرة (لأنها تغير توفيق معادلة الانحدار).

وبإضافة النقطتين D وE معاً إلى نقاط الأنموذج نلاحظ أن:

النقطة D هي مشاهدة شاذة ولكنها ليست مؤثرة وليست نقطة نفوذ عال.

أما النقطة E فهي مشاهدة مؤثرة لأنها تغير توفيق معادلة الانحدار ولكنها ليست شاذة (لها باقي صغير) وليست نقطة نفوذ (Chatterjee & Hadi,1990:135).

3. طرق الكشف عن المشاهدات الشاذة:

عرف الباحثان Hocking & Pendleton في عام (1983م) نقاط الانعطاف العالية (HLPs) بأنها (تلك المشاهدات في قيم المتغير Xi التي تكون بعيدة عن باقي البيانات)، ويمكن اعتبارها على أنها قيم خارجية مؤثرة في محور المتغير X والتي تسبب في انعطاف خط الانحدار نحوها، وبالتالي يرتبط مفهوم نقاط الانعطاف العالية (HLPs) بالكامل بالمتغيرات التوضيحية وليس بالمتغير التابع (Chatterjee & Hadi,1990:55).

إن وجود نقاط الانعطاف العالية في بيانات أنموذج الانحدار الخطي المتعدد يمكن أن يضلل الاستنتاج ويسبب بعض المشاكل الإحصائية الأخرى، قد لأيتم تحديدها من خلال ملاحظة بواقي المربعات الصغرى لأنها تميل إلى أن تكون لها بواقي صغيرة جداً، مما يكون لها تأثيراً قوياً على

معلمات الانحدار المتوقعة وتؤدي إلى مشاكل أكثر خطورة من القيم الشاذة باتجاه Y (Kim,2004:4).

وإن طرائق التشخيص التقليدية قد تفشل أحياناً في تحديد نقاط الانعطاف العالية بسب مشاكل الاخفاء والغرق، لهذا السبب تم تطوير العديد من طرائق التشخيص للمساعدة في تحديد نقاط الانعطاف العالية (HLPs) نذكر منها: (Kamruzzaman & Imon,2002:436)

1-3 مصفوفة القبعة (Hat Matrix):

تلعب مصفوفة القبعة دوراً مهماً في تحديد المشاهدات الشاذة التي تقع في فضاء X للمتغيرات التوضيحية، ويتم اكتشافها بواسطة عناصر القطر الرئيس (hii) لمصفوفة H، وهو مقياس للتأثيرات المحتملة في المتغيرات التوضيحية وتكون بالشكل الآتي:

$$H = X(X'X)^{-1}X' \quad \text{.....(4)}$$

$$\underline{\hat{Y}} = HY \quad \text{.....(5)}$$

وأيضاً يمكن كتابتها بدلالة عناصر هذه المصفوفة كالآتي:

$$\hat{Y} = \sum_{j=1}^{n} h_{ij} Y_j \quad \text{.....(6)}$$

أن مصفوفة H هي مصفوفة إسقاط (projection matrix) تكون قيمها المميزة أما صفر أو واحد، أن المعادلة (6) توضح أن عناصر h_{ij} تمثل مقياس وزن المشاهدة لـ Y_j نسبياً إلى كل القيمة التقديرية $\hat{Y}$ وعليه فإن (h_{ii}) هي عناصر القطر الرئيسي لمصفوفة (H)، تعد المشاهدة لها قوة انعطاف عالية إذا كانت ($h_{ii} > (\frac{2p}{n})$ ، إذ أن p يمثل عدد المعلمات في نموذج الانحدار (Hoaglin,*et.al.*,1978:17).

4. المقدرات الحصينة Robust Estimators:

عند انتهاك بعض الافتراضات الأساسية، الخاصة بتقدير معلمات أنموذج الانحدار الخطي المتعدد باستعمال طريقة المربعات الصغرى الاعتيادية تؤدي إلى تقديرات غير مستقرة بسبب وجود المشاهدات الشاذة في البيانات، ففي هذه الحالة يحل الانحدار الحصين محل طريقة المربعات الصغرى العادية (OLS) لتحقيق هذا الاستقرار، إذ تحد طرائق الانحدار الحصينة من تأثير الشواذ عن طريق تقليل أوزانها أو تغييرها، هنالك العديد من المقدرات الحصينة نذكر منها:

1-4 مقدر M (M estimator):

يعد مقدر M أحد أكثر طرق الانحدار الحصين شيوعاً التي اقترحها هوبر (1973م)، وعلى الرغم من أنه لا يتأثر بنقطة الانعطاف، إلا أنه يعد أبسط طريقة نظرياً، ويمكن للطريقة M تصغير دالة الهدف بدلاً من تقليل مجموعة مربعات الخطأ لدالة الهدف (Chen,2002:4).

إن مقدر M يتم الحصول عليه من خلال تقليل الصيغة الآتية:

$$\min \sum_{i=1}^{n} p\left(\frac{e_i}{S}\right) = \min \sum_{i=1}^{n} p\left(\frac{y_i - \acute{x}_i \beta}{S}\right) \quad \text{.....(7)}$$

إن S هي تقدير القياس يتم ايجادها وفقاً للمعادلة الآتية:

$$S = \frac{MAD}{0.6745} = \frac{\text{median}|e_i - \text{median}(e_i)|}{0.6745} \quad \text{.....(8)}$$

MAD : متوسط الانحراف المطلق.

يكون نظام حل المعادلات الطبيعية بالنسبة لمشكلة التقليل (Minimization problem) عن طريق أخذ المشتقات الجزئية بالنسبة لـ β ومساواتها للصفر لنحصل على المعادلة الآتية:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \acute{x}_i\hat{\beta}}{S}\right) x_i = 0 \quad \text{.....(9)}$$

p: هو مشتق من ψ

$$\psi = p'$$

يعتمد اختيار الدالة ψ على افضلية مقدار الوزن لتحديد القيم الشاذة.

إن الدالة p لها العديد من الخصائص منها:

$$p(e) \geq 0\,, p(0) = 0\,, p(e) = p(-e)\,, p(e_i) \geq p(\acute{e}_i) \text{ for } |e_i| \geq |\acute{e}_i|$$

وقد أعطى كل من (Draper & Smith) الحل للمعادلة (2-19) من خلال تعريف دالة الوزن الآتية:

$$w(e_i) = \frac{\psi\left(\frac{y_i - \acute{x}_i\hat{\beta}}{S}\right)}{\left(\frac{y_i - \acute{x}_i\hat{\beta}}{S}\right)} \quad \text{.....(10)}$$

عليه فإن المعادلة (9) ممكن أن تكتب وفقاً للصيغة الآتية:

$$\sum_{i=1}^{n} w_i\left(\frac{y_i - \acute{x}_i\hat{\beta}}{S}\right) x_i = 0$$

إن المعادلة أعلاه ممكن أن تكتب بصيغة المصفوفات وكالآتي:

$$X'WX\hat{\beta}_M = X'Wy \quad \text{.....(11)}$$

إذ أن:

W: مصفوفة قطرية عناصرها القطرية هي الاوزان (w_i).

ومن خلال حل المعادلة (11) نحصل على مقدر M الذي يكون وفقاً للصيغة الآتية:

$$\hat{\beta}_M = \left(\acute{X}WX\right)^{-1}\acute{X}WY \quad \text{.....(12)}$$

2-4 مقدر MM (MM estimator):

يعد مقدر MM أحد أكثر الطرق شيوعًا في مجال انحدار الحصين، وقد اقترحه Yohai في عام 1987. للمقدر العديد من الخصائص الجيدة، بما في ذلك الكفاءة العالية في حالة التوزيع الطبيعي للأخطاء مع نقطة الانهيار العالية، المقدر (MM) في الاسم يشير إلى عمليات متعددة باستخدام مقدر M، ويمكن الحصول على مقدر الانحدار (MM) وفقًا للخطوات الآتية: (Yuliana, *et.al.*,2014:356)

1. يتم تحديد مقدر أولي ذو نقطة انهيار عالية، ولكن ليس بالضرورة أن يكون كفوء، نرمز له بالرمز $(\hat{\beta}_s)$ وباستعماله يتم حساب البواقي الأولية وفقاً للصيغة الآتية:

$$r_i(\hat{\beta}_s) = y_i - x_i'\hat{\beta}_s \quad , \quad 1 < i < n \quad \text{.....(13)}$$

2. يتم حساب مقدر M القياس (S_n) للبواقي الأولية $r_i(\hat{\beta}_s)$ وفق معادلة M التقديرية لمعلمة القياس وبالشكل الآتي:

$$\frac{1}{n}\sum_{i=1}^{n} p\left(\frac{r_i(\hat{\beta}_s)}{S}\right) = 0.5 \quad \text{.....(14)}$$

$$S_n = S\left(r_1(\hat{\beta}_s), \ldots, r_n(\hat{\beta}_s)\right)$$

3. مقدر(MM) يعرف كمقدر M لـβ باستعمال دالة (re-descending score):

$$\psi_1(u) = \frac{\partial\, p_1(u)}{\partial u} \quad \text{.....(15)}$$

عليه فإن مقدر(MM) الذي نرمز له بالرمز $\hat{\beta}_{MM}$ يكون الحل للمعادلة الآتية:

$$\sum_{i=1}^{n} X_{ij}\psi_1\left(\frac{y_i - x_i'\beta}{S_n}\right) = 0 \quad , \quad j = 1,2,\ldots,n \quad \text{.....(16)}$$

إن تقدير القياس S_n يتم ايجاده في الخطوة (2).

إذ أن:

$$r_i(\beta) = y_i - x_i'\beta$$

3-4 مقدر GM2 (GM2 estimator):

يعد مقدر (GM2) إحدى طرائق التقدير الحصينة، إذ اقترح من الباحثة (Bagheri) عام (2011) أن فكرة هذه الطريقة تقوم على اساس التعديل لمقدر (Generalized Modified (GM) M-estimator) للحصول على مقدر حصين ذو خصائص عالية الكفاءة وأكثر مقاومة لنقاط الانعطاف العالية، يتم تلخيص الطريقة من خلال عدة خطوات على النحو الآتي: (Alguraibawi, *et.al.*,2015:311)

1. يتم حساب البواقي الأولية (r_i) من مقدر s، ومن ثم أيجاد مقياس للبواقي ($\hat{\tau}$) وكالآتي:

$$r_i = y_i - \hat{y}_i \qquad i=1,2...,n \quad \text{.....(17)}$$

$$\hat{\tau} = 1.4826\left(\frac{1+5}{n-p}\right) \text{Median}\,|r_i| \quad \text{.....(18)}$$

2. يتم ايجاد مصفوفة الأوزان القطرية (W)، وعناصر القطر هي الأوزان (W_i) من خلال المعادلة الآتية:

$$w_i = \min[1, \{\frac{x^2_{(0.95,\ p+1)}}{RMD^2}\}] \qquad i=1,2..., n \quad \text{.....(19)}$$

(RMD) تمثل مسافة (Mahalanobis Distance) الحصينة بالاعتماد على المقدار الأدنى للمقياس الإهليجي (MVE).

3. حساب دالة التأثير Ψ^* للبواق المعيارية، وذلك بحل المعادلة الآتية:

$$A = diag\ \Psi^*\left(\frac{r_i}{\hat{\tau} * w_i}\right) \quad \text{.....(20)}$$

Ψ^* : مشتقة من دالة التأثير لـ (Hubeŕs).

4. عليه فإن مقدر (GM_2) والذي نرمز له بالرمز ($\hat{\beta}_{GM2}$) يمكننا الحصول عليه عن طريق اشتقاق خطوة واحدة لطريقة (Newton Raphson) ويكون بالصيغة الآتية:

$$\tilde{\alpha}_{GM2} = \hat{\beta}_0 + (X'A\,X)^{-1}X'W\psi\left(\frac{r_i}{w_i\hat{\tau}}\right)\hat{\tau} \quad \ldots\ldots(21)$$

5. معيار المقارنة:

هناك العديد من المعايير التي يمكن استخدامها للمقارنة بين النماذج، وسنستخدم معيار متوسط مربعات الخطأ (MSE) كمقياس للدقة، ونحسبه من خلال مربع الفرق بين القيم الحقيقة والقيم التقديرية لمتغير الاستجابة، ومن ثم ايجاد المعدل أو متوسط القيم لمجموع هذه المربعات وتكون صيغته على النحو الآتي: (الراوي، 1987: 84)

$$MSE = \frac{1}{n-p}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad \ldots\ldots(22)$$

إذ أن:

y_i: يمثل القيم الحقيقية لأنموذج الانحدار.

$\hat{y}_i$: يمثل القيم التقديرية لأنموذج الانحدار.

p: يمثل عدد المتغيرات التوضيحية.

n: يمثل حجم العينة.

ثالثاً: الجانب التجريبي:

في هذا الجانب تم كتابة برنامج بلغة البرمجة الإحصائية R للمقارنة بين المقدرات الحصينة واختبار الأفضل منها بالاعتماد على معيار متوسط مربعات الخطأ، أما بالنسبة للأنموذج الذي تم الاعتماد عليه في البحث يكون بالشكل الآتي:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i \quad i = 1,2,..,n \ldots\ldots(23)$$

إذ يتم وصف تجارب المحاكاة على النحو الآتي:

يتم تحديد القيم الافتراضية للمعلمات وتعد هذه الخطوة من الخطوات الأساسية التي تعتمد عليها بقية المراحل، علماً بأن هذه القيم تم تحديدها من البيانات الحقيقية بعد تقديرها باستعمال طريقة المربعات الصغرى الاعتيادية وفقاً للقيم الآتية: $P = 5$ وبافتراض أن عدد المعالم هي:

$$\beta_1 = 0.255\,,\beta_2 = -0.167\,,\beta_3 = 0.101\,,\beta_4 = -0.055\,,$$
$$\beta_5 = -0.039 \quad ,\ \beta_0 = 0$$

إذ تم اختيار ثلاثة أحجام مختلفة للعينات المفترضة وهي: (n=20,50,100)

أما المرحلة الثانية فهي توليد المتغيرات التوضيحية X_{ij} ويتم توليد خمس متغيرات توضيحية واعتماد نسب مختلفة لتلوث قيم المتغيرات التوضيحية، وذلك بافتراض ثلاثة مستويات وهذه النسب وهي (30%, 20%, 10%=τ)، τ تمثل نسبة التلوث المفترضة للدراسة.

وثم بعد ذلك يتم توليد الأخطاء العشوائية، إذ يتم توليد الأخطاء العشوائية وفقاً للتوزيع الطبيعي بمتوسط (صفر) وتباين (σ^2) أي إن:

$$\boldsymbol{\varepsilon_i} \sim (0, \sigma^2) \quad , \quad i = 1,2,\ldots,n$$

وبالنسبة لقيم التباين هي ($\sigma^2 = 1\,,5\,,10$).

ثم بعد ذلك يتم تقدير معلمات انموذج الانحدار الخطي المتعدد وفق طرائق التقدير التي تم عرضها في الجانب النظري من البحث مقدر M ومقدر MM ومقدر S ومقدر(GM2) ثم تتم المقارنة بين هذه المقدرات بالاعتماد على معيار متوسط مربعات الخطأ (MSE) كمقياس لمقارنة متوسطات تقدير المعلمات وذلك تبعاً للصيغة رقم (22). وسوف يكون عدد مرات تكرار التجربة (R=1000) مرة.

الجدول (1) قيم معيار (MSE) للمقدرات

Sample Size	σ^2	% of Outliers	M	MM	S	GM2	Best
n=20	1	10	0.007748023	0.006400924	0.007523304	0.008922644	MM
		20	0.017268773	0.006374796	0.007203637	0.020369823	MM
		30	0.02432594	0.02081265	0.02445781	0.02357760	MM
	5	10	0.007628370	0.006462872	0.007934373	0.009403455	MM
		20	0.013824142	0.006459577	0.006855019	0.015900591	MM
		30	0.02114923	0.01276113	0.01324141	0.01972557	MM
	10	10	0.007407284	0.006564169	0.008183742	0.008756414	MM
		20	0.010039306	0.006255763	0.006592703	0.012342149	MM
		30	0.016470046	0.008534502	0.009372921	0.016736670	MM
n=50	1	10	0.002179405	0.002124349	0.002279036	0.002370762	MM
		20	0.002434983	0.002122546	0.002177015	0.004022463	MM
		30	0.004660105	0.002131653	0.002294961	0.004289055	MM
	5	10	0.004206366	0.002176325	0.002190369	0.004043173	MM
		20	0.002341734	0.002112709	0.002155791	0.003430618	MM
		30	0.004136220	0.002159433	0.002165065	0.003964218	MM
	10	10	0.002168139	0.002128804	0.002248239	0.002421531	MM
		20	0.002236559	0.002143895	0.002171270	0.003019414	MM
		30	0.003114207	0.002136617	0.002143158	0.003484995	MM
n=100	1	10	0.001055981	0.001051734	0.001085098	0.001089298	MM
		20	0.001080341	0.001060348	0.001069992	0.001453700	MM
		30	0.001693492	0.001056682	0.001058606	0.001565633	MM
	5	10	0.001064518	0.001057032	0.001089325	0.001131797	MM
		20	0.001079584	0.001063543	0.001073561	0.001347837	MM
		30	0.001570012	0.001064330	0.001064481	0.001535427	MM
	10	10	0.001070958	0.001067251	0.001093814	0.001130828	MM
		20	0.001069036	0.001059116	0.001066079	0.001258182	MM
		30	0.001255964	0.001066723	0.001067907	0.001415895	MM

بعد تنفيذ وإجراء تجربة المحاكاة استخلصت النتائج وفسرت بأخذ جميع العوامل المؤثرة بنظر الاعتبار فمن خلال ملاحظة الجدول (1)، نلاحظ إن أفضل مقدر عندما كانت حجم العينة (n=20) بمختلف مستويات التلوث والتباين هو مقدر MM، أما عندما زاد حجم العينة إلى (n=50)

فنلاحظ إن أفضل مقدر هو مقدر MM الحصين أيضاً وعندما كانت حجم العينة (n=100) فإن مقدر MM الحصين هو الأفضل، وبصورة عامة نلاحظ إن أفضل مقدر لتقدير انموذج الانحدار الخطي بوجود مشاهدات شاذة (نقاط انعطاف عالية) بمختلف نسب التلوث وبمختلف حجوم العينات ومعاملات الارتباط هو مقدر MM الحصين ثم يليه بالمرتبة الثانية مقدر S الحصين.

رابعاً: الجانب التطبيقي:

يتضمن هذا الجانب دراسة بيانات حقيقية مأخوذة من الجهاز المركزي للإحصاء خاصة بنتائج استمارة استبيان مسح تقييم الأمن الغذائي والهشاشة للأسرة في العراق لسنة 2016، من خلال وصف البيانات المتمثلة بالدخل الشهري التقديري لرب الاسرة على السلع والخدمات غير غذائية وبعينة حجمها (100) ولغرض تسهيل مهمة التحليل هذه البيانات فقد تم اعتبار المتغير التابع (Y) هو معدل الانفاق الشهري التقديري لرب الاسرة والمتغيرات التوضيحية ($X's$) هي:

1. المتغير (X_1) يمثل الانفاق الاجمالي لرب الاسرة على النقل والمواصلات.
2. المتغير (X_2) يمثل الانفاق الاجمالي لرب الاسرة على الصحة.
3. المتغير (X_3) يمثل الانفاق الاجمالي لرب الاسرة على التعليم.
4. المتغير (X_4) يمثل الانفاق الاجمالي لرب الاسرة على الملابس واحذية.
5. المتغير (X_5) يمثل الانفاق الاجمالي المسكن والكهرباء ومحروقات اخرى.

الكشف عن المشاهدات الشاذة (نقاط الانعطاف العالية)

سوف يتم الكشف عن وجود المشاهدات الشاذة نقاط الانعطاف العالية (HLPs) في المتغيرات التوضيحية باستعمال مصفوفة (Hat Matrix)، إذ يتم الكشف عنها من خلال عناصر القطر الرئيس لمصفوفة القبعة وكما مبين في الجدول الآتي:

الجدول (2) الكشف عن نقاط الانعطاف العالية باستعمال مصفوفة القبعة

obs	h_{ii}	obs	h_{ii}	obs	h_{ii}	obs	h_{ii}
1	0.062143	24	0.043935	51	0.067644	74	0.075192
2	0.042761	25	0.018858	52	0.012076	75	0.017975
3	0.051378	26	0.058646	53	0.140221	76	0.039802
4	0.064968	27	0.018587	54	0.028358	77	0.046113
5	0.060675	28	0.045973	55	0.018678	78	0.037873
6	0.070001	29	0.087643	56	0.089325	79	0.067255
7	0.056523	30	0.042064	57	0.031641	80	0.04672
8	0.041883	31	0.037213	58	0.011052	81	0.05463
9	0.086476	32	0.047058	59	0.056561	82	0.053082
10	0.06662	33	0.067031	60	0.048292	83	0.032039
11	0.047345	34	0.073905	61	0.042628	84	0.064929
12	0.035163	35	0.01232	62	0.059037	85	0.038267
13	0.051418	36	0.010008	63	0.023758	86	0.056615
14	0.070973	37	0.071374	64	0.08771	87	0.049618
15	0.031057	38	0.054702	65	0.04033	88	0.030131
16	0.084219	39	0.075819	66	0.043782	89	0.010387

obs	h_{ii}	obs	h_{ii}	obs	h_{ii}	obs	h_{ii}
17	0.06973	40	0.07644	67	0.033428	90	0.019471
18	0.015863	41	0.022819	68	0.017388	91	0.083725
19	0.059051	42	0.052679	69	0.052242	92	0.023117
20	0.06498	43	0.061605	70	0.049234	93	0.026627
21	0.05694	44	0.02205	71	0.048802	94	0.184951
22	0.058135	45	0.046068	72	0.017352	95	0.053015
23	0.119926	46	0.111466	73	0.05312	96	0.033233

نلاحظ من الجدول (2) الخاص بقيم عناصر القطر الرئيس لمصفوفة القبعة Hat Matrix وعندما تكون $h_{ii} > (\frac{2p}{n})$ تعد المشاهدات شاذة وذو نقاط انعطاف عالية، إذ كان $h_{ii} > 0.10$ لذلك فإن المشاهدات التي تعد بأنها نقاط انعطاف عالية هي: (23,46,53,94) كون قيم hii الخاصة بيها أكبر من 0.10.

1. تقدير معلمات الانموذج:

سوف يتم تقدير معلمات الانموذج باستعمال أفضل طريقة من طرائق التقدير التي تحققت في الجانب التجريبي للبحث وهو مقدر (MM) وذلك على أساس إنه أعطى أقل قيم لمعيار متوسط مربعات الخطأ مقارنة مع الطرائق الأخرى وبالاعتماد على عينة التطبيق المتمثلة بالبيانات الحقيقية لمعدل الانفاق الشهري التقديري لرب الاسرة في جوانب متعددة، تم كتابة برنامج تقدير المعلمات باستعمال برنامج بلغة (R)، وتم الحصول على النتائج وفق الجدول الآتي:

الجدول (3) يبين القيم التقديرية للمعلمات

Coefficients	Value	Std. Error	t value	p.value
Intercept	0.0165	0.0667	0.2472	0.3858025
X1	-0.0084	0.0677	-0.1236	0.3948323
X2	-0.0340	0.0678	-0.5012	0.3505180
X3	0.1590	0.0623	2.5521	0.0164974
X4	-0.0602	0.0673	-0.8952	0.2658626
X5	0.1785	0.0625	2.8560	0.0076316

من نتائج الجدول (3) نلاحظ معنوية متغير الانفاق الاجمالي لرب الأسرة على التعليم (x3)، إذ أن زيادة وحدة واحدة من الانفاق الاجمالي لرب الاسرة على التعليم يؤدي إلى زيادة معدل الانفاق الشهري لرب الاسرة بمقدار (10%)، وكما نلاحظ من الجدول معنوية متغير الانفاق الاجمالي المسكن والكهرباء ومحروقات أخرى، إذ أن زيادة متغير الانفاق الاجمالي المسكن والكهرباء ومحروقات أخرى بوحدة واحدة يؤدي إلى زيادة معدل الانفاق الشهري لرب الاسرة بمقدار (7%).

خامساً: الاستنتاجات:

بناءٍ على مخرجات الجانب التجريبي والتطبيقي يستنتج الباحث ما يأتي:

1. إن أفضل مقدر لتقدير أنموذج الانحدار الخطي المتعدد بوجود مشاهدات شاذة من نوع نقاط الانعطاف العالية هو مقدر MM الحصين بمختلف حجوم العينات ونسب التلوث كونه يمتلك أقل متوسط مربعات خطأ مقارنة ببقية المقدرات.
2. من نتائج الجانب التطبيقي نلاحظ وجود أربع مشاهدات ذات نقاط انعطاف عالية في البيانات وهذه المشاهدات هي (23,46,53,94)، إذ تم الكشف عنها بواسطة عناصر القطر الرئيس المصفوفة القبعة.
3. من نتائج الجانب التطبيقي نلاحظ معنوية متغير الانفاق الاجمالي لرب الاسرة على التعليم ومتغير الانفاق الاجمالي المسكن والكهرباء ومحروقات أخرى.

سادساً: التوصيات:

بناءٍ على مخرجات الجانب التجريبي والتطبيقي وما استنتجه الباحث نوصي بما يأتي:

1. ضرورة إعطاء الأولوية لطرق تقدير المعلمات لأنموذج الانحدار الخطي المتعدد بوجود المشاهدات الشاذة (نقاط الانعطاف العالية) بعدِّها أكثر أهمية من عملية الحذف لهذه المشاهدات لأن هذه المشاهدات قد تمتلك معلومات مهمة.
2. ضرورة توسيع طرائق تقدير المعلمات لأنموذج الانحدار الخطي المتعدد لتشمل نماذج الانحدار متعدد المتغيرات ونماذج الانحدار اللاخطية.
3. يمكن الاستفادة من مقدر MM الحصين في تقدير أي انموذج انحدار خطي متعدد يحتوي مشاهدات شاذة.

المصادر

أولاً: المصادر باللغة العربية:

1. الراوي، خاشع محمود، (1987)، المدخل الى تحليل الانحدار، مديرية دار الكتب للطباعة والنشر، جامعة الموصل.
2. كاظم، أموري هادي ومسلم، باسم شلبية، (2002)، القياس الاقتصادي المتقدم النظرية والتطبيق، مكتبة دنيا الأمل، بغداد.
3. ناسي، نبيل جورج والجبوري، شلال حبيب، (2001)، تقييم كفاءة طرق تقدير القيم الشاذة لنماذج الانحدار، رسالة ماجستير، قسم الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد.

ثانياً: المصادر باللغة الانكليزية:

4. Alguraibawi, M., Midi, H. & Rana, 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, 49(4), 305–322.
5. Chen C., (2002), "Robust regression and outlier detection with the Robustreg procedure" In Proceedings of the Twenty Seventh Annual SAS Users Group International Conference; SAS Institute: Cary, NC.
6. Chatterjee, S. & Hadi, A. S., (1990), Sensitivity Analysis in Linear Regression. In Journal of the Royal Statistical Society. Series A (Statistics in Society) (Vol. 153, Issue 1). Wiley. https://doi.org/10.2307/2983106

7. Hoaglin, D. C. & Welsch, R. E., (1978), "The Hat Matrix in Regression and Anova", The American Statistician, No. 32(1), PP 17-22
8. Kamruzzaman., MD. & Imon, A.H.M.R., (2002), "High leverage point: another source of multicollinearity", Pakistanian Journal of Statistics, No. 18(3), PP435-448
9. Kim, M. G., (2004), "Sources of high leverage in linear regression model", Journal of Applied Mathematics and Computing, No. 16(1-2), PP:509-513
10. Yuliana, S., Hasih, P., Sri Sulistijowati, H. & Twenty, L., (2014), M Estimation, S Estimation, and Mm Estimation in Robust Regression. International Journal of Pure and Applied Mathematics, 91(3), 349–360.