AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

تقدير معلمات الانحدار الخطي في ظل وجود مشكلة الارتباط الذاتي والقيم الرافعة العالية

محمد عبد الحسين محمد	بحر كاظم محمد
dw.moh2@atu.edu.iq	Bahr.mahemmed@qu.edu.iq
المعهد التقني ديوانية، جامعة الفرات الاوسط التقنية، القادسية، العراق	قسم الاحصاء، كلية الادارة والاقتصاد، جامعة القادسية، القادسية، العراق

زهراء حيدر حسين

zahra.hader@yahoo.com

قسم الاحصاء، كلية الادارة والاقتصاد، جامعة القادسية، القادسية، العراق

المستخلص

يعد تحليل الانحدار الخطي من أهم التقنيات الإحصائية في العديد من المجالات مثل الاقتصاد ودراسات البقاء وادارة الأعمال والطب والهندسة وغيرها, ولتقدير معاملات نموذج الانحدار الخطي، غالبًا ما يتم استخدام طريقة المربعات الصغرى لما تتمتع به من مزايا جيدة وسهولة حسابها. ومع ذلك، فإن وجود نقاط شاذة مفردة أو متعددة في مجموعة بيانات يمكن أن يؤدي إلى تدمير تقديرات OLS. ذكر العديد من الباحثين أن مجموعة البيانات الحقيقية تحتوي عادة على 1% إلى 10% من الملاحظات غير العادية. ان البيانات الشاذة في المتغيرات التوضيحية والتي تسمى بالنقاط الرافعة العالية (HLPs) والتي لها تأثيرات أكثر خطورة على تقديرات OLS من القيم الشاذة في المتغير y. ان القيم الرافعة العالية هي المسؤولة عن مشكلة الاخفاء Masking والإخفاء (Swamping) في الانحدار الخطي. تتسبب HLPs أيضًا في مشكلة التعددية الخطية ولها تأثير كبير على دقة التقديرات. وعليه، فمن الضروري اكتشاف تلك المشاهدات غير العادية. كذلك يمكن ان تعاني مقدرات طريقة المربعات الصغرى من تدهور كبير في ظل وجود مشكلة الارتباط الذاتي (Auto-correlated) . ولمعالجة هذه المشكلة تم اقتراح مجموعة من الطرق ومن بينها طريقة طريقة كوكرين- أوركوت برايس- وينستن التكرارية, ولكن للأسف معظم هذه الطرق التقليدية تفشل بشكل منفرد في معالجة مشكلة الارتباط الذاتي بين الاخطاء بوجود القيم الشاذة.

في هذا الدراسة تم اقتراح طريقة حصينة لمعالجة المشكلة المركبة للارتباط الذاتي في نموذج الانحدار الخطي المتعدد في ظل وجود القيم الشاذة والقيم الرافعة العالية. تم تقييم اداء الطريقة المقترحة باستخدام بيانات حقيقية.

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث: 22/2/2024
تاريخ قبول البحث: 12/4/2024
تاريخ رفع البحث على الموقع: 31/12/2024

الكلمات المفتاحية:
الارتباط الذاتي، الطرق الحصينة، قيم الرافعة العالية.

للمراسلة:
محمد عبد الحسين محمد
dw.moh2@atu.edu.iq

DOI: https://doi.org/10.55562/jrucs.v56i1.25

1. المقدمة

نموذج الانحدار الخطي هو احد النماذج الاحصائية الشائعة الاستخدام في مجموعة واسعة من التطبيقات العملية والتحليلات والتنبؤ بالقيم المستقبلية للظواهر المدروسة, يعطي تحليل الانحدار إجابات على الأسئلة المتعلقة بالعلاقة الدالية لمتغير الاستجابة (Response variable) مع واحد أو أكثر من المتغيرات التوضيحية (Explanatory variables)، بما في ذلك تقدير تأثير تغير قيم المتغير التوضيحي على قيم متغير الاستجابة، والتنبؤ بالقيم المستقبلية لمتغير الاستجابة (Montgomery et al., 2001; Gross, 2003; Kutner et al., 2005). لصياغة العلاقة الدالية تتم إضافة مصطلح الخطأ العشوائي ε_i إلى النموذج الإحصائي لمراعاة الفروق الفردية، كما في الصيغة التالية:

$$y_i = f\left(x_{i1}, x_{i2}, \dots, x_{ki}\right) + \varepsilon_i, \quad i = 1,2, \dots, n, \qquad (1)$$

حيث يكون ε_i هو الخطأ العشوائي الفردي للمشاهدة i وان n تمثل عدد المشاهدات في مجموعة البيانات مع عدد المتغيرات التنبؤية k. تسمى الدالة الخطية بين y وX نموذج الانحدار الخطي المتعدد (MLR)، ويمكن تعريفها بالصيغة التالية :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{pi} + \varepsilon_i \qquad (2)$$

كذلك يمكن التعبير عن نموذج الانحدار الخطي المتعدد بشكل مصفوفات حسب الصيغة التالية:

$$y = X\beta + \varepsilon \qquad (3)$$

حيث ان y متجهًا $(n \times 1)$ لمتغير الاستجابة، وX تكون مصفوفة $[(n \times p) + 1]$ لمتغيرات الانحدار (بما في ذلك متجه الحد الثابت)، وتكون β متجهًا $[(p+1) \times 1)]$ لمعلمات الانحدار (Coefficients) غير المعروفة إلى يتم تقديرها ويكون ε_i متجهًا عشوائيًا $(n \times 1)$ يُفترض أنه موزع بشكل مستقل (identical independent distribution) طبيعي بمتوسط صفر وتباين ثابت.

ويعتبر افتراض الأخطاء العشوائية غير المترابطة أمرا ضروريا في تطبيق طريقة المربعات الصغرى الاعتيادية لتقدير معلمات النموذج. ولكن قد لا يكون هذا الافتراض صحيحا في الكثير من الحالات عندما تكون اخطاء النموذج مرتبطة بالأخطاء السابقة. تسبب الاخطاء العشوائية المرتبطة تلقائيا مشاكل خطيرة في النموذج الخطي حيث انه ينتهك الخصائص المهمة لطريقه المربعات الصغرى (White and Brisbin,1980) حيث لم تعد تقديرات المعلمات التي يتم الحصول عليها هي افضل المقدرات الخطية غير المتحيزة (BLUE), وكذلك لم يعد اختبار t واختبار F ذات موثوقية عالية وذلك لان هذه الاختبارات تميل الى ان تكون ذات دلالة إحصائية في حين أنها ليست كذلك بالواقع. وبوجود مشكلة الارتباط الذاتي يصبح معامل التحديد (Coefficient of Determinant) متضخما بشكل كبير. وعلى هذا النحو, ستبدو المقدرات أكثر دقة بشكل مضلل مقارنة بقيمتها الفعلية. كل هذه المشاكل تساهم في فشل اختبار الفرضيات وبالتالي فأن الاخطاء المرتبطة تلقائيا ستعطي على الارجح استنتاجات مضللة حول الاهمية الاحصائية لمعاملات الانحدار المقدرة (Gujarati, D. N. and Porter, D. C.2009) . لذلك من المهم جدا الكشف عن وجود الأخطاء المرتبطة تلقائيا قبل البدء في عمليه التحليل الإحصائي.

غالبا ما تظهر الاخطاء المرتبطة تلقائيا في بيانات السلاسل الزمنية. حيث يتم جمع الملاحظات في بيانات السلاسل الزمنية وفقا لترتيب طبيعي مع مرور الوقت, ومن المرجح ان تظهر الملاحظات المتعاقبة ارتباطات متبادلة. ومن ناحية اخرى, يمكن ان تظهر الاخطاء المرتبطة تلقائيا في البيانات المقطعية, يتم جمع الملاحظات على اساس عشوائي في البيانات المقطعية بغض النظر عن التغيرات الزمنية. ومع ذلك، نلاحظ في معظم الدراسات دائما ما يتم انتهاك افتراض الأخطاء العشوائية غير المرتبطة.

يفترض النموذج الكلاسيكي أن حد الخطأ العشوائي المتعلق بأي مشاهدة لا يتأثر بالخطأ المتعلق بأي مشاهدة أخرى. إلا أنه يمكن ان نلاحظ ان الأخطاء قد تكون مرتبطة بالأخطاء السابقة, اي ان

$$E\left(e_i, e_j\right) \neq 0 \quad for\ \ i \neq j$$

على الرغم من أن الأخطاء المرتبطة تلقائيًا لا تسبب أي تحيز في مقدرات OLS ، ولكن تقديرات معاملات OLS تصبح أقل كفاءة في وجود أخطاء مرتبطة ذاتيا وكما اشرنا سابقا فان هذا يؤدي إلى استنتاج مضلل حول الإحصاءات.

من ناحية أخرى، من الواضح أن مقدرات طريقة OLS والتي تعتمد على منهجيه تقليل مجموع المربعات بين القيم الحقيقية والاستجابات المرصودة في مجموعة البيانات تتأثر بالقيم الشاذة (Outliers) وخاصة بنقاط الرافعة العالية (High Leverage Points). حيث أكدت الأبحاث التي أجراها هارتر (1974) أن تربيع البواقي يجعل طريقة المربعات الصغرى ضعيفة للغاية عند وجود نقاط رافعة عالية. وهذا يحدث خرق في افتراضات المربعات الصغرى. وفي الوقت نفسه، لا يمكن ضمان مجموعة البيانات الروتينية خالية من القيم الشاذة اونقاط الرفع العالية. لذلك من الضروري تقديم أساليب حصينه في الانحدار الخطي لمعالجة مشاكل الارتباط الذاتي ونقاط الرافعة العالية.

هناك عدة إجراءات علاجية مثل طريقة كوكرين- أوركوت التكرارية وإجراءات كوكرين-أوركوت برايس-وينستن ذات الخطوتين أو التكرارية، وإجراءات دوربين المكونة من خطوتين. يعتمد علاج المشكلة على معرفة بنية الارتباط الذاتي. فعلى سبيل المثال اذا كان الارتباط الذاتي من الرتبة الاولى AR(1)، اي ان:

$$u_t = \rho\, u_{t-1} + v_t \qquad (4)$$

هناك حالتان يجب النظر فيهما: الاولى عندما يكون معامل الارتباط ρ معروفًا و والثانية متى ما يكون ρ غير معروف وفي هذه الحالة يجب تقديره قبل استخدام طريقة المعالجة

1- عندما تكون معلمة الارتباط الذاتي ρ معلومة

إذا كان معامل الارتباط الذاتي معروفا، يمكن معالجة مشكلة الارتباط الذاتي بسهولة (Gujarati and Porter, 2009).فأذا كانت المعادلة (2) صحيحة عند الزمن t، فيجب أن تكون صحيحة أيضًا في الزمن (t-1). لذلك، فان :

$$y_{i-1} = \beta_0 + \beta_1 x_{1i-1} + \beta_2 x_{2i-1} + \cdots + \beta_k x_{ki-1} + \varepsilon_{i-1} \qquad (5)$$

بضرب المعادلة (5) بمعامل الارتباط الذاتي $\boldsymbol{\rho}$ سوف ينتج ما يلي :

$$\rho\, y_{i-1} = \rho\, \beta_0 + \rho\, \beta_1 x_{1i-1} + \rho\, \beta_2 x_{2i-1} + \cdots + \rho\, \beta_k x_{ki-1} + \rho\, \varepsilon_{i-1} \qquad (6)$$

بطرح المعادلة (6) من المعادلة (2) نحصل على ما يلي :

$$y_i - \rho\, y_{i-1} = \beta_0(1-\rho) + \beta_1(x_{1i-}\, \rho x_{1i-1)} + \cdots + \beta_k(x_{ki-}\, \rho x_{ki-1)} + \varepsilon_i - \rho\, \varepsilon_{i-1} \quad (7)$$

تسمى معادلة (7) بالنموذج المعمم او معادلة الفروق والتي هي انحدار قيم y على قيم X لقيم الفروق بين القيم الاصلية والقيم السابقة مضروبا في معامل الانحدار الذاتي. ويمكن التعبير عن المعادلة (7) بالصيغة الاتيه:-

$$y_i^* = \beta_0^* + \beta_1^* x_{1i}^* + \beta_2^* x_{2i}^* + \cdots + \beta_k^* x_{ki}^* + \varepsilon_i^* \quad (8)$$

حيث ان:

$$y_i^* = y_i - \rho\, y_{i-1}\ ;\ \beta_0^* = \beta_0(1-\rho);\ \beta_j^* = \beta_j, j = 1,2, \dots, k$$

$$x_{ji}^* = (x_{ji-}\, \rho x_{ji-1)};\ \ \varepsilon_i^* = \varepsilon_i - \rho\, \varepsilon_{i-1}$$

وبما أن حد الخطأ العشوائي في المعادلة (7) يحقق افتراضات طريقة OLS الاعتيادية، فيمكن تطبيق طريقة المربعات الصغرى للحصول على تقديرات المعلمات مع جميع الخصائص والتي هي افضل تقدير خطي غير متحيز BLUE.

2-عندما تكون معلمة الارتباط الذاتي ρ غير معلومة

عندما تكون معلمة الارتباط الذاتي ρ غير معلومة فمن الصعوبة تنفيذ طريقة الفرق المعمم الواردة في المعادلة (7) لأنها تعتمد بشكل كبير على معلمة الارتباط ρ. ولذلك تقديرها قبل أن يتم تنفيذ طريقة المعالجة (Gujarati and Porter, 2009). وهناك عدة طرق يمكن اتباعها في مثل هذه الحالة منها ما يلي:

- تقدير معلمة الارتباط الذاتي ρ بالاعتماد إحصاءه ديربن واتسن Durbin Watson statistics.

فاذا كانت قيمة ρ ليست كبيرة جدا فيمكن تقديرها بسهولة من خلال العلاقة بين قيمة المعلمة ρ وإحصاءه ديربن واتسن d وكما في ادناه (Maddala, 2001)

$$\hat{\rho} \approx 1 - \frac{d}{2} \quad (9)$$

ويمكن استخدام قيمة معامل الارتباط التقديرية $\hat{\rho}$ لتطبيق النموذج المعمم او معادلة الفروق في معادلة (7) ومن ثم تطبيق طريقة المربعات الصغرى الاعتيادية بالاعتماد على الصيغة (8).

2. طريقة كوكرين- أوركوت برايس- وينستن التكرارية Orcutt Cochrane-Winsten Prais Iterative

وهي واحدة من الطرق الشائعة الاستخدام في ايجاد قيمة ρ , حيث ان في هذه الطريقة يتم استخدام الطريقة التكرارية في ايجاد قيمة معامل الارتباط الذاتي التقديرية بالاعتماد على قيم ابتدائية لـ ρ . فعلى فرض ان معادلة الانحدار كما المعادلة (2) وعلى فرض ان نموذج الارتباط الذاتي من الدرجة الاولى حسب الصيغة (4), عندئذ يمكن ايجاد قيمة ρ التقديرية باتباع الخطوات التالية:

1- من معادلة 2 يتم تقدير معالم نموذج الانحدار الخطي وكذلك حساب البواقي $\hat{\varepsilon}_i$.

2- من نتائج الخطوة 1 يتم تطبيق طريقة المربعات الصغرى لإيجاد معالم انحدار البواقي $\hat{\varepsilon}_i$ على االبواقي بتخلف فترة زمنيه $\hat{\varepsilon}_{i-1}$ وكما في الصيغة التالية:

$$\hat{u}_t = \hat{\rho}\, \hat{u}_{t-1} + \nu_t \quad (10)$$

3- من خطوه 2, يتم استخدام قيمة $\hat{\boldsymbol{\rho}}$ لتطبيق النموذج المعمم او معادلة الفروق في معادلة (7) وتقدير معالم النموذج $\hat{\boldsymbol{\beta}}_0^*, \hat{\boldsymbol{\beta}}_1^*$ ومن ثم حساب القيم التنبؤية المقدرة (***Estimated fitted value***) وحسب الصيغة التالية:

$$\hat{y}_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* x_{1i}^* + \hat{\beta}_2^* x_{2i}^* + \cdots + \hat{\beta}_k^* x_{ki}^* \quad (11)$$

4- حساب البواقي الجديدة من العلاقة التالية:

$$\hat{u}_i^* = y_i - \hat{y}_i \quad (12)$$

5- يتم اعادة تطبيق معادلة (10) بالاعتماد على قيم $\hat{\boldsymbol{u}}_i^*$ التي تم الحصول عليها من خطوة 4 للحصول على قيمة $\hat{\boldsymbol{\rho}}$

ويتم اعادة احتساب قيم $\hat{\boldsymbol{\rho}}$ بشكل تكراري حتى يكون الفرق بين اخر دورتين اقل من 0.001 او بعدد دورات 20 كحد اقصى

3. طرق التقدير الحصينة

في هذه الفقرة سوف يتم عرض مجموعة مهمة من طرق التقدير الحصينة والتي تتمتع بمزايا جيدة من نقاط الانهيار العالية والتي تقترب من 0.50 والكفاءة النسبية والتي تقترب من 0.95

3.1. مقدر M : M-Estimator

يعتبر مقدر M التقنية الأكثر شعبية في نموذج الانحدار الخطي ، التي اقترحها هوبر (1964، 1973 و2004) والتي يمكن اعتبارها بمثابة تعميم لمقدر الاحتمال الأقصى. يقلل مقدر M من تأثير البيانات غير العادية عن طريق تغيير مجموع مربعات الخطأ في OLS بواسطة دالة أخرى. يتم تحديد مقدر M عن طريق تقليل مجموع دالة البواقي التي تتزايد بسرعة أقل، على النحو التالي :

$$\min_{\beta}\sum_{i=1}^{n}\gamma\left(r_i\right)=\min_{\beta}\sum_{i=1}^{n}\gamma\left(y_i-\sum_{i=1}^{n}x_{ij}\widehat{\beta}_j\right) \tag{15}$$

حيث γ دالة معينة تحدد مساهمات البواقي في دالة الهدف. يجب أن تتمتع دالة γ بما يلي:

- $\gamma(r)\geq 0,$ دالة موجبة التعريف
- $\gamma(r)=\gamma(-r)$ دالة متماثلة
- $\rho(0)=0$ لها حد ادنى وحيد عند الصفر
- $\left|r_i\right|>\left|r_i^t\right|$ for $\gamma\left(r_i\right)\geq\gamma\left(r_i^t\right)$. دالة متناغمة عند $\left|r_i\right|$

نظرًا لأن حل مقدر M ليس مكافئًا للقياس (not scale equivariant) ، يتم تعديل مشكلة التقليل عن طريق قسمة الدالة γ على تقدير قوي للقياس $\widehat{\sigma}$ كما في :

$$\min_{\beta}\sum_{i=1}^{n}\rho\left(\frac{r_i}{\widehat{\sigma}}\right)=\min_{\beta}\sum_{i=1}^{n}\rho\left(\frac{y_i-x_{ij}^t\widehat{\beta}_j}{\widehat{\sigma}}\right) \tag{16}$$

المقياس الحصين MAD هو التقدير الأكثر شيوعًا لـ $\widehat{\sigma}$. (Huber, 1964; Hampel, 1974)

3.2 طريقة M المعممة GM- Estimators

هامبل وآخرون. (1986) اثبتو أن مقدر M لا يملك دالة تأثير محدودة Bounded influential function ، خاصة اذا كانت البيانات تحتوي على قيم رافعة عالية. للحد من هذا القصور، تم اقتراح التأثير المحدود لمقدرات M المعممة (GM-estimator) من قبل Schweppe (انظر Hill، 1977؛ Andersen، 2008). الغرض الرئيسي من مقدر GM هو إنتاج أوزان تأخذ في الاعتبار كلا الاتجاهين y وX. الشكل العام لمقدرات GM بالشكل التالي:

$$\sum_{i=1}^{n}\gamma_i\psi\left(\frac{y_i-x_i^t\widehat{\beta}}{\widehat{\sigma}\pi_i}\right)x_i \tag{17}$$

حيث ان γ هي دالة وزن أولية تتحكم في الأوزان المعطاة للروافع العالية. يمكن حل المعادلة في (17) باستخدام تقنية المربعات الصغرى الموزونة التكرارية IRLS ومن ثم، يمكن كتابة مقدر GM بشكل متقارب كما في الصيغة التالية:-

$$\widehat{\beta}_{GM}=(X^tWX)^{-1}X^tWy \tag{18}$$

حيث ان W هي مصفوفة قطرية بالأوزان ω_i والتي تاخذ الصيغة التالية:

$$\omega_i=\frac{\psi\left[\left(y_i-x_i'\widehat{\beta}_{GM}\right)/\pi_i\widehat{\sigma}\right]}{\left(y_i-x_i'\widehat{\beta}_{GM}\right)/\pi_i\widehat{\sigma}} \tag{19}$$

في الادبيات الاحصائية, هناك مجموعة من الدوال التي تم اقتراحها لاستخدامها كدالة وزن اولية initial weight مثل دالة (Mallows (1975 بالصيغة $\pi_i=\sqrt{1-h_{ii}}$ ودالة (Krasker and Welsch (1982 بالصيغة $\pi_i=\sqrt{(1-h_{ii})/h_{ii}}$.

حيث ان h_{ii} هي العناصر القطرية لمصفوفة الاسقاطات او ما تسمى Hat matrix

3.3 مقدرات-MM المكيفة : MMorestimat-

أحد أكثر الأساليب شعبية وشائعة الاستخدام في مجال الانحدار الحصين هو مقدر MM الذي قدمه يوهاي (1987) فهو يجمع بين نقطة انهيار مرتفعة BP= 0.5 والكفاءة الفائقة (95% من كفاءة OLS في ضل تحقق الافتراضات الأساسية). تم استخدام تقنية المربعات الصغرى الموزونة التكرارية IRLS للحصول على مقدر MM. يشير "MM" في الإشارة إلى أنه يتم استخدام أكثر من عملية تقدير M لإيجاد التقديرات النهائية. يتم تلخيص إجراء تقديرات MM على النحو التالي (انظر Rousseeuw and Leroy, 2003):

1- حساب التقديرات الأولية للمعاملات $\widehat{\beta}^{(1)}$ والبواقي المقابلة $r_i^{(1)},\ i=1,2,\dots,n$ من خلال استخدام مقدرله نقطة انهيارمرتفعه مثل مقدرات S مع دالة وزن مناسبة مثل Huber أو bisquare.

2- نقوم بحساب تقدير M لتباين البواقي $\widehat{\sigma_r}$ باستخدام النتائج من الخطوة 1.

3- يتم استخدام البواقي (من الخطوة 1) والتباين (من الخطوة 2) في التكرار الأول للمربعات الصغرى الموزونه WLS لايجاد تقديرات M لمعلمات الانحدار

$$\sum_{i=1}^{n} \omega_i \left(r_i^{(1)} / \widehat{\sigma_r} \right) x_i = 0 \qquad (20)$$

حيث ان ω_i دالة وزن *Huber* أو *bisquare*.

4- حساب الأوزان الجديدة $\omega_i^{(2)}$ باستخدام البواقي من الخطوة 3.

5- يتم الحفاظ على $\widehat{\sigma_r}$ ثابتة من الخطوة 2، ويتم تكرار الخطوتين 3 و4 حتى التقارب.

4. طريقة GM- estimator المعدلة - Modified GM estimator

على الرغم من أن مقدرات GM لها دالة تأثير محدودة BIF، إلا أن لديها نقطة انهيار معتدلة تبلغ على الأكثر $\frac{1}{k}$، حيث k هو عدد معاملات الانحدار بما في ذلك الحد الثابت (Hampel et al., 1986; Andersen, 2008). قدم ويلكوكس (2005) مقدر GM6 حيث تعتمد وظيفة الوزن π في معادلة 19 على مصفوفة التباين والتباين المشترك الحصينة MVE . تكمن نقطة ضعف MVE في أنها تميل إلى إغراق Swamping بعض نقاط الرافعة المنخفضة (Bagheri and Habshah, 2015). وبعد ذلك، سيتم إعطاء بعض الروافع الجيدة أوزانًا منخفضة. ومن ثم فإن تقدير GM6 سيكون أقل كفاءة. وهذا يحفزنا على تحسين دقة مقدر GM6 من خلال استخدام دالة MRCD بدلا من مصفوفة MVE لصياغة مقدر GM الذي نسميه مقدر MGM. تم تلخيص MGM على النحو التالي:

1- حساب البواقي لدالة الانحدار باستخدام مقدر S ومن ثم الانحراف المعياري للبواقي، من خلال تطبيق المعادلة التالية:

$$\hat{\tau} = c(1+5/(n-p))\,\text{Median}|r_i| \qquad (20)$$

حيث ان c هي ثابت وقيمته تساوي تقريبا c = 1.4826

2- حساب دالة الوزن الاولية $\omega_\circ$ من الصيغة التالية:

$$\omega_i = \min\left[1, \left\{\frac{\chi^2_{0.95,K}}{RMD^2_{MRCD}}\right\}\right] \qquad (21)$$

حيث ان:

RMD^2_{MRCD} هي مسافة مهنلوبس Mahalanobis Distance المحسوبة باستخدام مصفوفة التباين والتباين المشترك RMD

وان $\chi^2_{0.95,K}$ هي قيمة مربع كاي بدرجة حرية 0.95 وk حيث k تمثل عدد المتغيرات المستقلة.

3- حساب مشتقة دالة GM في معادلة 17 والتي يرمز لها Ψ

$$\Psi = diag\left[\acute{\psi}\left(\frac{r_i}{\hat{\tau} \times \pi_i}\right)\right] \qquad (22)$$

حيث ان هي المشتقة الاولى لدالة هيوبر ψ

4- تكون الصيغة النهائية لطريقة GM المعدلة والتي نرمز لها MGM (Modify GM-estimator) والتي تعتمد على خوارزمية نيوتن رافسون بالشكل التالي:

$$\hat{\beta}_{MGM} = \hat{\beta}_0 + (X^t \Psi X)^{-1} X^T W \psi(\frac{r_i}{w_i \hat{\tau}})\hat{\tau} \qquad (23)$$

حيث ان W هي مصفوفو مربعة بالأوزان المحسوبة من خطوة 2 وان β_0 هي القيم الاولية للمعلمات والتي يتم تقديرها بالاعتماد على طريقة ذات نقطة انهيار عالية مثل طريقة LMS او LTS او S.

5. طريقة كوكرين- أوركوت برايس- وينستن الحصينة RCOPW

طرق التقدير التي تمت مناقشتها أعلاه تعطي تقديرات حصينة في ظل وجود القيم الشاذة ولا تعالج المشكلة المركبة في ظل مشكلة الارتباط الذاتي بوجود القيم الشاذة. وتعد طريقة كوكرين-أوركوت برايس-وينستن التكرارية - Orcutt-Winsten Prais Cochrane والتي سوف نرمز لها COPW واحدة من أكثر الطرق التكرارية شيوعًا المستخدمة في هذا المجال. ولكن للاسف فان هذه الطريقة ليست حصينة بوجود القيم الشاذة وبالتالي يمكن التأثير عليها بسهولة من خلال نقاط الرافعة العالية (انظر حبشه (1999)، رنا وآخرون (2012) وساني وآخرون (2019). في هذا البحث، نقترح طريقة كوكرين- أوركوت برايس- وينستن التكرارية الحصينة بالاعتماد على مقدر MGM الذي تم اقتراحه في الفقرة السابقة بحيث يمكن تصحيح الأخطاء المرتبطة ذاتيا ونقاط الرافعة المالية العالية في وقت واحد. تم تطوير هذه الطريقة من خلال دمج الكفاءة التقاربية العالية والانهيار العالي وخاصية

التأثير المحدود لمقدر MGM في طريقة كوكرين-أوركوت برايس-وينستن التكرارية. نطلق على هذه الطريقة الجديدة اسم طريقة Robust Orcutt Cochrane-Winsten Prais التكرارية, واختصارا نرمز لها RCOPW

من المهم الإشارة إلى أنه في كل خطوة من خطوات الطريقة التكرارية COPW، يتم استخدام طريقة المربعات الصغرى العادية للحصول على تقديرات المعلمات لنموذج الانحدار الخطي المتعدد. بالإضافة إلى ذلك، فإن التحويل لتصحيح مشكلة الارتباط التلقائي لا يزيل تأثير HLPs على طريقة OLS. وبالتالي، من المتوقع أن تظهر COPW نتائج غير صحيحة في ظل وجود HLPs. لذلك تقترح تعديل COPW بحيث اكثر مقاومتا ولا يتأثر بـ HLPs. تمت صياغة RCOPW الحصينة من خلال دمج مقدر MGM الذي لديه نقطة انهيار قريبة من 0.5، ودالة تأثير محدودة وكفاءة مقاربة عالية للنموذج العادي (انظر Coakley and Hettmansperger (1993)) في RCOPW يتم استبدال مقدر OLS بمقدر MGM. من المتوقع أن يكون RCOPW أكثر اتساقًا ودقة مقارنةً بـ COPW الاعتيادية .

على فرض ان نموذج الانحدار الخطي المتعدد كما في معادلة 2 والذي يتبع نموذج الانحدار الخطي من الدرجة الاولى كما في معادلة 4، عندئذ يتم تلخيص خطوات الطريقة المقترحة RCOPW بالشكل التالي:

1- تقدير معاملات الانحدار الخطي المتعدد في المعادلة (2) باستخدام مقدر MGM والحصول على البواقي ε_i.

2- باستخدام البواقي ε_i الذي تم الحصول عليه في الخطوة 1، يتم تقدير معلمات نموذج الارتباط الذاتي في معادلة 4 والحصول على تقدير معامل الارتباط الذاتي ρ.

3- باستخدام تقدير ρ من الخطوة 2، نقوم بتقدير معادلة الانحدار في معادلة (8) بالاعتماد على مقدر MGM:

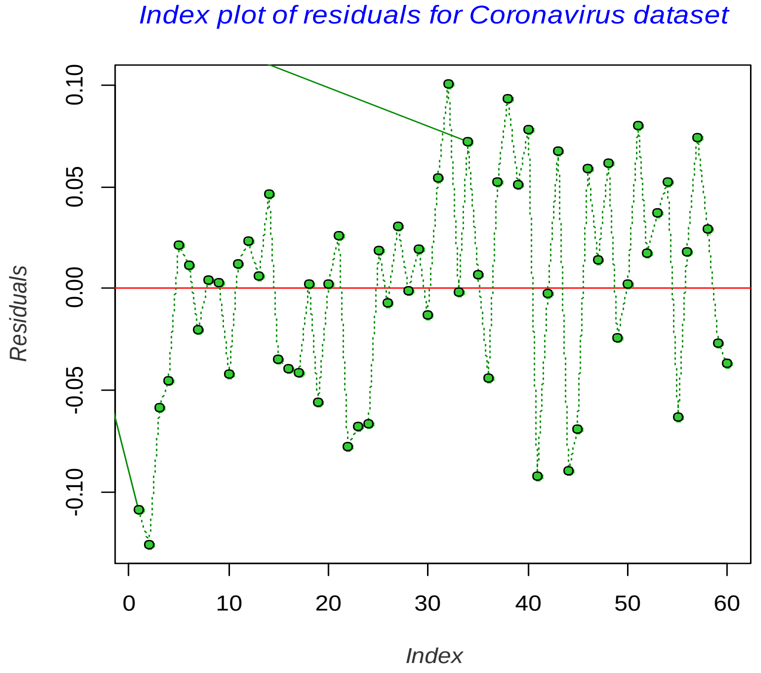
$$y_i^* = \beta_0^* + \beta_1^* x_{1i}^* + \beta_2^* x_{2i}^* + \cdots + \beta_k^* x_{ki}^* + \varepsilon_i^*$$

4- من خلال الخطوة 3 نستخدم المعلمات المقدرة $\beta_0^*,\ \beta_1^*\ ,\ \beta_2^*, \dots, \beta_k^*$ لحساب البواقي الجديدة.

5- نكرر الخطوة 2 إلى الخطوة 4 حتى تصل التقديرات المتعاقبة لـ ρ بأقل من 0.0001 أو بحد أقصى 20 تكرارًا إذا لم يتم استيفاء معايير التقارب.

6. بيانات كورونا

بيانات كورونا جمعت من مستشفى الديوانية العام للفترة من 2019-2021 وتضمنت البيانات متغير معتمد يتمثل بنسبة المرض للمصابين ومتغيرين مستقلين هما نسبة الاوكسجين في الدم والخثرة. عدد الحالات التي تم دراستها هو 60 حالة والذي يمثل حجم العينة المدروسة. الشكل 1 يمثل شكل المشاهدات للبواقي حسب طريقة المربعات الصغرى, حيث يمكن ملاحظة ان هناك نمط واضح في البيانات والذي يدل على وجود مشكلة الارتباط الذاتي



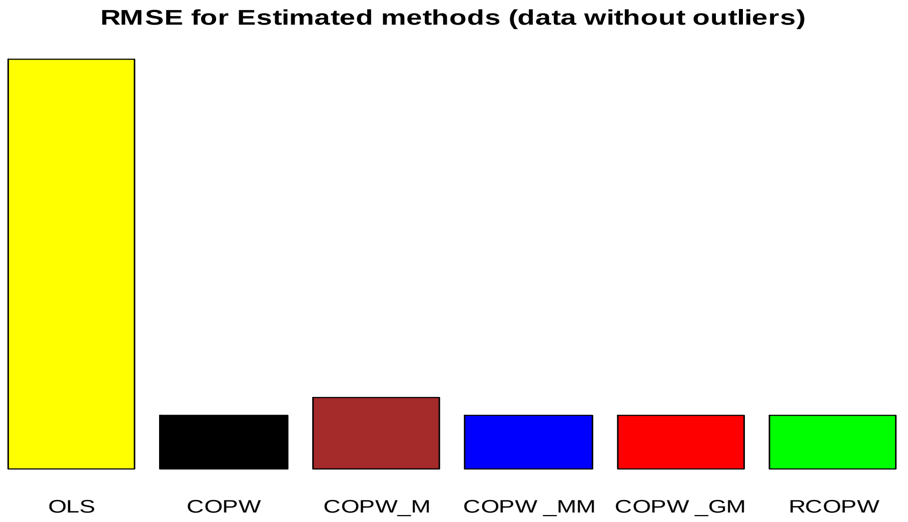
شكل 1: يمثل شكل الانتشار للبواقي

والجدول 1 يمثل متوسط مجموع مربعات الاخطاء (RMSE) لمجموعة من طرق التقدير في حالة البيانات الاصلية والتي تحتوي على مشكلة الارتباط الذاتي فقط وفي الحالة الثانية بوجود مشكلة الارتباط الذاتي والقيم الرافعة العالية

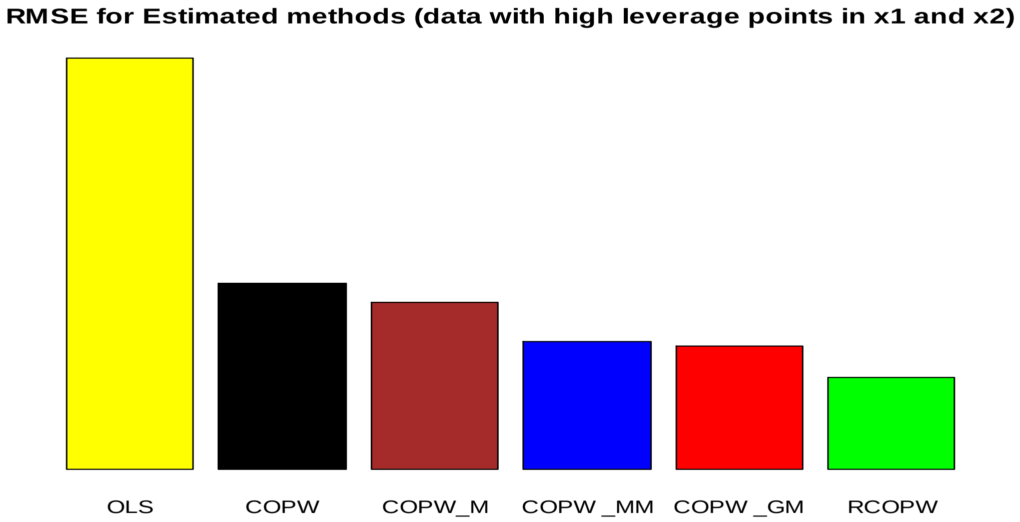
جدول 1: يوضح قيم RMSE لطرق التقدير لبيانات كورونا للفترة 2019 - 2021

Methods	Data with auto-correlated errors	Data with auto-correlated errors and high leverage points in x_1 and x_2
	RMSE	RMSE
OLS	0.3334	2.3334
COPW	0.04311	1.0541
COPW_M	0.0578	0.9447
COPW_MM	0.04367	0.7225
COPW_GM	0.04351	0.7017
RCOPW	0.04342	0.5189

من الجدول اعلاه نلاحظ ان طرق التقدير جميعها اعطت نتائج متقاربة لمتوسط مربعات الخطأ عدا طريقة المربعات الصغرى والتي كانت ذات اداء ضعيف جدا مقارنة بالطرق الاخرى وكما موضح في الشكل 2.



شكل 2: الاشرطة البيانية لقيم RMSE لطرق التقدير لبيانات كورونا الاصلية



شكل 3: الاشرطة البيانية لقيم RMSE لطرق التقدير لبيانات كورونا المعدلة بوجود قيم الرافعة العالية

اما في الحالة الثانية والتي فيها تم اضافة قيم رافعة عالية للبيانات في المتغيرين x1 و x2 فنجد ان اداء الطرق الكلاسيكية ينهار بشكل كبير وهذا واضح من خلال ارتفاع قيم RMSE، وبالمقابل نجد ان الطرق الحصينة كان ادائها جيدا في معالجة المشكلة

المركبة من الارتباط الذاتي وقيم الرافعة العالية وكما موضح في الشكل 3 حيث نجد ان الطريقة المقترحة RCOPW كان لها افضل اداء مقارنة بالطرق التقليدية والطرق الحصينة.

7. الاستنتاجات

في هذه الدراسة تم التطرق الى معالجة المشكلة المركبة لنموذج الانحدار الخطي والذي يعاني من مشكلة الارتباط الذاتي بوجود القيم الرافعة العالية حيث ان معظم الطرق التقليدية تفشل في تقدير معلمات النموذج ولهذا تم اقتراح طريقة لتقدير المعلمات والتي تتمتع بمميزات جيدة مقارنة بالطرق التقليدية. .وفي ادناه اهم الاستنتاجات التي تم الحصول عليها من تطبيق الطريقة المقترحة على البيانات الحقيقية:

1. ان الطرق التقليدية في معالجة مشكلة الارتباط الذاتي تنهار بشكل كبير بوجود القيم الرافعة العالية.
2. إن الطريقة المقترحة RCOPW وطريقة MM-COPW هما أفضل طرق التقدير مقارنة بالطرق الأخرى بالاعتماد على معيار المقارنة RMSE والتي امتلكت اقل القيم لمعيار المقارنة. ومع ذلك، فإن RCOPW تتفوق بشكل ملحوظ على طريقة MM-COPW
3. ان طريقة M-COPW تنهار بوجود القيم الرافعة العالية.

المصادر

[1] Alguraibawi, M., Midi, H., and Imon, A. H. M. R. (2015). A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. Mathematical Problems in Engineering, Volume 2015. Article ID 279472, 12 pages, https://doi.org/279472/2015/10.1155.

[2] Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. Australian Economic Papers, 17(31):334 355.

[3] Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. Econometrica, 48(6):1293 1301.

[4] Green, W. H. (2008). Econometric Analysis. New Jersey: Pearson Education Inc.

[5] Gujarati, D. N. and Porter, D. C. (2009). Basic Econometrics. New York: McGraw-Hill Education.

[6] Habshah, M. (1999). Preliminary estimators for robust non-linear regression estimation. Journal of Applied Statistics, 26(5):591 600.

[7] Habshah, M., Lim, H. A., and Rana, S. (2013). On the robust parameter estimation for linear model with autocorrelated errors. Advanced Science Letters, 19(8):2494 2496.

[8] Hekimoglu, S. and Erenoglu, R. (2013). A new GM-estimate with high breakdown point. Acta Geodaetica et Geophysica, 48:419 437. https://doi.org/10.1007/s1-0029-013-40328.

[9] Lim, H. A. and Midi, H. (2014). The performance of robust modification of breusch-godfrey test in the presence of outliers. Science and Technology, 22(1):81 93.

[10] On the Robust Parameter Estimation Method for Linear Model with Autocorrelated Errors

[11] Rana, S., Midi, H., and Imon, A. (2012). Robust wild bootstrap for stabilizing the variance of parameter estimates in heteroscedastic regression models in the presence of outliers. Mathematical Problems in Engineering, 12(7):1 14. https://doi.org/730328/2012/10.1155.

[12] Rousseeuw, P. J. and Leroy, A. M. (2003). Robust regression and outlier detection. New York: John Wiley and Sons.

[13] Sani, M., Midi, H., and Arasan, J. (2019). The performance of robust heteroscedasticity consistent covariance matrix estimator. Malaysian Journal of Mathematical Sciences, 13(April):71 88.

[14] White, G. and Brisbin, I. (1980). Estimation and comparison of parameters in stochastic growth models for barn owls. Growth, 44(2):97 111.

[15] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. .The Annals of Statistics, 15(20):642 656

Journal of AL-Rafidain University College for Sciences (2024); Issue 56; pp. 274-282

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

Estimating Linear Regression Parameters in the Presence of the Autocorrelation Problem and High Leverage Values

Mohammed A. Mohammed	Baher K. Mohammed
dw.moh2@atu.edu.iq	Bahr.mahemmed@qu.edu.iq
Diwaniyah Technical Institute, Al-Furat Al-Awsat Technical University, Al-Qadisiyah, Iraq	Department of Statistics, College of Administration and Economics, Al-Qadisiyah University, Al-Qadisiyah, Iraq

Zahraa H. Hussain

zahra.hader@yahoo.com

Department of Statistics, College of Administration and Economics, Al-Qadisiyah University, Al-Qadisiyah, Iraq

Article Information

Article History:
Received: February, 22, 2024
Accepted: April, 12, 2024
Available Online: December, 31, 2024

Keywords:
Autocorrelation, robust methods, high leverage values.

Correspondence:
Mohammed A. Mohammed
dw.moh2@atu.edu.iq

Abstract

Linear regression analysis is one of the most important statistical techniques in many fields such as economics, survival studies, business administration, medicine, engineering, etc. To estimate the coefficients of a linear regression model, the least squares method is often used because of its good advantages and ease of calculation. However, the presence of single or multiple outliers in a data set can destroy OLS estimates. Many researchers have mentioned that a real data set typically contains 1% to 10% of anomalous observations. Outliers in explanatory variables, called high leverage points (HLPs), have more serious effects on OLS estimates than outliers in the variable y. High leverage values are responsible for the problem of masking and swapping in linear regression. HLPs also cause the problem of multi-co-linearity and have a significant impact on the accuracy of the estimates. Therefore, it is necessary to detect these unusual observations. Also, the estimates of the least squares method can suffer from a significant deterioration in the presence of the auto-correlation problem. To address this problem, a group of methods have been proposed, including the Cochrane-Orcutt-Price-Winston iterative method, but unfortunately most of these traditional methods fail individually to address the problem of autocorrelation between errors in the presence of outliers. In this study, a robust method was proposed to address the complex problem of autocorrelation in a multiple linear regression model in the presence of outliers and high leverage values. The performance of the proposed method was evaluated using real data.

DOI: https://doi.org/10.55562/jrucs.v56i1.25