ISSN (1681- 6870)

المقارنة بين الطريقة الاعتيادية (OLS) والطريقة الحصينة الموزونة ذات المرحلتين (TSRWLS) في تقدير معلمات انموذج الانحدار الخطي المتعدد بوجود مشكلة عدم تجانس تباين الخطأ وظهور القيم الشاذة في متغير الاستجابة[1]

أ.م. غفران اسماعيل كمال
ghufranka62@gmail.com

شيماء ابراهيم خليل
shaimaa595@yahoo.com

كلية الادارة والاقتصاد - جامعة بغداد، بغداد، بغداد - العراق

المستخلص

الهدف الرئيسي من هذا البحث هو استعمال بعض الطرائق لتقدير معالم انموذج الانحدار الخطي المتعدد، الطريقة الاولى هي الطريقة الكلاسيكية طريقة المربعات الصغرى الاعتيادية (OLS) والطريقة الثانية هي طريقة المربعات الصغرى الحصينة ذات المرحلتين (TSRWLS)، لبيان اثر كل منهما في تقدير المعالم في ظل وجود مشكلة عدم تجانس تباين الخطأ وظهور القيم الشاذة في البيانات التي تعاني من هذه المشكلتين معاً. وتم ذلك باستعمال محاكاة مونتي كارلو ومن خلال معيار المقارنة متوسط الخطأ النسبي المطلق (MAPE)، وتطبيقها على بيانات حقيقية في مجال عسرة المياه مأخوذة من أمانة بغداد - دائرة ماء بغداد - قسم السيطرة النوعية للعام (2019 م)، وقد تم التوصل الى أن طريقة المربعات الصغرى الحصينة ذات المرحلتين (TSRWLS) هي الافضل لمعالجة مشكلة عدم تجانس تباين الخطأ بدون التأثر بالقيم الشاذة.

الكلمات المفتاحية: طريقة المربعات الصغرى الاعتيادية، طريقة المربعات الصغرى الحصينة ذات المرحلتين، عدم تجانس تباين الخطأ، القيم الشاذة، محاكاة مونتي كارلو، متوسط الخطأ النسبي المطلق.

Comparison of Ordinary Least Square (OLS) with Two-Step Robust Weight Least Square (TSRWLS) in Estimating the Parameters of the Multiple Linear Regression Model with Heteroscedastic and Outliers in the Response Variable

Assist. Prof. Ghufran I. Kamal
ghufranka62@gmail.com

Shaimaa I. Khalil AL-Obaidi
shaimaa595@yahoo.com

Statistics Department - College of Administration and Economics -University of Baghdad
Baghdad – Iraq

Received 24/8/2020 **Accepted 1/9/2020**

Abstract: *The main objective of this research is to use some methods to estimate the parameters of the multiple linear regression model. The first method is the classical method, the method of ordinary least square (OLS) and the second method is the Two-Step Robust Weighted Least Squares (TSRWLS). To show the effect of each of them in estimating the parameters in light of the problem of heterogeneity of error variance and the appearance of anomalies in the data that suffer from these two problems together. This was done using Monte Carlo simulation and through the Mean Absolute Percentage Error (MAPE) comparison parameter, and applied it to real data in the field of water hardness taken from the Municipality of Baghdad-Baghdad Water Directorate-Department of Quality Control on 2019. It has been found that the Two-Step Robust Weighted Least Squares method is best method for addressing the problem of heterogeneity of error variance without being affected by anomalies values.*

Keywords: The classical methods, Two-Step Robust Weighted Least Squares, Heterogeneity, monte-carlo simulation, The Mean Absolute Percentage Error (MAPE), Outliers.

[1] بحث مستل من رسالة ماجستير

ISSN (1681- 6870)

1. المقدمة

إن أنموذج تحليل الانحدار الخطي المتعدد من الاساليب الاحصائية الأكثر انتشاراً واستعمالاً لتحليل العلاقة بين متغير الاستجابة والمتغيرات التوضيحية، وينتج عن هذه العلاقة معادلة احصائية تضم هذه المتغيرات والتي من خلالها نستطيع الاعتماد عليها لقياس درجة التوافق بينهم، وذلك لغرض معرفة مدى دقة النتائج والاستنتاجات التي يتم التوصل اليها في نهاية الدراسة، وهنالك عدة طرائق لتقدير معالم الانموذج الاحصائي سواء كان بسيطاً او متعدداً، وتعتبر طريقة المربعات الصغرى (OLS) الاكثر انتشاراً في هذه النماذج لما تتميز به من السهولة في الحساب والدقة في التقديرات، الا أن هذه الطريقة تعتبر غير كفوءة في حالة عدم تحقق احدى الشروط الخاصة بأنموذج الانحدار وهو تجانس تباين الخطأ فعند التقدير يتم الحصول على تقديرات واستنتاجات مضللة وغير دقيقة، ان طريقة المربعات الصغرى الموزونة (WLS) هي الطريقة البديلة لـ(OLS) والتي من خلالها يتم معالجة مشكلة عدم تجانس تباين الخطأ والتوصل الى تقديرات جيدة، لكن نتيجة للأزمات والكوارث التي ترافق سائر الامم والشعوب في جميع مراحل الحياة، ظهرت مشكلة جديدة هي ظهور قيم غير اعتيادية ومختلفة عن سائر باقي القيم من حيث كبرها او صغرها سميت بالقيم الشاذة، وعندها ايضاً اصبحت هذه الطريقة الموزونة لمعالجة عدم تجانس تباين الخطأ تفشل بشكل كبير عند اصطدامها بوجود هذه القيم الشاذة (Outliers) والتي يجب التعامل معها بحذر كبير لأنها قد تشوه ادوات التشخيص، و تؤثر على كافة عمليات التقدير للأنموذج، ولتفادي مشكلة عدم تجانس تباين الخطأ في حالة وجود القيم الشاذة كان لابد من استعمال طرائق حصينة تعالج المشكلتين معاً.[15] [11].

الهدف من هذه الدراسة بالدرجة الأساس تسليط الضوء على البيانات التي تحتوي على بعض من المشكلات التي تواجه الباحثين والعاملين في مختلف القطاعات العلمية والحياتية، وهما مشكلتا عدم تجانس تباين الخطأ والقيم الشاذة (الملوثة)، والتوصل الى تقديرات لمعلمات انموذج الانحدار الخطي المتعدد الذي يعاني من تلك المشكلتين اعلاه، والتوصل الى افضل التقديرات للأنموذج وذلك من خلال (طرائق حصينة) تعالج مشكلة عدم تجانس تباين الخطأ بدون التأثر بوجود القيم الشاذة والمقارنة بين هذه الطرائق لاختيار افضلها، وتم ذلك باستعمال محاكاة مونتي كارلو ومن خلال معيار المقارنة متوسط الخطأ النسبي المطلق (MAPE)، ومن ثم تطبيقها على بيانات حقيقية مأخوذة من أمانة بغداد / دائرة ماء بغداد /قسم السيطرة النوعية للعام (2019م).

2. مشكلة عدم تجانس تباين الخطأ العشوائي Heteroscedasticity Problem

احدى الفرضيات الاساسية والتي تعتبر ركيزة من الركائز التي يقوم عليها النموذجان الخطيان (البسيط والعام) هو تجانس تباين الخطأ (ثبات التباين لحدود الخطأ), ويصبح الفرض فيهما كالآتي [1] :

في الأنموذج البسيط:

$$E(U_i^2) = \sigma_u^2$$

وفي الأنموذج العـام:

$$E(U\grave{U}) = \sigma_u^2 In$$

ولكن ما يحدث في الواقع التطبيقي والذي يواجهه اغلب الباحثين هو عدم تحقق الشرط اعلاه، اي يصبح التباين غير ثابت لجميع المشاهدات، وهذا يظهر جليا" في القطر الرئيسي لمصفوفة التباين والتباين المشترك (معلمة القياس) والذي يصبح متضمن قيم مختلفة له، ويكون الفرض هو كالآتي [15]

$$E(U\grave{U}) \neq \sigma_u^2 In$$

وهو ما يطلق عليه (Heteroskedasticity)، والمقصود هو عدم تجانس تباين الخطأ العشوائي أو ما يسمى احيانا بالـ (العنصر الاضطرابي Disturbances terms) :[6]

وتظهر أغلب هذه المشكلات بصورة خاصة في الدراسات التي تعتمد على البيانات المقطعية (Cross−Sectional data) كون إن كل مشاهدة فيها يختلف اختلافا كبيرا في قيم المتغيرات التوضيحية الأمر الذي يؤثر على مشاهدات متغير الاستجابة.

وهنالك عدة اختبارات وطرق للتشخيص تستعمل للكشف عن مشكلة عدم تجانس التباين للخطأ:

- **طرق تخطيطية:** تعتمد على رسم العلاقة بين المتغيرات الداخلة في البيانات.
- **طرق تحليلية:**
  - "اختبار بارك ـ كليجسر" - Park-Glejser Test .
  - "اختبار باغان كودفري"- Breusch−Pagan−Godfrey Test .

3. القيم الشاذة The Outliers

إن وجود قيم غريبة في العينة المختارة قد اصطلح على تسميتها علميا بالمشاهدات أو القيم الشاذة، وهي تعتبر واحدة من المشكلات الاحصائية المعروفة لدى الباحثين، وان اغلب المقاييس والوسائل الاحصائية والمعروفة لدى الاحصائيين مثل (الوسط الحسابي، المنوال، الانحراف المعياري،...) تكون حساسة للغاية تجاه القيم الشاذة، والتي تكون ذات اثر واضح على تغيير نتائج التحليل المعتمد ويكون هذا التغيير كبيرا كلما زاد عدد هذه القيم، وعلى الرغم من هذا كله لا يمكن اسقاطها او اهمالها لمجرد كونها

ISSN (1681- 6870)

قيمة شاذة، لأنها يمكن ان تكون الاكثر اثارة للاهتمام ومن المهم التحري عنها ودراستها وتحليلها قبل اتخاذ القرار، وفي العديد من البحوث والدراسات نوقشت هذه المشكلة لأهميتها، وإن البيانات التي لا تحتوي على شواذ تعتبر حالة استثنائية في الواقع العملي والتطبيقي، لأنها تحدث بسبب اخطاء شائعة غير مقصودة مثل اخطاء (القياس، التسجيل، المعاينة)، او تحدث بسبب (ظروف طبيعية او حدوث ازمات او كوارث ...) وبشكل عام نستنتج من ذلك بانها اما تحدث بسبب اخطاء او انها قيم حقيقية لكنها شاذة (متطرفة) ومن هذه الاهمية تم التطرق الى العديد من التعريفات التي تخص المشاهدات الشاذة لمحاولة تفسيرها وفهمها نذكر بعضاً منها: .[8][13] [14].

- عرف (Bross) عام (1961):"القيم الشاذة بأنها تلك القيم التي تظهر منحرفة بصورة كبيرة عن سائر مكونات قيم العينة التي اخذت منها" .
- والعالم (Barnett) عام (1978): فقد عرف القيمة الشاذة "هي تلك القيمة التي تظهر غير منسقة إذا ما قورنت بسائر قيم البيانات الاخرى".
- والعالم (Freeman) فقد عرفها في العام (1980): "بأنها تلك القيمة التي لم تتولد بالطريقة الاعتيادية التي ولدت الأغلبية العظمى من قيم باقي البيانات".

وهنالك عدة طرق للكشف عن القيم الشاذة نذكر بعضاً منها:[10] [13]

- "الرسم الصندوقي" – .(Box Plot).
- "بواقي ستيودنت المحذوفة" - .(Studentized Deleted Residual).

4. اختبار الكشف عن مشكلة عدم تجانس تباين الخطأ بوجود القيم الشاذة

Heteroscedasticity Problem Test With The Outlier Values

اختبار (كولد فيلد كوانت) المعدل Modification of the Goldfeld –Quandt Test.

إن الاختبارات التقليدية للكشف عن وجود مشكلة عدم تجانس تباين الخطأ، تكون غير كفوءة عندما تحتوي البيانات على مشاهدات شاذة، لذا اصبح من الحاجة تطوير اختبار لا يتأثر كثيرًا عندما تتضمن البيانات على تلك المشاهدات، لذا اقترح (Habshah) واخرون،اختبارًا جديدًا يعد بمثابة تعديل لاختبار Goldfeld-Quandt التقليدي، وذلك من خلال تحديد مكونات اختبار Goldfeld-Quandt التي تتأثر بالمشاهدات الشاذة ثم تستبدل هذه المكونات ببدائل حصينة أي تحصينه باستعمال طريقة المربعات الصغرى المشذبة (LTS) ليكون بذلك اكثر حصانة في التشخيص في ظل وجود النسب المختلفة للقيم الشاذة، وسمي هذا الاختبار اختبار كولد فليد كوانت المعدل Modified Goldfield-Quandt (MGQ) ، والذي سيكون اكفأ من الاختبارات التقليدية ، ويتضمن هذا الاختبار الخطوات الاتية: [14]

- **الخطوة الاولى:** نرتب قيم المشاهدات تصاعدياً او تنازليا" بناءً على قيم مصدر الاختلاف.
- **الخطوة الثانية:** تحذف المشاهدات الوسطية (c) ، حيث يتم تحديد واستبعاد (c) من مشاهدات المركز في حدود ربع المشاهدات الكلية:

$$C \cong \frac{1}{4} * n$$

والمتبقي من العينة (n − c) تقسم الى مجموعتين ، كل مجموعة منهما تشتمل على ($\frac{n-c}{2}$) من المشاهدات .

- **الخطوة الثالثة:** استعمال أحد اساليب الانحدار الحصين (طريقة المربعات الصغرى المشذبة (Least Trimmed Squares Method (LTS)) المقترحة من قبل (Rousseeuw and Leroy) لتوفيق خط الانحـــــدار، وإن هذه الطريقة لا تتأثر بوجود مثل هذه القيم الشاذة لكونها تمتاز بالحصانة.[13]
- **الخطوة الرابعة:** استخراج البواقي المحذوفة، اي ايجاد البواقي المحذوفة للمجموعتين الاولى والثانية على التوالي ، ويتم بعد ذلك حذف القيم الشاذة الموجودة في البيانات لكلا العينتين الجزئيتين ،ثم حساب الوسيط لمربعات البواقي المحذوفة (Median of the Squared Deletion Residuals) (MSDR) من خلال الصيغة الاتية:

$$MGQ = \frac{\mathrm{MSDR}_2}{\mathrm{MSDR}_1} \tag{1}$$

$MSDR_1$و$MSDR_2$ هما الوسيطين لقيم مربعات البواقي على التوالي ، لأصغر واكبر تباين للمجموعة على التوالي، تحت افتراض التوزيع الطبيعي الاحصاءة (MGQ) تتبع توزيع (F) بدرجات الحرية لكل من البسط والمقام $\frac{(n-c-2K)}{2}$. لاختبار صحة فرضية العدم بوجود مشكلة عدم التجانس اذا كانت قيمة (F)المحتسبة اقل من قيمة (F) الجدولية في الجداول الاحصائية.

ISSN (1681- 6870)

5. مفهوم الحصانة Robustness Notion

"القوة والحصانة ضد الانحرافات" هو المفهوم العام للحصانة، والذي تم استعماله لأول مرة من قبل الباحث (Box) لعام (1953) وعرفها "بأنها تدل على قوة التقدير والحصول على افضل النتائج في حالة عدم توفر الشروط الاساسية للطرائق المعتمدة في التقدير"، ومن ثم قدمت اول نظرية عامة في مقال بعنوان "التقدير الحصين لمعلمة الموقع" للباحث (Huber) عام (1964).[9]

وكلمة الحصانة تطلق على المقدرات التي لا تتأثر في حالة عدم تحقق الافتراضات الأساسية ،وتكمن قوة تلك المقدرات بنقاط الانهيار العالية والكفاءة .

ويمكن تعريف نقطة الانهيار للمقدر بانها بداية اصغر جزء مقاوم لتلوث البيانات والذي يمكن ان يدمر المقدر بالكامل اذا صغر عن ذلك الجزء، ويصبح بعدها التقدير عديم الفائدة، ويعرف كذلك بانه مقياس القوة التي كلما كانت نقطة الانهيار اكبر كان المقدر افضل واطلق عليه بالمقدرات الحصينة .[17].

6. طرائق التقدير الحصينة Robust Estimation Method

6.1 المربعات الصغرى الاعتيادية (OLS) Ordinary Least Square

هي احدى اهم الطرق واكثرها استعمالاً عند تحقق جميع فرضياتها، ولكن عند عدم تحقق واحدة او اكثر من تلك الفرضيات فأنها ستفقد كافة مميزاتها التي من ضمنها واهمها الحصول على افضل مقدر خطي غير متحيز "BLUE"، وفي ظل الفروض الاساسية يتم الحصول على تلك المقدرات وكما في الصيغة الاتية:[1]

$$b_{ols} = (X'X)^{-1}X'Y \tag{2}$$

6.2 طريقة المربعات الصغرى الحصينة ذات المرحلتين

(Two-Step Robust Weighted Least Squares (TSRWLS))

اذا كان لدينا انموذج الانحدار الخطي المتعدد كما في الصيغة التالية:

$$Y_i = \beta_0+\beta_1\ X_{i1}+\beta_2X_{i2} + \cdots + \ \beta_kX_{ik} + U_i \tag{3}$$

حيث يكون المقدر التقليدي في تقدير هذا الانموذج هو مقدر المربعات الصغرى (OLS) ويمتلك هذا المقدر خاصية افضل مقدر خطي غير متحيز عند تحقق الافتراضات الخاصة بأنموذج الانحدار الخطي ولكن عند خرق الافتراض الخاص بثبات تباين قيم لخطأ

$$\mathrm{var}(U_i) = \sigma_u^2$$

سوف يعاني الانموذج من مشكلة عدم ثبات تجانس تباين الخطأ، فـي هذه الحالة فـان استخدام مقـدر (OLS) سوف يكون متحيزاَ وغير كفوء، وان مصفوفة التباين والتباين المشترك له تأخذ الصيغة التالية :

$$cov(b) = (X'X)^{-1}X'\Omega X(X'X)^{-1} \tag{4}$$

اذ ان:

$$E(e'e) = \ \Omega$$

Ω : هي مصفوفة محددة موجبة تحت ثبات التجانس من درجة $(n \times n)$.

وعندما يكون التباين ثابتاً تكون $\Omega = \sigma^2 I_n$ اي ان $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$ ففي هذه الحالة فان مصفوفة التباين والتباين المشترك تأخذ الصيغة الاتية:[15] .

$$cov(b) = \sigma^2\ (X'X)^{-1}$$

يمكن حساب ($\hat{\sigma}^2$) والتي يمكن تقديرها من الصيغة الاتية:

$$\hat{\sigma}^2 = \frac{e'e}{n-p}$$

$e = \ e_1\,, e_2\,, \dots, e_n$ هو موجه بدرجة n لبواقي (OLS).

ISSN (1681- 6870)

وعند وجود مشكلة عدم ثبات تجانس تباين الخطأ فأن $\Omega = \sigma^2 Z$ حيث ان Z مصفوفة قطرية. وان تباين المعلمات المقدرة يصبح كالاتي:

$$V(b) = \sigma^2 (X'X)^{-1} X'ZX (X'X)^{-1}$$

تعرف $w = (Z^{-1})$ بأنها مصفوفة قطرية مع عناصر قطرية للأوزان ($w_1, w_2, \dots, w_n$) وبذلك يصبح مقدر المربعات الصغرى الموزونة بالصيغة الاتية :

$$b_{wls} = (X'WX)^{-1} X'WY \tag{5}$$

وان مصفوفة التباين والتباين المشترك لمقدرات WLS تأخذ الصيغة الاتية :

$$cov(b_{wls}) = \sigma^2_{wls} (X'WX)^{-1}$$

$$\hat{\sigma}^2_{wls} = \frac{\sum w_i \, e_i^2}{(n-p)} \tag{6}$$

وعند وجود قيم شاذة في متغير الاستجابة فإن الاوزان الخاصة بطريقة المربعات الصغرى الموزونة سوف تتأثر كثيراً بوجودها اذا لم يتم معالجتها بشكل صحيح لأنها سوف تؤثر على مقدرات المعالم مما تكون هذه المقدرات متحيزة وغير كفؤة، لذا لابد من ايجاد مصفوفة اوزان مناسبة وذات اداء جيد بوجود مشكلة (عدم ثبات تجانس تباين الخطأ والقيم الشاذة).
ولإيجاد مصفوفة اوزان حصينة سوف يتم استعمال طريقة المربعات الصغرى الحصينة ذات المرحلتين (TSRWLS) المقترحة من قبل (Habshah)في عام (2009) [11] وذلك من خلال تطوير لخوارزمية KNN للعالم Kutner واخرون [Kutner et al,2004] وذلك من خلال استخدام مقدر (LTS) بدلاً من مقدر (OLS) في خوارزمية KNN وذلك للحصول على اوزان حصينة اولية لذا فان طريقة المربعات الصغرى الموزونة الحصينة ذات المرحلتين تتألف من الخطوتين الآتيتين: [15]

- الخطوة الاولى: تتضمن تحديد اوزان اولية.
- والخطوة الثانية: تتضمن تحديد اوزان نهائية .

وكما يأتي:

- **الخطوة الاولى**

1. ايجاد القيم التقديرية لـ (y_i) ثم ايجاد قيم البواقي (e_i) من انموذج الانحدار الخطي باستعمال طريقة المربعات الصغرى المشذبة (LTS) .
2. انحدار القيم المطلقة للبواقي | e_i| والتي يشار اليها | $si=|e_i$ على(y_i) كذلك باستعمال طريقة LTS.
3. ايجاد القيم التقديرية لـ s_i اي ايجاد $\hat{s}_i$ (من الخطوة الاولى -2).
4. حساب الاوزان الاولية الحصينة من خلال صيغة معكوس مربعات القيم المقدرة لـ ($\hat{S}i$) حسب الصيغة التالية:

$$w_{1i} = \frac{1}{(\hat{s}_i)^2} \tag{7}$$

- **الخطوة الثانية**

لتحديد الاوزان النهائية لابد من استعمال دوال اوزان حصينة كدالة هوبر Huber function او دالة Bisquare function ، وفي هذه الحالة سوف يتم استعمال Huber function كدالة اوزان والتي تعرف كالاتي : [10]

$$w_{2i} = \begin{cases} 1 & |e_i| \le 1.345 \\ \dfrac{1.345}{|e_i|} & |e_i| > 1.345 \end{cases} \tag{8}$$

حيث ان 1.345 ثابت يدعى ثابت التناغم او الضبط .
e_i : البواقي القياسية التي تم الحصول عليها من طريقة LTS بالخطوة الاولى (1).
وبضرب الاوزان الاولية w_{1i} بالأوزان w_{2i} لإيجاد الاوزان النهائية W_i اي :

ISSN (1681- 6870)

$$W_i = w_{1i} * w_{2i} \tag{9}$$

ان معاملات الانحدار التي تم الحصول عليها من هذه الطريقة هي التقدير المطلوب لأنموذج الانحدار الخطي المتعدد لمشكلة عدم تجانس تباين الخطأ وظهور القيم الشاذة، واخيراً سوف يكون اداء المربعات الصغرى الموزونة WLS باستعمال الاوزان النهائية جيد في تقدير معلمات انموذج الانحدار الخطي المتعدد العام بوجود مشكلتي عدم التجانس والقيم الشاذة.

$$b_{Tswls} = (X'W_iX)^{-1}X'W_iY \tag{10}$$

7. معيار المقارنة Compare Criterion

لغرض التوصل للمقدر الاكفأ وجدت عدة معايير (مقاييس) للمقارنة بين طرائق التقدير ، وإن افضلها من يمتلك اقل خطأ ممكن وهذا يقودنا بالتالي للأنموذج الافضل لغرض التقدير و التنبؤ للظاهرة تحت الدراسة. وأحد هذه المعايير هو متوسط الخطأ النسبي المطلق (Mean Absolute Percentage Error) المعتمد للمقارنة في دراستنا ، ويمكن حسابه بشكل نسبي من خلال الصيغة التالية: [2][3]:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \mathrm{y}_i}{Y_i}\right| \times 100 \tag{11}$$

إذ أن:

$\boldsymbol{Y_i}$: تمثل قيم متغير الاستجابة.

y_i : يمثل قيم المتنبأ بها لمتغير الاستجابة $\boldsymbol{Y_i}$.

وبمقارنة المقياس بالنسبة للأنموذج الأفضل فالمعادلة التي تمتلك اقل قيمة للمعيار (متوسط للخطأ النسبي المطلق) يكون هو الافضل .

8. مراحل وصف تطبيق تجربة المحاكاة

سيتم الاعتماد على أنموذج الانحدار الخطي المتعدد حسب الصيغة (3) ، لوصف المراحل التجريبية للمحاكاة والمقارنة بين الطرائق الحصينة لاختيار افضلها حسب معيار متوسط الخطأ النسبي المطلق (MAPE). تمت كتابة البرنامج الاحصائي بالاعتماد على لغة الـR وهي "عبارة عن مجموعة متكاملة من البرمجيات التي تسمح بمعالجة البيانات والقيام بالعمليات الحسابية واظهار البيانات الرسومية " ،وهو من اللغات الحديثة التي صعد نجمها وبشكل متزايد في مجال البرمجة العلمية في قطاعي الاحصاء والمعلوماتية الحيوية، اذ يتم وصف تجارب المحاكاة من خلال المراحل والخطوات الآتية :-

- تم توليد عينات بأحجام مختلفة (n=50,99,150) للمتغير المعتمد (Y) وفق أنموذج الانحدار الخطي المتعدد.

$$\mathrm{Yi} = \mathrm{B}_0 + \mathrm{B}_1\mathrm{X}_{\mathrm{i}1} + \mathrm{B}_2\mathrm{X}_{\mathrm{i}2} + \mathrm{B}_3\mathrm{X}_{\mathrm{i}13} + \mathrm{B}_4\mathrm{X}_{\mathrm{i}4} + \mathrm{B}_5\mathrm{X}_{\mathrm{i}5} + U_{\mathrm{i}} \quad i = 1,2,3,\ldots,n \tag{12}$$

- لتوليد نموذج الانحدار الغير متجانس يتم من خلال الصيغة التالية:

$$\sigma_i^2 = \sigma^2 \ \mathrm{Exp}(\mathrm{a}x_{1i} + \mathrm{a}\,x^2{}_{2i} + \mathrm{a}\,x^3{}_{3i} + \mathrm{a}\,x^4{}_{4i} + \mathrm{a}\,x^5{}_{5i}) \tag{13}$$

حيث ان $\sigma^2 = 1$.

a : يمثل ثابت اعتباطي (عشوائي).

- ولتحديد مستوى عدم التجانس تباين الخطأ، كان من خلال المقياس التالي:

$$\sigma = \max(\sigma_i^2) / \min(\sigma_i^2) \qquad \mathrm{i}=1,2,\ldots,\mathrm{n} \tag{14}$$

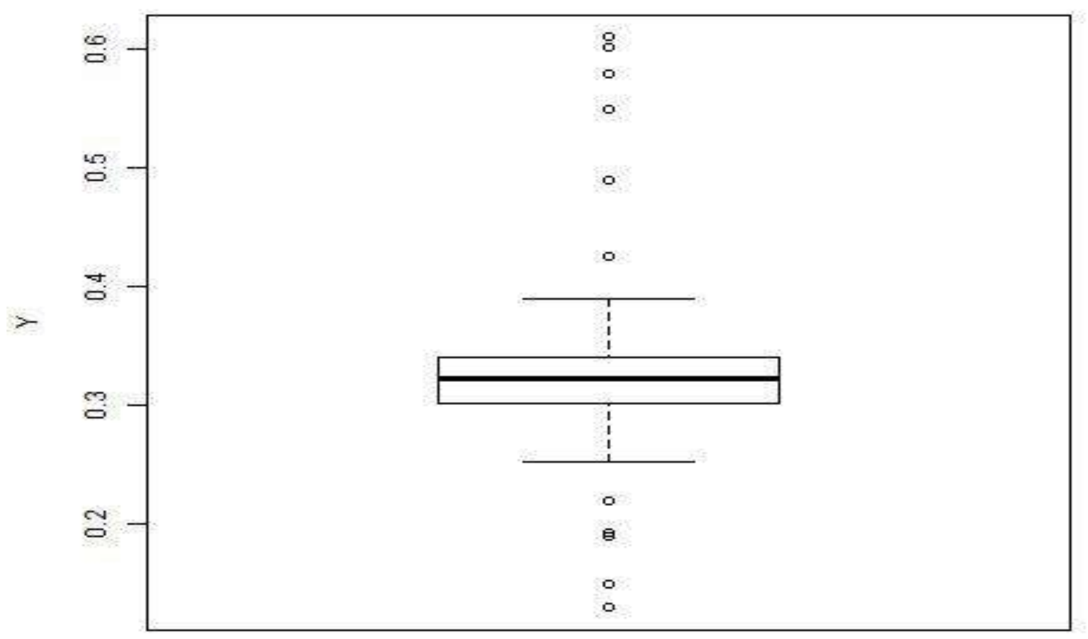
ISSN (1681- 6870)

وتم تحديد نسبة عدم تجانس تباين الخطأ ولكل حجم عينة بين { a=0 و a=1.5 و a=2.1 }، فعندما تكون قيمة (a=0) يكون (σ=1) وهي النسبة التي تكون عندها البيانات متجانسة ، وعند(a=1.5 , a=2.1) على التوالي يكون (σ=2.8 و σ=3.8) وهي النسب التي تدل على وجود مشكلة عدم تجانس تباين الخطأ، ويتم هذا كله بهدف ايجاد النموذج غير المتجانس وبعد تكرار التجربة ولحجوم العينات المختلفة (n=50,99,150)، وبتطبيق الصيغ (13)و(14) كانت النتيجة للبيانات هي (σ =2.8) التي دلت في دراستنا هذه على وجود مشكلة عدم تجانس التباين وبنسبة وجود للقيم الشاذة (10%) وحجم عينة (n=99) .

- تحديد واختيار القيم الافتراضية للمعالم بالاعتماد على المعالم الحقيقية، وتعد هذه المرحلة من اهم المراحل التي يعتمد عليها لاحقاً .
- توليد المتغيرات التوضيحية (X_{ij}) الخمسة، وتم هذا بالاعتماد على توزيعها في البيانات الحقيقية، وباستعمال الدوال الجاهزة في برنامج (R) وكما يلي :

X_1~N(0.0925,0.01)
X_2~N(0.0737,0.005)
X_3~N(0.065,0.034)
X_4~N(0.168,0.02)
X_5~N(0.542,0.01)

- e_i~ N (0,1) + Cauchy (0,10)
- تعويض المتغيرات التي تم توليدها في اعلاه لأجل الحصول على متغير الاستجابة (y) .
- يتم حساب إحصاءه الطرائق الحصينة والمقارنة بينهم عن طريق المعيار (MAPE)، وبتكرار (1000) مرة ...[15].



شكل (1): يظهر القيم الشاذة في متغير الاستجابة (Y) بواسطة Box-Plot

ISSN (1681- 6870)

جدول (1): يبين مقدرات المعلمات وقيم (MAPE) للأنموذج الخطي المتعدد (n=50) و(σ =1)

المعلمات	القيم الافتراضية	نسب القيم الشاذة					
		%0		%10		%20	
		OLS	TSRWLS	OLS	TSRWLS	OLS	TSRWLS
b_0	-0.38967	-0.38450	-0.38386	0.01786	-0.37299	0.19849	-0.38480
b_1	0.75359	0.76788	0.76668	0.00335	0.74951	0.30634	0.74508
b_2	3.54956	3.52446	3.51932	4.02347	3.55103	3.67233	3.63170
b_3	-0.07543	-0.07269	-0.07254	-0.04009	-0.05987	-0.08408	-0.07574
b_4	2.62533	2.63098	2.63013	2.48998	2.59476	2.88914	2.62836
b_5	0.21939	0.22348	0.22300	0.22494	0.21661	0.31416	0.22492
MAPE		0.10434	0.10207	0.70183	0.12441	1.21344	0.11039

جدول (2): يبين مقدرات المعلمات وقيم (MAPE) للأنموذج الخطي المتعدد (n=99) و(σ =1)

المعلمات	القيم الافتراضية	نسب القيم الشاذة					
		%0		%10		%20	
		OLS	TSRWLS	OLS	TSRWLS	OLS	TSRWLS
b_0	-0.38967	-0.37638	-0.37651	-0.10460	-0.38019	0.13410	-0.37778
b_1	0.75359	0.70334	0.70280	1.22203	0.75126	1.14084	0.73698
b_2	3.54956	3.57158	3.57182	3.27373	3.59906	3.59857	3.50597
b_3	-0.07543	-0.07822	-0.07831	0.04321	-0.07582	-0.05077	-0.07292
b_4	2.62533	2.62348	2.62437	2.56391	2.61815	2.96692	2.62999
b_5	0.21939	0.22165	0.22169	0.23412	0.22123	0.26961	0.21933
MAPE		0.07912	0.06751	0.52986	0.10695	0.92014	0.10896

جدول (3): يبين مقدرات المعلمات وقيم (MAPE) للأنموذج الخطي المتعدد (n=150) و(σ =1)

المعلمات	القيم الافتراضية	نسب القيم الشاذة					
		%0		%10		%20	
		OLS	TSRWLS	OLS	TSRWLS	OLS	TSRWLS
b_0	-0.38967	-0.37574	-0.37604	-0.09622	-0.38343	0.20569	-0.37795
b_1	0.75359	0.73427	0.73519	0.99921	0.76140	0.89736	0.73630
b_2	3.54956	3.55531	3.55616	3.94245	3.59967	3.12221	3.60315
b_3	-0.07543	-0.07860	-0.07835	-0.07153	-0.07214	0.14404	-0.07094
b_4	2.62533	2.61482	2.61550	2.52035	2.63003	2.63519	2.61807
b_5	0.21939	0.21949	0.21959	0.25767	0.22115	0.27153	0.21873
MAPE		0.06341	0.05180	0.42007	0.08309	0.74983	0.10708

جدول (4): يبين مقدرات المعلمات وقيم (MAPE) للأنموذج الخطي المتعدد (n=50) و(σ =2.8)

المعلمات	القيم الافتراضية	نسب القيم الشاذة					
		%0		%10		%20	
		OLS	TSRWLS	OLS	TSRWLS	OLS	TSRWLS
b_0	-0.58967	-0.60519	-0.59185	-1.05608	-0.61115	-1.65364	-0.57671
b_1	0.55359	0.70119	0.50757	5.40694	0.61617	3.62859	0.53301
b_2	3.00000	3.24177	3.17669	1.95346	3.19235	5.47515	2.98886
b_3	-0.09543	-0.01555	-0.04524	2.04187	-0.03819	3.36225	-0.06755
b_4	2.00000	2.01005	2.05630	2.52614	2.05531	5.96604	1.92411
b_5	0.11939	0.16994	0.16657	1.04116	0.18479	4.07302	0.18834
MAPE		2.46627	0.72808	33.96492	0.72065	38.47339	0.70951

ISSN (1681- 6870)

جدول (5): يبين مقدرات المعلمات وقيم (MAPE) للأنموذج الخطي المتعدد (n=99) و(σ =2.8)

المعلمات	القيم الافتراضية	نسب القيم الشاذة					
		%0		%10		%20	
		OLS	TSRWLS	OLS	TSRWLS	OLS	TSRWLS
b_0	-0.58967	-0.59532	-0.59528	-1.36009	-0.60050	-1.63426	-0.59025
b_1	0.55359	0.63911	0.63683	3.12235	0.61536	4.96038	0.57009
b_2	3.00000	3.18920	3.20829	2.06886	3.32015	7.84201	3.15428
b_3	-0.09543	-0.03611	-0.03689	0.86628	-0.03351	3.47390	-0.08774
b_4	2.00000	2.00781	2.01194	5.40699	2.04586	5.28273	1.98837
b_5	0.11939	0.16536	0.16399	2.14149	0.16195	3.95657	0.18399
MAPE		1.46458	0.56917	20.97561	0.57257	46.96676	0.61093

جدول (6): يبين مقدرات المعلمات وقيم (MAPE) للأنموذج الخطي المتعدد (n=150) و(σ =2.8)

المعلمات	القيم الافتراضية	نسب القيم الشاذة					
		%0		%10		%20	
		OLS	TSRWLS	OLS	TSRWLS	OLS	TSRWLS
b_0	-0.58967	-0.61140	-0.61052	-1.23602	-0.61265	-1.86031	-0.60053
b_1	0.55359	0.71905	0.71024	2.22645	0.56555	3.29239	0.52555
b_2	3.00000	2.85511	2.84984	3.95298	3.16054	4.81556	3.22972
b_3	-0.09543	-0.01316	-0.01274	2.30539	0.00152	4.21201	-0.04153
b_4	2.00000	2.04642	2.04889	3.81797	2.03977	6.27659	2.03027
b_5	0.11939	0.17744	0.17677	2.27287	0.19438	4.30828	0.18483
MAPE		1.33710	0.48855	19.79399	0.43993	34.60029	0.48350

جدول (7): يبين مقدرات المعلمات وقيم (MAPE) للأنموذج الخطي المتعدد (n=50) و(σ =3.8)

المعلمات	القيم الافتراضية	نسب القيم الشاذة					
		%0		%10		%20	
		OLS	TSRWLS	OLS	TSRWLS	OLS	TSRWLS
b_0	-0.18967	-0.25643	-0.24937	-3.31190	-0.17958	-4.22824	0.14542
b_1	0.95399	1.32240	1.30312	6.89649	1.15106	7.42969	1.04403
b_2	4.00000	3.62570	3.60429	21.90240	3.46570	5.11339	3.86291
b_3	-0.05543	0.07119	0.06166	6.30013	0.06786	11.73614	0.27101
b_4	3.20000	3.33338	3.33551	8.74505	3.58650	10.72370	2.94753
b_5	0.31939	0.45158	0.44426	5.57962	0.53742	10.10190	0.45229
MAPE		2.06646	1.72484	83.59299	1.91832	12.09026	0.96556

جدول (8): يبين مقدرات المعلمات وقيم (MAPE) للأنموذج الخطي المتعدد (n=99) و(σ =3.8)

المعلمات	القيم الافتراضية	نسب القيم الشاذة					
		%0		%10		%20	
		OLS	TSRWLS	OLS	TSRWLS	OLS	TSRWLS
b_0	-0.18967	-0.28240	-0.27633	-2.38120	-0.15619	-5.94204	-0.27430
b_1	0.95399	1.23245	1.22266	3.42885	1.33031	19.68543	1.21104
b_2	4.00000	4.17144	4.11505	17.50950	4.07317	12.76553	4.51434
b_3	-0.05543	0.13094	0.13183	5.64516	0.05021	9.94915	0.05838
b_4	3.20000	3.25821	3.24600	8.79699	3.48980	13.41299	3.29584
b_5	0.31939	0.50499	0.50154	4.77104	0.47713	10.30337	0.48244
MAPE		1.17627	0.92528	8.63349	1.68164	15.13158	1.72045

ISSN (1681- 6870)

جدول (9): يبين مقدرات المعلمات وقيم (MAPE) للأنموذج الخطي المتعدد (n=150) و(σ =3.8)

المعلمات	القيم الافتراضية	نسب القيم الشاذة					
		%0		%10		%20	
		OLS	TSRWLS	OLS	TSRWLS	OLS	TSRWLS
b_0	-0.18967	-0.26212	-0.25828	-2.73581	-0.16219	-6.07170	-0.29584
b_1	0.95399	0.91803	0.91580	7.95493	1.33200	17.33128	1.28217
b_2	4.00000	4.20369	4.21258	14.38062	4.65786	15.12201	4.16496
b_3	-0.05543	0.09677	0.08797	5.13544	0.10379	10.72427	0.15596
b_4	3.20000	3.33690	3.32695	7.17710	3.36510	13.67727	3.39537
b_5	0.31939	0.49764	0.49521	5.26830	0.49113	10.62600	0.48095
MAPE		1.28158	1.17512	7.88989	1.56968	11.29274	1.29731

التحليل العملي لنتائج تجربة المحاكاة

- من ملاحظتنا للجداول اعلاه تبين تناقص قيم متوسط الخطأ المطلق النسبي (MAPE) كلما زادت احجام العينات المختلفة عند ثبات مقياس التجانس (σ) وهذا دليل على الخصائص الجيدة للمعيار " عندما تقترب قيمة المقدر من القيمة الحقيقية للمعالم عند زيادة حجم العينة.
- ومن الجداول (4)،(5)،(6)،(7)،(8)،(9) والتي تكون فيها البيانات تعاني من مشكلة عدم تجانس تباين الخطأ (σ =2.8,3.8) وبزيادة حجوم العينات ولكافة النسب للقيم الشاذة (%0,%10,%20) تمت ملاحظة تناقص قيم متوسط الخطأ المطلق النسبي المطلق (MAPE) لطريقة (TSRWLS) وهذا يدل على صحة العمل الاحصائي، والتي اثبتت كفاءتها في اغلب تجارب المحاكاة بوجود تلك المشكلتين.
- يلاحظ من ان قيم متوسط الخطأ المطلق النسبي (MAPE) تكون اصغر في حالة الطريقة الحصينة(TSRWLS)، مما هي عليه في الطريقة الكلاسيكية (OLS) ولكافة الجداول وبزيادة نسب الشواذ ونسب عدم التجانس للخطأ ولكافة حجوم العينات.

الجانب التطبيقي: تهيئة البيانات وتعريف متغيرات الدراسة

تم الحصول على البيانات من امانة بغداد / دائرة ماء بغداد / قسم السيطرة النوعية / مختبر مشروع ماء الوحدة ، واحتوت البيانات على ستة متغيرات منها خمسة أملاح اساسية تسبب عسرة المياه تمثل المدخلات التوضيحية للعملية اضافة للهدف وهو نسبة العسرة كمتغير تابع مخرج العملية تم استخدامها في هذه الدراسة لسنة 2019 ولعينة متكونة من (99) مفردة، والمتغيرات هي كالتالي :

- الكالسيوم المتمثل بـ(X1).
- المغنسيوم المتمثل بـ(X2).
- الكلورايد المتمثل بـ(X3).
- قاعدية الماء المتمثل بـ(X4).
- الاملاح المذابة في الماء المتمثل بـ(X5).
- عسرة الماء المتمثل بـ(Y).

الطريقة الحصينة في تقدير المعلمات:

في هذا الجزء تم تقدير معلمات الأنموذج باستعمال افضل طريقة من طرائق التقدير الحصينة التي تطابقت مع الجانب التجريبي (المحاكاة) وهي طريقة (TSRWLS) وذلك بالاعتماد على انها اظهرت اقل (MAPE) للأنموذج وفي اغلب الحالات، وبالاعتماد على عينة التطبيق المتمثلة بالبيانات الحقيقية لمعدلات عسرة المياه والعوامل المؤثرة عليه ، ومن خلال استعمال برنامج (R) الذي يعد من البرامج ذات الامكانية المتقدمة في اغلب مجالات الاحصاء ، تمت كتابة البرنامج لتقدير المعلمات للأنموذج ، وكانت النتائج كما في الجدول التالي:

جدول (10): يبين القيم التقديرية لمعلمات افضل طريقة (TSRWLS)

Param. / Methods	B_0	B_1	B_2	B_3	B_4	B_5
TSRWLS	-0.60050	0.61536	3.32015	-0.03351	2.04586	0.16195

ISSN (1681- 6870)

الاستنتاجات Conclusions

ان اختبار (Goldfeld-Quadt) المعدل الحصين، قدم تحسينات كبيرة على الاختبارات الشائعة الاستعمال، من خلال تطبيقه على البيانات الحقيقية والمحاكاة لمونت كارلو، واظهر اداءً رائعاً في الكشف عن مشكلة عدم تجانس التباين للخطأ في ظل وجود القيم الشاذة. واظهرت الطريقة الحصينة (TSRWLS) كفاءة اعلى من الطريقة التقليدية (OLS) في تقدير معلمات انموذج الانحدار الخطي المتعدد في حالة وجود مشكلة عدم تجانس تباين الخطأ وظهور القيم الشاذة .

التوصيات

1. يتم الاحتفاظ بسجل جيد للتجارب المؤقتة، لتسجيل جميع البيانات مع اي تفسير ممكن لها او معلومات اضافية تخص البيانات.
2. البحث عن طرائق حصينة اخرى (اي لا تتأثر بالقيم الشاذة) ومحاولة تطويرها، لمعالجة مشكلة عدم تجانس تباين الاخطاء في وجود القيم الشاذة.
3. استعمال متغيرات اخرى، غير تلك المعتمدة في بيانات الرسالة، مثل (كبريتات الكالسيوم، كلوريد المغنيسيوم) وغيرها من العوامل المعروفة بمدى تأثيرها على عسرة المياه ، حيث جرى استبعادها لعدم اكتمال البيانات حولها.

المصـــــادر

[1] أموري، هادي كاظم والقيسي، باسم شليبة، القياس الاقتصادي المتقدم النظرية والتطبيق، مطبعة الطيف، بغداد، العراق، 2002 م.

[2] اموري، هادي كاظم والدليمي، محمد مناجد، تحليل الانحدار بالأمثلة، جامعة بغداد، كلية الادارة والاقتصاد، مطابع التعليم العالي، العراق،1990.

[3] عبدالمجيد حمزة الناصر و صفاء يونس الصفاوي، "مقارنة بين المقدرات الاعتيادية والحصينة لنماذج السلاسل الزمنية المختلطة الثنائية من الرتب الدنيا"، المجلة العراقية للعلوم الاحصائية، العدد 8، (2005): 1- 19.

[4] ناسي، نبيل جورج، "تقييم كفاءة طرق تقدير القيم الشاذة لنماذج الانحدار"، رسالة ماجستير، اكلية الادارة والاقتصاد، جامعة بغداد، (2001).

[5] ابراهيم، بسام يونس واخرون، الاقتصاد القياسي، دار العزة للنشر والتوزيع، الخرطوم، السودان، (2002م) .

[6] السواعي، خالد محمد، "موضوعات متقدمة في القياس الاقتصادي"، المجلد الاول (486)، النشر الدار العربية للعلوم ، (2015).

[7] UN Department of Economic and social Affairs, “The human right to water and sanitation”, available at: https://www.un.org/waterforlifedecade/human_right_to_water.shtml

[8] Bross, I.D.J., "Outliers in Patterned Experiments: strategic Re-Appraisal", Technometrics, Vol. 3, No. 1 (Feb., 1961), pp. 91-102 (12 pages)

[9] Huber, P. J., Robust Statistics, Wiley, New York, (1981), pp183-184.

[10] M. Habshah, S. Rana, A. H. M. R. Imon, (2009), “The Performance of Robust Weighted Least Squares in the Presence of Outliers and Heteroscedastic”, WSEAS Transition of Mathematics, 8 (2008), pp.351 – 361.

[11] Özlem, Gürünlü Alma, "Comparison of Robust Regression Methods in Linear Regression", Muğla University, Faculty of Arts & Sciences Department of Statistics, Muğla, Turkey Int. J. Contemp. Math. Sciences, Vol. 6, No. 9, 2011, pp.409 – 421.

[12] Rousseeuw, P.J. and A. Leroy, ”Robust Regression and Outlier Detection”, 1st Ed., Wiley, New York, 1987, pp: 329.

[13] Midi, H. A. B. S. H. A. H., Rana, S. O. H. E. L., & Imon, A. H. M. R., ”Robust Estimation of Regression Parameters with Heteroscedastic Errors in the Presence of Outliers”, In WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering (No. 8), (2008), World Scientific and Engineering Academy and Society.

[14] 14-Midi, H. A. B. S. H. A. H., Rana, S. O. H. E. L., & Imon, A. H. M. R.,”On a Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers”, Proceedings of the World Congress on Engineering, Vol. I, July 3 - 5, 2013, London, U.K

ISSN (1681- 6870)

[15] White, Halbert. "A Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." Econometrica: journal of the Econometric Society (1980): 817-838.

[16] Coakley, Clint W., and Thomas P. Hettmansperger. "A bounded influence, high breakdown, efficient regression estimator." Journal of the American Statistical Association 88.423 (1993): 872-880.