المجلة العراقية للعلوم الإحصائية

http://stats.uomosul.edu.iq

الكشف عن القيم المتطرفة في نموذج الانحدار الخطي مع التطبيق على بيانات تلوث مياه الآبار أطراف مدينة الموصل

سجى مروان إسماعيل صفوان ناظم راشد

قسم الاحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

معلومات النشر

تاريخ المقالة:
تم استلامه في 22 شباط 2022
تم القبول في 30 نيسان 2022
تم القبول في 9 ايار 2022
متاح على الإنترنت في 1 حزيران 2022

الكلمات الدالة:
القيم المتطرفة، الكشف التشخيص، المعالجة، طريقة الحذف، مقدر M الحصين ،مقدر MM الحصين ومقدر M الحصين الموزونة

المراسلة:
سجى مروان إسماعيل
saja.20csp120@student.uomosul.edu.iq

الخلاصة

تهتم فكرة البحث في التعرف عن تأثير القيم المتطرفة على معلمات نموذج تحليل الانحدار الخطي المتعدد، حيث يتم الكشف عن القيم المتطرفة وتشخيصها والموجودة في البيانات إن كانت في المتغيرات المستقلة أو المتغير المعتمد مما يتسبب في التأثير على تقدير معلمات النموذج المدروس . وقد تم التعرف على أنواع البيانات المتطرفة وطرق معالجتها للحصول على نموذج أفضل ذات كفاءة عالية أو التقليل من تأثير هذه القيم على النموذج ، وتم وضع معيار متوسط مربعات الخطأ MSE لغرض المقارنة بين طرق المعالجة وتم تطبيقها على بيانات حقيقية مأخوذة من مركز بحوث السدود والموارد المائية جامعة الموصل ، وظهرت النتائج توضح أفضلية الكشف عن القيم الشاذة بطريقة الرسم الصندوقي Box-Plot كذلك أفضلية طرق المعالجة بطريقة مقدر M الحصين الموزونة المقترحة من قبل (Shaker ، 2009) الافضل في الكشف من بين الطرق التي تم استخدامها.

DOI: 10.33899/IQJOSS.2022.174334 , ©Authors, 2022, College of Computer Science and Mathematics, University of Mosul.

مقدمة

من المعلوم أن التقدير في الطرائق الاحصائية يعتمد على مجموعة من الفروض المهمة للحصول على نموذج انحدار دقيق ، وتعد معلومية التوزيع الاحتمالي للبيانات أحدى أهم الأحيان تأخذ البيانات الموزعة نمطاً مختلفاً وقد لا تتمثل بنمط معين من التوزيعات والسبب يعود أحياناً الى وجود القيم الشاذة (Outlier) وهو الأمر الذي يؤدي الى عدم التحقق في فروض المربعات الصغرى وعندها ستفقد خصائصها الجيدة وعليه يتم البحث عن طرق بديلة حصينة لمعالجة هذه المشكلة وتعطينا مقدرات كفوءة ، وقد تم البحث عن طرق للكشف عن القيم الشاذة التي ظهرت في البيانات منها طريقة الرسم الصندوقي ، وطريقة الـHat Matrix ، وطريقة حذف ستيودنت ، وقد وضعت أساليب عديدة لمعالجة هذه القيم عند تقدير معلمات نموذج الانحدار أهمها أسلوب الحذف ، طريقة مقدر M الحصين الذي يعالج الخلل في المتغير التابع دون معالجة الخلل الموجود في المتغيرات التوضيحية ، كذلك استخدام طريقة مقدر MM الحصين الذي يعالج الخلل في المتغيرات المستقلة والمتغير التابع ، كذلك طريقة مقدر M الحصين الموزونة المقترحة من قبل(شاكر ، 2009) .

تضمن هذا البحث جانبين ، الأول مفهوم القيم الشاذة وطرق الكشف عنها وتشخيصها ، وطرق كشف التأثير عن القيم الشاذة ، وأساليب معالجة القيم الشاذة .أما الجانب الثاني تمثل بتطبيق الطرائق الثلاثة للكشف على بيانات حقيقية المتمثلة بتلوث مياه الابار أطراف مدينة الموصل وتم الحصول على البيانات من مركز بحوث السدود والموارد المائية –جامعة الموصل وتم وضع معيار متوسط مربعات الخطأ MSE للمقارنة .

1. نموذج الانحدار الخطي المتعدد :

تعد طريقة المربعات الصغرى الاعتيادية أحد الطرق في تقدير معلمات النموذج التي تأخذ العلاقة :

$$Y = X\beta + e \quad (1)$$

تشترط توفر عدد من الشروط من أجل الحصول على مقدرات دقيقة لمعلمات نموذج والذي يتمثل بالمعادلة الأتية:

$$\hat{\beta}_{ols} = (X'X)^{-1}X'Y \quad (2)$$

ومن هذه الشروط هي التوزيع الطبيعي لمتجه الاخطاء وتجانس تباين الاخطاء (Heteroscedastic) إلا أن هذه الشروط لاتتحقق في كثير من الحالات والسبب يعود إلى وجود القيم الشاذة في متجه الأخطاء أو في قيم المتغيرات التوضيحية ، لذا فأن كشف هذه القيم ومعالجتها قبل تحليل البيانات أمر في غاية الأهمية من أجل الحصول على نتائج واقعية دقيقة (Maronna and Martia ,2006) وقد وضعت أساليب للكشف عن هذه القيم منها:

2. بعض طرق الكشف عن المشاهدات الشاذة:

a) الرسم الصندوقي :_هو أحد الطرائق الاستكشافية الحديثة لتعيين القيم الشاذة وهي طريقة العرض بالرسم الصندوقي Box_Plot بأستخدام خمس ملخصات المقترحة من قبل العالم (Tukey , 1977) ، وهو أحد الأساليب الرسومية التي توضح القيم ذات التطرف القوي (Extrem Outlier) والقيم ذات التطرف المعتدل وتحتويها البرامج والحزم الرسومية (الصائغ ,2013).

b) فحص عناصر قطر المصفوفة (Hat- Matrix): هي أحدى الطرق الأحصائية التي من خلالها يتم الكشف عن القيم الشاذة في المتغيرات المستقلة والتي تسمى أحياناً بقيم قوة الرفع (Leverage values) (يوسف ،2015)إذا إن:

$$H_{(n\times n)} = X(X'X)^{-1}X' = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \vdots & \vdots & & \vdots \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix} \quad (3)$$

فإذا كانت $h_{ii} > \frac{2p}{n}$ or $h_{ii} > \frac{3p}{n}$ or $h_{ii} > 0.2$ فهذا يدل على ان المشاهدة i ()هي مشاهدة متطرفة (Belsly and Walsch, 1980) .

❖ **طرق تشخيص القيم الشاذة في المتغيرات المستقلة:**

هناك ثلاث طرق لتشخيص المتغيرات المستقلة:

○ **طريقة بليسلي وآخرون:**أوضح (Belsly et al,1980) في هذه الطريقة عن المشاهدات الشاذة وبين فيما اذا كانت مؤثر ام لا على قوة الرافعة في المعادلة (3) كالأتي :

$$h_{ii} > \frac{2p}{n} \quad (4)$$

○ **طريقة نيتر وآخرون:** تشخيص القيمة الشاذة للمشاهدة i حيث اعتبر (Neter et al.,1990) إن الرافعة التي تزيد قيمتها عن 0.5 كبيرة كالاتي:

$$h_{ii} \geq 0.5 \quad (5)$$

○ **طريقة جون فوكس:** اقترح (Fox,1997) دراسة كل الحالات التي تزيد قيم رافعاتها عن ثلاثة أضعاف متوسط الرافعة.

$$h_{ii} > \frac{3p}{n} \quad (6)$$

c) **طريقة بواقي ستيودنت المحذوفة:**_ تمثل أحدى طرق الكشف عن القيم الشاذة في المتغير التابع ، والتي يتم الحصول عليها بايجاد القيمة المعيارية للبواقي المحذوفة (Deleted Residual)، حيث ان البواقي المحذوفة d_i للمشاهدة يساوي الفرق بين قيم (y_i) الفعلية والقيم المقدرة لها ($y_{i(i)}$) باستخدام نموذج الانحدار الخطي الذي تقديره باستبعاد المشاهدة (i)، مما يجعل تحليل البواقي أكثر فاعلية في الكشف عن المشاهدات القاصية في المتغير التابع (y) وتم التوصل إلى أفضل معادلة تم الاعتماد عليها في الدراسة حيث تم حساب بواقي ستيودنت أستناداً إلى المعادلة أدناه والتي تتبع توزيع t بدرجة حرية (n-k-1) وتستند في حسابها على الخطأ e_i ومجموع مربعات الخطأ SS_{Res} فضلاً عن قيم الرفع h_{ii} لمصفوفة الـ Hat-Matrix (المطيري، 2010):

$$d_i^* = e_i\left[\frac{n-k-1}{SS_{Res}(1-h_{ii})-e_i^2}\right] \sim t(n-k-1) \qquad (7)$$

❖ **طرق تشخيص القيم الشاذة في المتغير المعتمد:**

يتم تشخيص القيم الشاذة لمتغير التابع بمقارنة القيمة المطلقة لباقي ستيودنت المحذوفة d_i^* بقيمة توزيع t عند درجة حرية n-k -1 ومستوى معنوية α حيث تعتبر الحالة y_i حالة شاذة لابد من دراستها وتحديد مدى تأثيرها على مقدرات المربعات الصغرى(المطيري ,2010) .

$$\left|d_i^*\right| = t_{\alpha/2}, n-k-1 \qquad (8)$$

3. **المشاهدات الشاذة المؤثرة وطرق الكشف عنها:**

هنالك مقاييس يتم من خلالها معرفة إذا كانت المشاهدات الشاذة مؤثرة أم لا من هذه لمقاييس:

a. **مقياس DFFITS لتأثير القيم الشاذة:**

يستخدم مقياس DFFITS لقياس اثر المشاهدة i على القيمة المقدرة ، وتم الاعتماد على الصيغة أدناه لقياس أثر المشاهدة i على القيم المقدرة

$$(DFFITS)_i = d_i^*\left[\frac{h_{ii}}{1-h_{ii}}\right]^{\frac{1}{2}} \qquad (9)$$

❖ **طرق كشف التأثير لقيمة DFFITS على النموذج:**

يمكن تشخيص المشاهدة i باعتبارها مؤثرة على نتائج نموذج تحليل الانحدار الخطي المتعدد وهناك طرق كشف منها:-

○ **طريقة بليسلي وآخرون (Belsley et al., 1980):**

أختبر بليسلي الحالة مؤثرة على نتائج تحليل الانحدار الخطي المتعدد أستناداً إلى عدد معلمات النموذج p الموضح بالصيغة الأتية:

$$\left|DFFITS\right| > 2\sqrt{\frac{p}{n}} \qquad (10)$$

○ **طريقة شاترجي وهادي(Chatterjee and Hadi, 1988):**

اقترح كل من شاترجي وهادي معيار لمقارنة القيمة المطلقة لـ DFFITS بقيمة اكبر قليلاً من القيمة التي اقترحها بليسلي واخرون تعرف كالآتي:

$$\left|DFFITS\right| > 2\sqrt{\frac{p}{n-p}} \qquad (11)$$

b. **مقياس COVRATIO لتأثير القيم الشاذة:**

نستخدم مقياس الاثر على الأخطاء المعيارية Influence on standard Error الذي طوره بليسلي اثر على حالة مصفوفة تباين تغاير معاملات الانحدار المقدرة. وبعد مراحل من التطور تم التوصل إلى معادلة حسابية تعتمد على قيم الرافعة h_{ii} وقيم بواقي ستيودنت المحذوفة d_i^* وعدد معلمات النموذج p وعدد المتغيرات k وفق الصيغة الأتية:

$$COVRATIO = \frac{1}{(1-h_{ii})\left(\frac{(n-k-1)+d_i^{*2}}{n-k}\right)^p} \quad (12)$$

نلاحظ في المعادلة تزيد قيمة COVRATIO بزيادة قيمة الرافعة وانخفاض بواقي ستيودنت المحذوفة ويكون ذلك مؤثر جيد لاكتشاف قيم COVRATIO المؤثرة على الأخطاء المعيارية.

❖ **طريقة كشف التأثير لقيمة (COVRATIO):**

اقترح (Belsley et al. , 1980) مقارنة قيمة COVRATIO بالقيمة $1\pm\frac{3p}{n}$ لتشخيص اثر الحالة رقم (i) على الأخطاء المعيارية لمعاملات الانحدار أي إذا كانت قيمة COVRATIO خارج هذه الفترة فان المشاهدة رقم (i) تعتبر مؤثرة على قيم الأخطاء المعيارية لمعاملات نموذج الانحدار أعتماداً على الفترة الاتية:

$$1-\frac{3p}{n} < COVRATIO < 1+\frac{3p}{n} \quad (13)$$

c. **مقياس Cook's Distance لتأثير القيم الشاذة:_**

يستخدم مقياس (مسافة كوك) لقياس أثر المشاهدة (i) على كل معاملات نموذج الانحدار المقدرة، وقد تم اعتماد مقياس مسافة كوك على القيمة e_i, h_{ii} لأنه عندما تكون احد هاتان القيمتان كبيرة أو كلاهما فان قيمة مسافة كوك ستصبح كبيرة أيضاً.

$$D_i = \frac{e_i^2}{p * MSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right) \quad (14)$$

❖ **طريقة كشف التأثير لقيمة مسافة كوك Di :_**

طريقة فوكس: حيث اقترح فوكس طريقة في عملية الكشف القيمة الشاذة مؤثرة على قيم معاملات الانحدار وبخلاف ذلك تكون القيمة غير مؤثرة من خلال الصيغة الأتية:_

$$D_i > \frac{4}{n-p} \quad (15)$$

d. **مقياس (DFBETAS) لتأثير القيم الشاذة:**

يستخدم مقياس DFBETAS لقياس الفرق بين معاملات الانحدار المقدرة باستخدام كل المشاهدات وقيم معاملات الانحدار المقدرة بعد حذف المشاهدة رقم (i) في كل مرة. وهناك معادلة يجب حسابها في كل مرة يتم فيها توفيق النموذج بعد حذف المشاهدة رقم i مستنداً إلى الخطأ المعياري S_i وعنصر القطر k من مصفوفة $(X'X)^{-1}_{kk}$ وفي الصيغة الاتية :

$$DFBETAS_{k(i)} = \frac{b_k - b_{k(i)}}{S_i\sqrt{(X'X)^{-1}_{kk}}} \quad \text{for } k=0,1,2,\ldots,p \quad (16)$$

❖ **طريقـــة كشـــف التـــأثير لقيمة(DFBETAS):**

لتشخيص الحالات المؤثرة على قيمة معامل الانحدار .

- **طريقة نيتر وأخرون (Neter et al,1990):**

اقترح نيتر معياراً لتحديد الحالات المؤثرة في حالة العينات الصغيرة والمتوسطة $\left|DFBETAS_{k(i)}\right|>1$ أما في حالة العينات الكبيرة $\left|DFBETAS_{k(i)}\right|>\frac{2}{\sqrt{n}}$

4. طرق معالجة القيم الشاذة:

a. طريقة حذف المشاهدات

الشاذة : تستخدم هذه الطريقة أذا كان حجم العينة كبيرا (إسماعيل،2001) حيث يتم بناء نموذج جديد في كل مرة يتم فيها حذف مشاهدة شاذة(i) لحين الحصول على نموذج يبلغ فيه قيمة الخطأ (MSE) أقل ما يمكن فضلاً عن قيمة معامل التحديد R^2 والمختبر الاحصائي F .

b. **طريقة مقدر M الحصين :** Robust M Estimation

هو أحد طرق المعالجة الذي يتم من خلاله معالجة القيم المتطرفة في المتغير المعتمد وذالك بأستعمال الاساليب الحصينة التي أقترحها (Huber,1973) .وتم الاعتماد على الصيغة أدناه يمكن الحصول على معلمات النموذج وكما يلي:

$$\hat{\beta}_M=(X'W_MX)^{-1}X'W_MY \qquad (17)$$

هدف أسلوب M الحصين هو أعطاء أوزان صغيرة للمشاهدات غير الاعتيادية (المتطرفة) من خلال عناصر القطر للمصفوفة القطرية W، ولتطبيق أسلوب M يتطلب الأمر مقدر ابتدائي وأسلوب تكراري للوصول في النهاية إلى تقارب في مقدرات M للمعلمة β ، ويدعى هذا الأسلوب بأسلوب المربعات الصغرى الموزونة (IRLS) (RousseeandLeroy,1987).

c. طريقة مقدر MM الحصين:

هو أحد الأساليب الحصينة ذات الخصائص الجيدة والأكثر استخداما ، وهو أحد طرق المعالجة يستخدم لمعالجة القيم المتطرفة في المتغيرات المستقلة والمتغير التابع وهو مقدر يجمع بين الكفاءة النسبية المحاذية العالية لمقدر M مع نقطة انهيار عالية لنوع معين من مقدرات S ، حيث أن مقدر MM يقوم بحساب تباين الأخطاء من مقدر ابتدائي ذو نقطة انهيار عالية الا وهو مقدر S-Estimation ، بينما أسلوب M الذي كان يستخدم المربعات الصغرى الموزونة يأخذ الاخطاء من مقدر ابتدائي وهو مقدر المربعات الصغرى الاعتيادية الذي تكون نقطة انهياره 0%(شاكر,2017).

وقد وصف (Yohai , 1987) ثلاث مراحل للحصول على مقدر MM:

1. استخدام مقدر ابتدائي ذو نقطة انهيار عالية مثل مقدر S ، إذ يرمز له بالرمز $\tilde{\beta}$ ومن ثم استخدام أخطاء هذا المقدر من خلال المعادلة

$$e_i(\tilde{\beta})=y_i-x_i'\tilde{\beta}$$

2. يتم أستخدام أخطاء المقدر $\tilde{\beta}$ والتي هي $S(e_1(\tilde{\beta}),...,e_n(\tilde{\beta}))$ويرمز لها بالرمز S_n . وأن دالة الهدف المستخدمة في هذه المرحلة يرمز لها ب ρ_0 .

3. يتم في هذه المرحلة ايجاد مقدر MM من خلال

$$j=1,...,k\ ,\ \sum_{i=1}^{n}x_{ij}\psi_1\left(\frac{y_1-x_i'\beta}{S_n}\right)=0$$

إذ أن $\psi_1(e)=\frac{\partial\rho_1(e)}{\partial e}$

كما أن دالة ρ_1 في المرحلة الثالثة يجب أن لا تكون نفس دالة ρ_0 ، ولكن يجب أن تحقق الشروط الثلاثة (Yohai , 1987) .

والصيغة أدناه تمثل مقدر معلمات النموذج MM الحصين التي تم الاعتماد عليها في العمل.

$$\hat{\beta}_{MM} = (X'W_{MM}X)^{-1}X'W_{MM}X \qquad (18)$$

d. طريقة مقدر M الحصين الموزونة (R.M.W):

تعتبر هذه الطريقة أحد طرق معالجة القيم الشاذة الموجودة في متجه الأخطاء الأ أنه في حالة وجود مثل هذه القيم في مصفوفة المتغيرات التوضيحية يصبح هذا الأسلوب غير قادر على معالجة هذه القيم والتخلص من أثرها ، لذا أقترح (شاكر،2009) أولاً تعديل القيم المتطرفة الموجودة في مصفوفة المتغيرات التوضيحية باستعمال مصفوفة الأوزان لطريقة المربعات الصغرى الموزونة weighted least squares (W.L.S) ومن ثم معالجة القيم المتطرفة الموجودة في متجه متغير الاستجابة من خلال استخدام (متجه أخطاء المربعات الصغرى الموزونة) باستعمال أسلوب M الحصين ثانياً، وأخيراً إيجاد المقدرات الجديدة بعد التعديل الأخير وهذه المقدرات سيطلق عليها اسم M الحصينة الموزونة Robust M-Weighted Estimator(R.M.W).

وبنفس طريقة المربعات الصغرى الموزونة الاعتيادية يمكن إيجاد M الحصين الموزون المقترح (شاكر،2009)،نحصل على المعادلة الأتية:

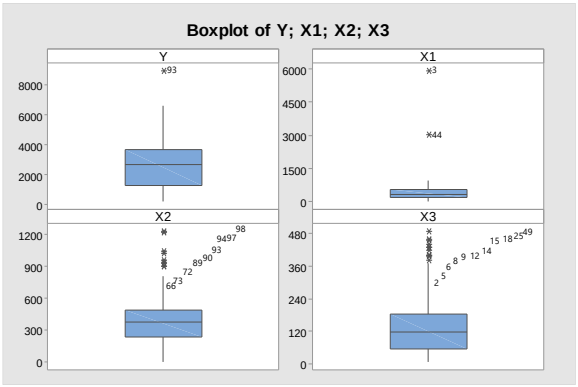
$$\therefore \hat{\beta}_{RMW} = (X'W_{RMW}X)^{-1}X'W_{RMW}Y \qquad (19)$$

المعادلة(19) تمثل صيغة أسلوب M الحصين الموزون المقترح (R.W.M) والذي يتم بواسطته معالجة التطرف الموجود في المتغيرات التوضيحية أو متغير الاستجابة أو كليهما معاً.

5. الجانب التطبيقي:

في هذا الجانب من البحث تم تطبيق ما ورد في الجانب النظري على بيانات الدراسة تتمثل عن تلوث مياه الابار والعناصر الداخلة في التركيب والتي تتمثل بمتغيرات X وعناصر اخرى تتمثل بالمتغير Y والتى تحمل اعلى نسبة من الشواذ (AL-Youzbakey and Sulaiman , 2020) ومن هذه العناصر التي تم اعتمادها في الدراسة كمتغيرات مستقلة (الكالسيوم (Ca^{2+}) والبيكربونات (HCO_3) والمغنيسيوم (Mg^{+})) ،والاملاح الذائبة (T.D.S) كمتغير معتمد. ومن خلال دراستنا استنتجنا بأن العلاقة بين كل متغير مستقل مع متغير معتمد علاقة طردية موجبة.

1. في الخطوة الاولى تم اختبار وجود القيم الشاذة في كلا المتغيرين (المتغيرات المستقلة والمتغير التابع) بأستعمال الرسم الصندوقي وهو أفضل طرق الكشف كما موضح في الشكل الاتي:



الشكل (1) يوضح القيم الشاذة في المتغرات المستقلة X والمتغير التابع Y

من خلال الشكل (1) تبين إن المتغير y ظهرت فيه قيمة شاذة واحدة أما في المتغير X_1 ظهرت فيه قيمتين شاذتين كذلك المتغير X_2 ظهرت فيه تسعة قيم شاذة والمتغير X_3 يحتوي احدى عشر قيمة شاذة .

2. الخطوة الثانية معالجة القيم الشاذة : يتم معالجة القيم الشاذة بعدة طرق منها

- طريقة الحذف حيث تبين أن افضل نموذج تم الحصول عليه عند حذف المشاهدتين (3)،(93).

نتائج النموذج المقدر بعد حذف المشاهدة رقم (3)(93):

$$\hat{Y} = 0.0185 + 0.4988x_1 + 0.3342x_2 + 0.6651x_3$$

$$R^2 = 79.88\% \quad MSE = 0.1851 \quad F = 134.98$$
$$VIF_{x_1} = 1{,}16; VIF_{x_2} = 1.12; VIF_{x_3} = 1.06 \quad D.W = 1.47435$$

حيث تم الحصول على أقل MSE عند حذف المشاهدتين (3)،(39) مقارنتةً بباقي النماذج .

• أما المعالجة بطريقة مقدر M الحصين : حيث تم اختيار نقطة انهيار مختلفة ولكل نقطة لها ثابت توليف معين (شاكر،2017) ، وتم التوصل الى افضل نموذج مقدر عند نقطة انهيار 50% كما موضح في الجدول الاتي:

جدول(1) :مقدر معلمات النموذج والجذر التربيعي لتباين الاخطاء ومعامل التحديد باستخدام مقدر Mالحصين

Obs	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	R^2	MSE	F
1	-0.0076	0.8951	0.2052	0.6272	0.9910	0.8730	2971.80

حيث ظهرت قيمة ال MSE=0.8730 وهي قيمة صغيرة مقارنة بباقي النماذج كذالك قيمة R^2 0.9910= وهي قيمة تفسيرية تبين أن متغيرات الانموذج (الكالسيوم ،والبيكربونات ، والمغنيسيوم) قد فسرت بنسبة (0.991) من أجمالي التغيرات الحاصلة بالمتغير التابع (T.D.S الأملاح الذائبة الكلية) .

• المعالجة بطريقة مقدر MM الحصين : تبين من هذه الطريقة التي تعالج مشكلة القيم الشاذة في المتغيرات المستقلة والمتغير التابع لتقليل أثر القيم الشاذة كما تم دراستها في الجانب النظري وفق المعادلة (18) حيث تم أختيار أنسب معالجة للقيم بناءا على الدراسات السابقة (شاكر،2017)عند نقطة أنهيار 50%وكفاءة قدرها 3.42 بمعامل توليف قدره 1.547 وكانت النتائج كما موضح في الشكل الاتي:

جدول(2): يوضح معاملات الانحدار ومتوسط مربعات الخطأ و معامل التحديد

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	R^2	F	MSE
0.0906	1.2398	0.1891	0.5353	0.9626	1295.69	1.5638

حيث تم الحصول على نموذج تكون فيه قيمة ال1.5638=MSE وقيمة أقل ما يمكن وقيمة معامل التحديد $R^2 = 0.9626$ وهي قيمة تفسيرية تبين أن متغيرات الثلاثة (الكالسيوم، البيكربونات ،المغنيسيوم) للنموذج قد فسرت ما نسبته (0.9626) من أجمالي التغيرات الحاصلة بالمتغير التابع (الاملاح الذائبةT.D.S).

• المعالجة بطريقة مقدر M الحصين الموزونة: في هذه الطريقة تم اختبار معلمات نموذج الانحدار M الموزونة استناداً إلى المعادلى (19) وفق أوزان وختبرات مختلفة فقد تم التوصل الى أفضل نموذج تم الحصول عليه عند أضافة الوزن $W_{RMW\,4}$ حيث كانت قيمة الـ $MSE = 0.022$ وهي قيمة صغيرة مقارنةً بباقي النماذج كما في الجدول الاتي:

جدول(3):معلمات النموذج بطريقة المقدر M الحصين الموزونة عند $\frac{4}{n-p}$

Dependent variable	Coefficients	Std.Error	T	p-value
Constant	0.048	0.049	0.981	0.329
$X_1(Ca^{2+})$	0.587	0.121	4.869	0.000
$X_2(HCO_3^-)$	0.347	0.047	7.368	0.000
$X_3(Mg^+)$	0.659	0.047	13.953	0.000

❖ **المقارنة بين طرق معالجة القيم الشاذة :**

بعد عدة طرق معالجة اجرية على بيانات التلوث تم التوصول الى اربع نماذج أنحدار يتم المقارنة بين هذه النماذج للحصول على افضل مقدر ذو كفاءة عالية تكون فيه قيمة الMSE أقل مايمكن كما هو مبين في الجدول الاتي:

جدول (4) : نتائج المقارنة بين طرق معالجة القيم الشاذة

	طريقة الحذف/(3+93)	الحصين /عند نقطة Mمقدر أنهيار (50%)	الحصين/عنده MMمقدر نقطة الانهيار وكفاءة (3.42)	الحصينة Mمقدر الموزونة/0.038
R-Squer	0.7988	0.9910	0.9528	0.9772
MSE	0.1851	0.8730	1.5638	0.022
F	134.98	2971.805	1295.696	117.646

من ملاحظة النتائج في الجدول (4) يمكننا القول بأن R.M.W قد عالج نسبة كبيرة من القيم الشاذة مقارنة بكل من طريقة الحذف ومقدر M الحصين ومقدر MM الحصين .

6. الأستنتاجات :

كل ما تم الحصول عليه من نتائج وأستنتاجات وكفاءة مقدر وفق معيار المقارنة حسب طبيعة البيانات المدروسة حالياً والمأخوذة من مركز بحوث السدود والموارد المائية. أن أفضل طريقة لكشف المشاهدات الشاذة في بيانات نموذج الانحدار المتعدد باستخدام الرسم الصندوقي Box-plot. وأن معالجة القيم الشاذة أدى الى تحسين كبير جداً في أداء نموذج الانحدار الخطي المتعدد بالتنبؤ بقيم المتغير التابع بشكلها العام وتم التوصل الى أفضل نموذج عند معالجة القيم الشاذة بطريقة الحذف عند حذف المشاهدتين (3)،(93) حيث كانت قيمة MSE أقل مايمكن عند هذه الحالتين ، أما بطريقة مقدر M الحصين ظهر أفضل نموذج عند نقطة أنهيار قدرها 50% أما بطريقة مقدرM الحصين الموزونة حيث ظهر النموذج المفضل عند $\frac{4}{n-p}$ وظهر تفوق لـMM الحصين بكفاءة 3.42 ونقطة أنهيار قدرها 50% وهذا كان واضحا لحصوله على أقل MSE ، كما تبين أن أفضل طريقة لمعالجة القيم الشاذة والتي اظهرت كفاءتها العالية عن باقي الطرق هي طريقة مقدرM الحصين الموزونة عند $\frac{4}{n-p}$ عن باقي المختبرات.

Reference

1. Shaker ,Saleh Muayad,(2009)," Improving the hippocampal M method in estimating multiple linear regression model parameters", Iraqi Journal of Statistical Science, No.16,Pp.219-242.
2. Shaker ,Saleh Muayad,(2017)," Proposed robust methods for the median analysis of the linear regression model and their comparison with the ordinary least squares estimators using simulation",PHD. Thesis, University of Mosul, Mosul, Iraq.
3. AL-Mutery, Abed Al-Aziz Mnahe, (2010)," Methods for discovering anomalous and affecting observations on linear regression",King Saud University, AL-Reyad.
4. AL-Saeg, Mumen Amer Hsan,(2013)," The effect of outliers on the results of some statistical hypotheses",BSc. Thesis, University of Mosul, Mosul ,Iraq.
5. Yusef, Isaam Al-deen Yusef Abd Alla,(2020)," The effect of outliers on the parameters of the multiple linear regression analysis model",PHD. Thesis, AL-Sudan University.
6. Al-Youzbakey, K.T. and Sulaiman, A.M. (2020). "Hydrochemical Evaluation for Al-Sada Area Wells and their Suitability for Agricultural Usages", Journal of Umm Al-Qura University for Applied Science, Dams and Water Resources Researches Center, University of Mosul, Mosul, Iraq.
7. Belsley,D. et al. (1980). "Regression Diagnostics: Identifying Infuential Data and sources of Collinearity", Wiley, New York, p:105.
8. Chatterjee, S. and Hadi, A. S.(1988)."Sensitivity Analysis in Linear Regression", New York: john Wiley.
9. Fox, John,(1997). "Applied Regression Analysis, Linear Models, and Related Methods", Sage publications.
10. Neter,J. et al. (1990). "Applied Linear statistical Models:Regression, Analysis of Variance, and Experimental Designs".(3rd edition). Irwin, Homewood, IL 60430, Boston, MA 02116.
11. Rousseew P.J. and Leroy, A.M. (1987). Robust Regression and Outlier Detection. Wiley-Interscience, New York.
12. Tukey, J.W.(1977). "Exploratory Data Analysis", Addison-Wesley reading, MA.
13. Yohai, V.J. (1987). "High breakdown-point and high efficiency estimates for regression", The Annals of Statistics 15, 642-65.

Detection Of Outliers In The Linear Regression Model With Application To Well Water Pollution Data On The Outskirts Of The City Of Mosul

Saja Marwan Esmaeel & Safwan Nathem Rashed
Department of Informatics & Statistic, College of Computer Science, & Mathematics University of Mosul, Mosul, Iraq

Abstract : The research idea is concerned with identifying the effect of outliers on the parameters of the multiple linear regression analysis model. Where the outliers values that are present in the data are detected and diagnosed if they are in the independent variables or the dependent variable, which causes an impact on the estimation of the parameters of the studied model. The extreme data types and methods of processing them have been identified to obtain a better model with high efficiency or reduce the impact of These values on the model,The MSE standard was developed for the purpose of comparing treatment methods and was applied to real data taken from the Dams and Water Resources Research Center, University of Mosul. Suggested by (Shaker,2009) is the best in detection among the methods that have been used .
Keyword: Outliers, Diagnostic Detection, Treatment, Elimination Method, Hippocampal M Estimator, Hippocampal MM Estimator and Weighted Hippocampal M Estimator