المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

توظيف أسلوب مقدرات MM الحصينة في تقدير نموذج انحدار المكونات الرئيسة: دراسة مقارنة

اسراء نجيب سعيد الصراف و بشار عبد العزيز الطالب

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

معلومات النشر

تاريخ المقالة:
تم استلامه في 16 شباط 2021
تم القبول في 11 اذار 2021
متاح على الإنترنت في 1 حزيران 2021

الكلمات الدالة:
انحدار المكونات الرئيسية، القيم الجاذبة، الطرائق الحصينة، مقدرات MM، المربعات الصغرى الموزونة.

المراسلة:
بشار عبد العزيز الطالب
basharaltalib@gmail.com

الخلاصة

يركز البحث على اقتراح استخدام مقدرات MM الحصينة في تقدير معلمات نموذج انحدار المكونات الرئيسية والذي يستخدم عادة في تقدير نموذج الانحدار عندما تكون المتغيرات التوضيحية غير مستقلة، حيث يعتبر مقدر MM الذي تم اقتراحه من قبل (Yohai, 1987) ذو نقطة انهيار (Breakdown) عالية حتى عن وجود نقاط جذب (Leverage Points)في البيانات ويعطي مقدرات ذات كفاءة جيدة، وقد سمي هذا المقدر بمقدر MM إشارة إلى حقيقة أن أكثر من مقدر M واحد يستخدم للحصول على المقدر النهائي حيث يتم التقدير باستخدام طريقة المربعات الصغرى المعاد وزنها (IRLS) ، وسيتم تطبيق النماذج التقليدية والمقترحة على بيانات تم الحصول عليها من معمل الاسمنت في منطقة بادوش وتتم على هذا الأساس المقارنة بين الطرق التقليدية المعروفة والطريقة المقترحة باستخدام مقدرات MM من خلال دراستين تجريبية باستخدام المحاكاة وتطبيق بيانات معمل الأسمنت.

DOI: https://doi.org/10.33899/iqjoss.1970.170002 , ©Authors, 2021, College of Computer & Mathematical Science, University of Mosul, Mosul.
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

1- المقدمة Introduction

يعتبر استخدام انحدار المكونات الرئيسية واحدة من طرق التقدير التي تعالج مشكلة تعدد العلاقة الخطية والتي تحدث في كثير من النماذج القياسية، وجوهر هذه الطريقة هو أنها تستبدل المتغيرات المستقلة التي تعاني من مشكلة التعدد الخطي بمكونات رئيسية تعبر عنها لتوفيق نموذج الانحدار، وطريقة المكونات الرئيسية واحدة من النماذج الخطية المتحيزة واسعة الاستخدام وتقوم على فكرة تحويل المتغيرات التفسيرية إلى متغيرات جديدة متعامدة دون حذف أي منها تسمى بالمكونات (المركبات) الرئيسية، وكل مكون رئيسي عبارة عن تركيب خطي في المتغيرات المستقلة الأصلية. وتعتبر عملية إيجاد المكونات الرئيسية خطوة مهمة لإزالة أثر التعدد الخطي تمهيداً لاستخدام طريقة المربعات الصغرى الاعتيادية لتقدير معالم نموذج الانحدار الخطي الأصلية للمتغيرات التفسيرية التي يستعاض عنها بالمكونات الرئيسية.(الشيخ،2014) وقد كان اهتمام الباحثين لإيجاد طرائق ذات كفاءة أعلى ولا تتأثر كثيراً بالقيم الشاذة أو بالانحراف عن الافتراضات كبدائل عن الطرق التقليدية سواء المعلمية واللامعلمية والتي تفقد كفاءتها في حالة جود قيم شاذة في بياناتها، وفي بعض الحالات يتحتم عدم حذف قيم المشاهدات الشاذة بل توظيف طرائق بديلة للتقدير تكون غير حساسة Insensitive أو حصينة تجاه وجود هذه القيم في العينة، أي أنها تستطيع التعامل معها آخذة بنظر الاعتبار طبيعة البيانات والعلاقة بين متغيراتها، وهذه الطرائق تسمى بالطرائق الحصينة (Robust Methods) أما المقدرات الناتجة عنها فتسمى بالمقدرات الحصينة (Robust Estimators). إذ أن الطرائق الحصينة تعطي أوزان أقل للمشاهدات الشاذة وذلك للتقليل من تأثيرها على عملية التقدير، ويتم غالباً استخدام الأسلوب التكراري Iteration في الحساب للتقليل من تأثير وجود الارتباط الذاتي والتعددية الخطية في البيانات. ومن أهم هذه الطرائق التي تم استخدامها في هذه الدراسة هو أسلوب MM الحصين (Huber , 1964). ومن أجل تحقيق الغرض من البحث تم استخدام برنامج R في كتابة أكواد التحليل، وقد تم اختبار

تحقيق البيانات للافتراضات الخاصة بتحليل المكونات الرئيسية أولاً بالبيانات الأصلية ومن ثم تم حساب المكونات الرئيسية وأُجريت بعدها المقارنة بين نتائج النموذج الأصلي والنماذج الأخرى فضلاً عن الأسلوب المقترح.

2- هدف البحث

يهدف هذا البحث إلى تحصين مقدرات نموذج الانحدار ضد وجود القيم الشاذة في البيانات سواءً في المتغير المعتمد أو في المتغيرات المستقلة، فضلاً عن التعامل مع وجود مشكلة تعدد العلاقة الخطية بين المتغيرات التوضيحية.

3- الجانب النظري

(3-1) مشكلة تعدد العلاقة الخطية Multi-Collinearity Problem

تظهر مُشكلة تعدُد العلاقة الخطية عندما يرتبط اثنان أو أكثر من المتغيرات التوضيحية (المستقلة) بعلاقة خطية قوية جداً بحيث يُصبح من الصعب فصل أثر كُل مُتغير عن المُتغير المُعتمد، وان مُصطلح تعدُد العلاقة الخطية هو مُصطلح مُركب من كلمة multi (مُتعدد) وco (مُشترك) وlinearity (خطي)، وتختلف درجة التعدُد الخطي تبعاً لقوة الارتباط بين المُتغيرات التوضيحية، فإذا كان مُحدد مصفوفة المعلومات مُساوياً للصفر $\left(\left|X'X\right|=0\right)$ فإننا نحصل على تعدد خطي تام (Perfect Multi-collinearity) وبهذا لا نستطيع تقدير معلمات الانحدار باستخدام طريقة المُربعات الصُغرى الاعتيادية، حيث أن تبايُن مُقدرات معلمات الانحدار ستكون قيمها غير مُعرفة (لأن قيمة المحدد في المقام مساوية للصفر)، أي أن (الخطيب ،2010):

$$\operatorname{var}(\underline{\hat{\beta}})=\sigma^2(X'X)^{-1}=\frac{adj(X'X)}{\left|X'X\right|}\sigma^2 \rightarrow \text{Undefined} \qquad (1)$$

أما إذا كان المحدد قريباً من الصفر فإننا نحصل على تعدُد خطي غير تام (ضعيف) (Weak Multicollinearity)، وعند استخدام طريقة المُربعات الصُغرى الاعتيادية لغرض تقدير معلمات الانحدار فإننا نحصل على مُقدرات غير دقيقة ما يؤدي إلى الحصول على تبايُن عالٍ (لأن المقام $|X'X|$ سيكون صغيراً)، وكما يلي:

$$\operatorname{var}(\underline{\hat{\beta}})=\frac{adj(X'X)}{\left|X'X\right|}\sigma^2 = \text{ Large value} \qquad (2)$$

وقد اهتم الباحثون بالتخلُص من مُشكلة تعدُد العلاقة الخطية أو التقليل من أثرها، وذلك باستخدام أساليب متعددة وطرائق مختلفة لمعالجة المُشكلة، ومن هذه الطرائق ما يأتي:

1- إضافة بيانات جديدة إلى البيانات الأصلية ، حيث أن زيادة عدد المُشاهدات يؤدي إلى التقليل من قيمة الأخطاء المعيارية، ويعود السبب في ذلك إلى الزيادة الحاصلة في قيمة مُحدد مصفوفة المعلومات.

2- استبعاد أحد المتغيرات ذي الارتباط المرتفع مع أحد المتغيرات التوضيحية الأخرى، مع مراعاة أن أي استبعاد للمتغيرات المهمة يؤدي إلى حصول مشكلة في توصيف النموذج.

3- استخدام المعلومات المُسبقة، إذ يتُم وضع معلومات مُسبقة حول معلمات الانحدار لغرض تقديرها، حيث تأخذ المعلومات المُسبقة صيغاً ثابتة وعشوائية ومُختلطة.

4- استخدام تحويلات على المُتغيرات التي لها ارتباط عالٍ وهناك نوعان من التحويلات القياسية وهما الأكثر شيوعاً: التحويل إلى المتغير الطبيعي(Normal variable form) والتحويل الثابت بطول واحد (Unit length scaling).

5- استخدام طرائق تقدير مُتحيزة، منها انحدار الحرف (Ridge Regression)، وانحدار المركبات الرئيسية (Principal Components Regression)، وانحدار الجذور الصماء (Latent Root Regression)، والمُقدر المُتقلص (Shrunken Estimator) أو إحدى الطرق غير المتحيزة مثل طريقة المُربعات الصُغرى المُقيدة (Restricted Least Square).

وهناك عدة معايير للكشف عن وجود التداخل الخطي منها ما هو عام في اختبار وجود المشكلة ومنها ما هو خاص يتناول كل متغير على حدى، ومن أهمها:

أ- المؤشرات العامة (للمتغيرات بالعموم) لوجود مشكلة تعدد العلاقة الخطية

أولاً: المحدد Determinant

يكون محدد المصفوفة $|X'X|$ مفرداً Singular اذا احتوى على أعمدة وصفوف مرتبطة خطياً. لذلك، فإن محدد مصفوفة الارتباط ($R=X'X$) كون المتغيرات محولة إلى الصيغة القياسية يشير إلى وجود مشكلة تعدد العلاقة الخطية. ومع ذلك، فإن المحدد لا يوفر معلومات حول الترابط بين معلمات الانحدار. والمحدد $X'X$ قيمته محصورة بين الصفر والواحد ($0\leq |X'X| \leq 1$) (Cooley & Lohnes, 1971). فإذا كانت قيمته تقترب من الصفر $|X'X|\sim 0$ فهذا يعني وجود تداخل خطي بين المتغيرات (Asteriou & Hall, 2007).

ثانياً: معامل التحديد R^2 Coefficient of Determination

يعرف معامل التحديد R^2 بأنه نسبة تأثير المتغيرات المستقلة X_j في التغيرات الحاصلة في المتغير التابع Y (Gujarati & Porter, 2008; Stock &Watson, 2010)، ومن ناحية اخرى كلما كانت قيمة معامل التحديد عالية زادت فرصة وجود تعدد علاقة خطية بين المتغيرات وصيغته الرياضية كالآتي:(Asteriou & Hall, 2007; Gujarati & Porter, 2008; Maddala, 1988)

$$R^2 = \frac{SSR}{SST} \qquad (3)$$

ثالثاً: Farrar χ^2:

إنه اختبار Chi-Square لاكتشاف قوة العلاقة الخطية بين مجموعة من المتغيرات المستقلة في نموذج الانحدار، ويتم حسابه وفقاً للصيغة الآتية:

$$\chi^2 = -[n-1-\frac{1}{6(2p+5)}] \times \log_e[X'X] \quad \sim\chi^2_{\frac{1}{2}p(p-1)} \qquad (4)$$

فإذا كان $\chi^2 > \chi^2_{\frac{1}{2}p(p-1)}$ فهذا يعني وجود مشكلة تعدد العلاقة الخطية (Farrar& Glauber,1967).

رابعاً: الدليل الشرطي Condition Index

يستخدم هذا المؤشر للكشف عن وجود التعدد الخطي ومن خلال القيم المميزة يتم حساب هذا المؤشر وفقاً للصيغة الآتية (المولى،2016):

$$CI_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}} \qquad , j = 1,2,\ldots,p \qquad (5)$$

حيث ان: λ_j : القيمة المميزة ذات التسلسل (j).

ويستفاد من الدليل الشرطي في قياس درجة تعدد العلاقة الخطية بالاعتماد على قيمة الجذر المميز j ، حيث اقترح كل من Belsley,Kuh Welsch& في عام 1980 أنه اذا كانت قيمة الدليل الشرطي C_j بحدود العشرة تكون العلاقة ضعيفة ، أما عندما ($30\leq CI_j\leq 100$) تكون العلاقة متوسطة أما إذا كانت ($CI_j > 100$) فهذا يعني إن العلاقة قوية.

خامساً: مجموع مقلوب الجذور المميزة Sum of reciprocal of eigenvalues

$$\sum_{j=1}^{p} \frac{1}{\lambda j} \quad > 5p \qquad (6)$$

اذا كان هذا المجموع أكبر من خمسة مرات عدد المتغيرات التفسيرية فهذا يعني وجود تداخل خطي بين المتغيرات التفسيرية ايضاً.(سالم،2018)

سادساً: Theil's indicator

اقترح (Theil, 1971) مقياساً للعلاقة الخطية بين المغيرات المستقلة بناءً على المساهمة التزايدية ($R^2 - R_j^2$) لمربع الارتباط المتعدد ، حيث أن R_j^2 هو عبارة عن معامل التحديد لانحدار المتغير X_j على بقية المتغيرات المستقلة في النموذج.

$$m = R^2 - \sum_{i=1}^{p}(R^2 - R_{-i}^2) \qquad (7)$$

فاذا كانت m=0 فعندها تكون كل المتغيرات التوضيحية غير مرتبطة لأن المساهمة المتزايدة تضاف جميعها إلى R^2 ، ومع ذلك إذا كانت m~1 فهذا يعني وجود مشكلة تعدد علاقة خطية بين المتغيرات.

سابعاً: Red indicator

قدم (Kovács et al., 2005) مؤشراً موحداً لتشخيص تعدد العلاقة الخطية باستخدام الجذور المميزة أو تحديد معدل الارتباط في البيانات، وصيغته الرياضية كالآتي:

$$\text{Red} = \frac{\sqrt{\sum_{j=1}^{p}(\lambda_j-1)^2}}{\frac{p}{\sqrt{p-1}}} \qquad (8)$$

اذا كانت قيمة المؤشر صفر (Red=0) يشير إلى عدم وجود تعدد علاقة خطية واذا كانت قيمته تقترب من الواحد (Red~1) يشير الى وجود تعدد علاقة خطية بين المتغيرات.

ب- المؤشرات الفردية (لكل متغير على حدى) لوجود مشكلة تعدد العلاقة الخطية

اولاً: معامل تضخم التباين Inflation Variance Factor (VIF)

تم اقتراح هذا المقياس عام 1967 من قبل Farrar&Glauber وقد أطلق عليه Marquardt عام 1970 بمعامل تضخم التباين Variance Inflation Factors ويرمز له بالرمز (VIF) ومعاملات تضخم التباين يمكن التعبير عنها بصيغة رياضية تأخذ الشكل الآتي(النعيمي،2005):

$$\text{VIF} = \frac{1}{1-R_j^2}\ , \quad j=1, 2,\ldots, m \qquad (9)$$

حيث أن: m: تمثل عدد المتغيرات التوضيحية ، R_j^2: تمثل معامل التحديد لانحدار المتغير X_j على بقية المتغيرات التوضيحية وإن $0 \le R_j^2 \le 1$ ، فاذا كان هناك ارتباط بين X_j مع بقية المتغيرات فهذا يؤدي الى ان قيمة R_j^2 سوف تقترب من الواحد $R_j^2 \approx 1$ مما يؤدي أن قيمة (VIF) تكون كبيرة جداً وهذا يؤشر على وجود مشكلة تعدد العلاقة الخطية بين المتغيرات التوضيحية ، أما إذا كان المتغير X_j مستقل تماماً عن بقية المتغيرات التنبؤية الاخرى فان (R^2=0) وبذلك تكون قيمة (VIF=1) ويعني هذا عدم وجود مشكلة تعدد العلاقة الخطية بين المتغيرات التنبؤية. أوضح Montgomery أنه إذا كانت قيمة $\text{VIF} \ge 10$ فانه يعد سبباً كافياً لإهمال المتغير X_j من التحليل.

ثانياً : TOL

يستخدم هذا المؤشر للكشف عن وجود مشكلة تعدد العلاقة الخطية وصيغته الرياضية هي:

$$TOL_j = \frac{1}{VIF_j} = 1 - R_j^2 \qquad (10)$$

فاذا كانت قيمة $TOL_j \sim 0$ فهذا يعني وجود تعدد علاقة خطية بين المتغيرات وكما وضح كل من (et al., 2004; Marquardt, 1970 Kutner).

ثالثاً: الجذور المميزة Eigenvalues

اقترح كل من الباحثين Kendall (1957) وSilvey (1969) استخدام قيم الجذور المميزة لمصفوفة الارتباط $X'X$ للتحقق من وجود تعدد علاقة خطية و حيث تشير قيم الجذور المميزة الصغيرة والقريبة من الصفر الى وجود علاقة قوية(Kendall, 1957; Silvey, 1969).

رابعاً: CVIF

هو مقياس تم اقتراحه من قبل كل من الباحثين Curto and Pinto (2011) للكشف عن وجود مشكلة تعدد العلاقة الخطية ولتقييم تأثير الارتباط بين معلمات الانحدار في تباين $OLSE_s$ ويتم حساب هذا المقياس وفقاً للصيغة الآتية:

$$CVIF_j = VIF_j \times \frac{1-R^2}{1-R_\circ^2} \qquad (11)$$

حيث أن: $R_\circ^2 = R_{yx_1}^2 + R_{yx_2}^2 + \cdots + R_{yx_p}^2$

فإذا كانت قيمة $CVIF_j \ge 10$ فهذا يعني وجود مشكلة تعدد العلاقة الخطية بين المتغيرات المستقلة (Curto&Pinto, 2011).

خامساً: Leamer's method

اقترح Leamer (Greene,2002) مقياساً لتأثير تعدد العلاقة الخطية للمتغير ذو التسلسل j على بقية المتغيرات وصيغته الرياضية هي :

$$C_j = \left\{ \frac{\left(\sum_i^n (X_{ij} - \bar{X}_j)^2\right)^{-1}}{(X'X)_{jj}^{-1}} \right\}^{\left(\frac{1}{2}\right)} \qquad (12)$$

علماً أن هذا المقياس هو بمثابة الجذر التربيعي لنسبة التباينات من المعلمات المقدرة $(\hat{\beta}_j)$ عند تقديرها بدون ومع معلمات الانحدار الأخرى، فإذا كان المتغير X_j غير مرتبط بالانحدارات الأخرى فإن C_j سيكون 1 وإلا فسيكون مساويًا لـ $(1 - R_j^2)^{\frac{1}{2}}$ ، واذا كانت $C_j \sim 0$ فهذا يعني وجود مشكلة تعدد العلاقة الخطية.

سادساً: F and R^2relation

يمكن استخدام العلاقة بين اختبار F وR^2 من انحدار X_j على الانحدارات المتبقية الاخرى لاكتشاف تعدد العلاقة الخطية، وتوصف العلاقة بأنها:

$$F_j = \frac{\frac{R^2_{x_j.x_1.....x_p}}{p-2}}{\frac{1-R^2_{x_j.x_1.....x_p}}{n-p+1}} \sim F(p-2.n-p+1) \qquad (13)$$

حيث $F^* = F_{p-2.n-p+1}$

فإذا كان $F_j > F^*$ فهذا يعني أن X_j تربطه علاقة خطية مع المتغيرات الاخرى ويجب إسقاطها من النموذج (Gujarati and Porter,2008).

سابعاً: Farrar W_j

هو اختبار F لتحديد الانحدارت التي تربطها علاقة خطية ببقية المتغيرات ويتم حساب احصاءة الاختبار من معاملات الارتباط المتعددة بين المتغيرات المستقلة وصيغته الرياضية هي:

$$w_j = \frac{R_j^2}{1-R_j^2}\left(\frac{n-p}{p-1}\right) \sim F(n-p.p-1) \qquad (14)$$

فإذا كانت$w_j > F(n-p.p-1)$ فهذا مؤشر على وجود مشكلة تعدد علاقة خطية كبيرة.(Farrar & Glauber, 1967) .

(2-3) الطرائق الحصينة Robust Methods

إن العديد من الباحثين توصلوا إلى أن الطرائق التقليدية للتعامل مع المشاكل تكون غير كفوءة في حالة عدم تحقق أحد الافتراضات أو الشروط التي تعتمد عليها. لذلك سعى الباحثون لإيجاد طرائق أكثر كفاءة ولا تتأثر كثيراً بالانحرافات عن الافتراضات المحددة، فكانت الطرائق الحصينة (Robust Methods) هي البديل الجيد وهي أقل تأثراً في حالة اختراق البيانات لشرط من شروط التحليل المستخدم، وتمتاز بأنها مناسبة لفئة واسعة من التوزيعات في تقدير معلمات النموذج الخطي. أما المقدرات الناتجة عن هذه الطريقة فتسمى بالمقدرات الحصينة وتكون غير حساسة تجاه الشواذ، وبالرغم من وجود طرق حصينة مختلفة إلا أن أغلبها تشترك بنقطتين رئيسيتين، الأولى هي بإعطاء وزن أقل للمشاهدة الشاذة (إن وجدت) وذلك للتقليل من تأثيرها، والنقطة الثانية استخدام أسلوب التكرار (Iteration) الذي يؤدي إلى تقليل تأثير الارتباط الذاتي والتعدد الخطي (Rousseeuw , 2005)، ومن أهم الطرائق الحصينة المستخدمة لإيجاد مقدرات الأنموذج هي:-

(2-3-1) مقدر MM الحصين MM-estimators

من المعلوم أن مقدرات M تصلح للتعامل مع مشكلة تقدير المعلمات للمتغير المعتمد والمتغيرات المستقلة لكن هذه المقدرات تكون غير حصينة (Non-Robust) ضد الحالات التي يكون فيها البعد x-direction X يحتوي على قيم شاذة (أي عندما توجد في مصفوفة التنبؤ قيم مخلة أو جاذبة Leverage Points). فيعتبر مقدر MM الذي تم اقتراحه من قبل (Yohai, 1987) ذو نقطة انهيار (Breakdown) عالية حتى عن وجود نقاط جذب (Leverage Points) في البيانات ويعطي مقدرات ذات كفاءة جيدة، علماً أن ثابت التوليف (Tunning Constatnt) يتم اختياره لتحقيق نقطة الانهيار والكفاءة المطلوبتين (Heritier et. al., 2009) و (Riazoshams, 2019). وقد سمي هذا المقدر بمقدرMM إشارة إلى حقيقة أن أكثر من مقدرM واحد يستخدم للحصول على المقدر النهائي حيث يتم

التقدير باستخدام طريقة المربعات الصغرى المعاد وزنها (IRLS) (Andersen, 2008). ويقوم مقدر MM على تقدير معلمات الانحدار باستخدام تقدير S (S-estimation) الذي يقوم بتصغير تباين الأخطاء (البواقي) من مقدرات M (M-estimation) ومن ثم إعادة تطبيق مقدر M مرة ثانية. حيث يتم تقدير المعلمات لمقدر MM كحل لدالة التصغير التالية (Almetwally and Almongy, 2008) و (Susanti et. al., 2014):

$$\sum_{i=1}^{n} \rho_1^*(u_i) X_{ij} = 0 \qquad (15)$$

or

$$\sum_{i=1}^{n} \rho_1^* \left(\frac{Y_i - \sum_{j=0}^{k} X_{ij} \hat{\beta}_j}{sMM} \right) X_{ij} = 0 \qquad (16)$$

حيث أن sMM يمثل الانحراف المعياري الذي يتم الحصول عليه من تقدير S وأن ρ هي دالة Tukey's biweight:

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2C^2} + \frac{u_i^6}{6C^2} & -C \le u_i \le C \\ \frac{C^2}{6}. & u_i < -C \quad or \quad u_i > C \end{cases} \qquad (17)$$

حيث أن

$$u_i = \frac{e_i}{\hat{\sigma}_i} \qquad (18)$$

ويقوم استخدام هذا المقدر من خلال الموازنة بين الكفاءة عند التوزيع الطبيعي وبين التحيز عندما تكون البيانات ملوثة بالقيم الشاذة، حيث أنه كلما زادت الكفاءة كلما زاد التحيز (Maronna et. al., 2019).

(2-3-2) بعض مقاييس التشتت الحصينة Some Robust Measures of Scale

ولأجل المقارنة بين أداء كل طريقة من الطرق المستخدمة في حالات عدم وجود شواذ وحالتي وجود قيم شاذة في المتغيرين المعتمد والمتغيرات المستقلة تم استخدام العديد من معايير المقارنة الشائعة ومنها الخطأ القياسي لقيم البواقي Residual standard error، معامل التحديد R-Square، متوسط مربعات الخطأ MSE، جذر متوسط مربعات الخطأ RMSE، وسيط نسبة الأخطاء المطلقة Median MDAPE Absolute Percentage Error، ووسيط الأخطاء المطلقة MDAE وهي من المعايير المعروفة في مقارنة الكفاءة ودقة التقدير بين النماذج المختلفة (Memmedli and Ozdemir , 2009) , (Boiroju and Reddy, 2012), (Willmott and Matsuura, 2005), (Sarwar and Sharma, 2014), (Woschnagg and Cipan, 2004), (Makridakis and Hibon, 1995)

سنتناول في هذه الفقرة بعض مقاييس التشتت الحصينة التي سيتم استخدامها في التحاليل الإحصائية وهي كما يلي:

1- الخطأ القياسي لقيم البواقي Residual standard error

هو الجذر التربيعي لحاصل قسمة مجموع مربعات البواقي على درجات الحرية (جذر متوسط مربعات الخطأ)، وكلما صغرت قيمة هذا المعيار دل ذلك على كفاءة النموذج المقدر. أي أنه مقياس يستخدم لتقييم مدى ملائمة نموذج الانحدار الخطي للبيانات، وصيغته الرياضية هي:

$$\text{RSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{df}} \qquad (19)$$

2- متوسط مربعات الخطأ Mean Square Error (MSE)

متوسط مربعات الخطأ العشوائي هو مقياس للدقة يتم حسابه بتربيع الخطأ لكل مشاهدة في مجموعة البيانات، ومن ثم إيجاد المعدل أو متوسط القيم لمجموع هذه المربعات، حيث أن الأخطاء يتم تربيعها قبل أخذ مجموعها ويمكن توضيح هذا المعيار كما يأتي:-

$$\text{MSE} = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} \qquad (20)$$

وباستخدام المصفوفات:

$$\text{MSE} = \frac{e'e}{n-p-1} \qquad (21)$$

$$e'e = y'y - \hat{\beta} x' y$$

$y'y$: يمثل حاصل ضرب مصفوفة قيم المتغير التابع ومدورها.

$\hat{\beta}$: يمثل مصفوفة معلمات الانحدار المقدرة.

$x'y$: يمثل حاصل ضرب مدور مصفوفة قيم المتغيرات المستقلة ومصفوفة قيم المتغير التابع.

n: حجم العينة ، p: عدد متغيرات الاستجابة.

3-جذر متوسط مربعات الخطأ Root Mean Square Error (RMSE or RSE)

هو مقياس يعبر عن الجذر التربيعي لمقياس آخر وهو متوسط مربعات الخطأ ((Mean Square Error (MSE)، وتكون صيغته الرياضية كما يلي:

$$\text{RMSE}= \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \qquad (22)$$

ولغرض المقارنة بين النماذج، من العادة استخدام مقياس جذر متوسط مربعات الخطأ(RMSE) بدلاً من متوسط مربعات الخطأ(MSE)، وذلك لأن جذر متوسط مربعات الخطأ قد تم قياسه بالوحدات الأصلية للقيم نفسها. ويعد هذا المقياس أكثر حساسية من المقاييس الأخرى لقيم البواقي الكبيرة، حيث إن عملية تربيع القيم تعطي أوزاناً غير متكافئة(غير متجانسة) لقيم البواقي الكبيرة. أما إذا كانت قيمة البواقي الكبيرة لا تشكل مشكلة في اتخاذ قرارك (مثلاً عندما تكون الكلفة الحقيقية للخطأ تتناسب تقريباً مع حجم الخطأ، وليس مع مربع قيمته) فعندها قد يكون من المناسب استخدام MAE أو MAPE كمعايير للمقارنة. إن جذر متوسط مربعات الخطأ(RMSE) ومتوسط الأخطاء المطلقة (MAE) يستخدمان فقط للمقارنة بين النماذج التي تقاس قيم البواقي فيها بنفس وحدات القياس. ومن المعلوم أن متوسط مربعات الخطأ(MSE) ومتوسط الأخطاء المطلقة(MAE) ربما يعالجان مشكلة وجود بواقي موجبة وسالبة كما هو معروف، وهما يعدان بديلين لمحدودية استخدام متوسط الأخطاء (Mean Error ME)، ولكنهما يفشلان في تزويدنا بمعلومات عن الدقة التنبؤية المتعلقة بوحدات القيم التي تحسب لها(العبيدي،2015).

4- وسيط الأخطاء المطلقة Median Absolute Error (MDAE)

هو متوسط الفرق المطلق بين متجهين رقميين، وصيغته الرياضية هي:(Alexel,2019)

$$\text{MDAE} = \text{median}\left(|e_j|\right) \qquad (23)$$

$$e_j = y_j - \hat{y}_j$$

حيث أن: y_j : القيم الفعلية $\hat{y}_j$: القيم المتوقعة

5- وسيط نسبة الأخطاء المطلقة Median Absolute Percentage Error (MDAPE)

هو وسيط النسبة المئوية لحاصل قسمة قيم البواقي الى قيم المشاهدات الحقيقية للمتغير المعتمد بعد ترتيب القيم المطلقة للأخطاء، وصيغته الرياضية هي: (Alexel,2019)

$$\text{MDAPE} = median\left(\frac{|e_j|}{|y_j|}\right) \times 100 \qquad (24)$$

4- الجانب التجريبي

إن تحليل المكونات أو المركبات الرئيسية هو الأساس في أسلوب انحدار المكونات الرئيسية حيث أنه في تحليل المكونات الرئيسية الاعتيادي يتم إيجاد المكونات أو المركبات ثم إيجاد قيم التحميلات والجذور المميزة أما انحدار المكونات الرئيسية فتكون فيه المتغيرات المستقلة على شكل مكونات رئيسية نقوم بدراسة تأثيرها على المتغير المعتمد، وعليه نحتاج إلى انحدار المكونات الرئيسية عندما يكون لدينا مشكلة تعدد العلاقة الخطية بين المتغيرات المستقلة، وهنالك بعض الحالات يكون فيها المتغير المعتمد أو المتغيرات المستقلة ملوثة بقيم شاذة فتسبب خفض في كفاءة النموذج المقدر وعندها تتداخل أهمية دمج انحدار المكونات الرئيسية مع الاساليب الحصينة للحصول على مقدرات كفوءة،

وقد تم في هذا البحث استخدام أوزان حصينة على نموذج مقدر بطريقة المربعات الصغرى متغيراتها المستقلة عبارة عن المكونات الرئيسية للنموذج الأصلي. ولمقارنة كفاءة الطرق تمت تجربة نماذج بثلاثة وخمسة وتسعة متغيرات على التوالي وبأحجام عينات 50 و 100 و 200 مشاهدة على التوالي في حالة عدم وجود شواذ في البيانات وأيضاً في حالات وجود 5% أو 6% و10% و 20% و 30% و 40% شواذ في البيانات في المتغير المعتمد Y وتم تطبيق النموذج المقترح لانحدار المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين بدوال أوزان ومنها Huber و Hampel و Bisquare على بيانات تحتوي على قيم جاذبة كي نتمكن من تمييز الطريقة الأكثر كفاءة مقارنة ببقية الطرق وذلك في حالات وجود 5% أو 6% و10% و 20% و 30% و 40% شواذ في البيانات وفي المتغيرات المستقلة وكما هو مبين في لاحقاً وتم تكرار كل التجارب (Replicates = 2970) مرة لغرض الحصول على نتائج متسقة، ولإعطاء صورة شاملة عن كفاءة الطرق. ولأجل المقارنة بين أداء كل طريقة من الطرق المستخدمة في حالات عدم وجود شواذ وحالتي وجود قيم شاذة في المتغيرين المعتمد والمتغيرات المستقلة تم استخدام العديد من معايير المقارنة الشائعة ومنها الخطأ القياسي لقيم البواقي Residual standard error، معامل التحديد R-Square، متوسط مربعات الخطأ MSE، جذر متوسط مربعات الخطأ RMSE، وسيط نسبة الأخطاء النسبية MDAPE Median Absolute Percentage Error، ووسيط الأخطاء المطلقة MADE وهي من المعايير المعروفة في مقارنة الكفاءة ودقة التقدير بين النماذج المختلفة (Memmedli and Ozdemir , 2009) , (Boiroju and Reddy, 2012), (Willmott and Matsuura, 2005), (Sarwar and Sharma, 2014), (Woschnagg and Cipan, 2004), (Makridakis and Hibon, 1995)

جدول (1): مقارنة كفاءة الطرق لنتائج المحاكاة لثلاثة متغيرات ولعينة بحجم 50 مشاهدة

	عند وجود 20% شواذ في المتغير المعتمد Y وعينة بحجم n = 50 وثلاثة متغيرات مستقلة m =3					
معيار المقارنة / الطريقة	Residual standard error	R-Square	MSE	RMSE	MDAE	MDAPE
Least Squares	1.2005474	0.09284980	1.326009	1.151525	0.6587	90.5583
LSPCRRobHuber	1.0285443	0.11606436	1.326733	1.151839	0.6786	90.7362
LSPCRRobHampel	1.1753073	0.09817641	1.326073	1.151552	0.6604	90.4818
LSPCRRobBisquare	1.0119709	0.11786001	1.326249	1.151629	0.6602	91.3032
LSRobMM	0.7204281	0.17159683	0.11757054	1.162424	0.6246	95.6406
التفسير	نلاحظ أنه وبوجود 20% شواذ في المتغير المعتمد أن مقدر LSRobMM قد حقق أفضل إنجاز وذلك من خلال اعلى قيمة لمعامل التحديد واقلها لمقاييس الكفاءة					

جدول(2) مقارنة كفاءة الطرق لنتائج المحاكاة لخمسة متغي ارت ولعينة بحجم 100مشاهدة

	عند وجود 5% شواذ في المتغير المستقل X عندما n = 100 و m =5 X-Leverage Points					
معيار المقارنة / الطريقة	Residual standard error	R-Square	MSE	RMSE	MDAE	MDAPE
Least Squares	2.9896447	0.02065423	8.401697	2.898568	0.8126	101.7935
LSPCRRobHuber	1.3464216	0.04332422	8.456752	2.908050	0.7217	98.5667
LSPCRRobHampel	1.0643818	0.06543584	8.490451	2.913838	0.7421	98.0982
LSPCRRobBisquare	0.8787052	0.07438228	8.487907	2.913401	0.7082	97.4617
LSRobMM	0.6945179	0.13751009	8.571837	2.927770	0.6400	98.9995
التفسير	نلاحظ أنه وبوجود 5% شواذ في المتغير المستقل X أن مقدر LSRobMM قد حقق أفضل إنجاز من خلال اعلى قيمة لمعامل التحديد واقلها لمقاييس الكفاءة (علماً ان اللون الرصاصي الفاتح لافضل قيمة لمقاييس الكفاءة واعلى قيمة لمعامل التحديد) والرصاصي الغامق للقيمة التي تليها في الأداء.					

جدول (3): مقارنة كفاءة الطرق لنتائج المحاكاة لتسعة متغيرات ولعينة بحجم 200

	عند وجود 10% شواذ في المتغير المعتمد Y وعينة بحجم n = 200 وتسعة متغيرات مستقلة m =9					
معيار المقارنة / الطريقة	Residual standard error	R-Square	MSE	RMSE	MDAE	MDAPE
Least Squares	1.0332342	0.03079962	1.014194	1.007072	0.6722	98.6964
LSPCRRobHuber	0.9368468	0.03382002	1.014792	1.007369	0.6747	98.5510
LSPCRRobHampel	1.0125223	0.03115766	1.014308	1.007128	0.6664	98.8208
LSPCRRobBisquare	0.9007909	0.03503423	1.014806	1.007376	0.6676	98.4796
LSRobMM	0.7087010	0.04996901	1.019959	1.009930	0.6265	98.1166
التفسير	نلاحظ أنه وبوجود 10% شواذ في المتغير المعتمد أن مقدر LSRobMM قد حقق أفضل إنجاز من خلال اعلى قيمة لمعامل التحديد واقلها لمقاييس الكفاءة					

جدول (4): تفسير النتائج لكل تجارب المحاكاة التي بلغت 2970 محاولة

		نسب التلويث / n	0%	5% or 6%	10%	20%	30%	40%	الخلاصه (الغالبيه)
			عندما يوجد في النموذج ثلاثه متغيرات مستقله m = 3						
Contamination with Outliers	Y-outliers	50	لنموذج بثلاثة متغيرات توضيحية ولكل أحجام العينات وكل نسب التلويث للمتغير المعتمد وكذلك للنموذج بدون شواذ كانت طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين قد حققت أفضل النتائج مقارنة بالطرق الأخرى.						LSRobMM
		100							
		200							
	leverage	50	لنموذج بثلاثة متغيرات توضيحية ولكل أحجام العينات وكل نسب التلويث للمتغيرات المستقله أن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين قد حققت افضل النتائج مقارنة بالطرق الاخرى.						LSRobMM
		100							
		200							
			عندما يوجد في النموذج خمسة متغيرات مستقله m = 5						
Contamination with Outliers and Leverage Points	Y-outliers	50	لنموذج بخمسة متغيرات توضيحية ولحجم عينة (50,100) وكل نسب التلويث للمتغير المعتمد وكذلك للنموذج بدون شواذ أن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين قد حققت أفضل النتائج مقارنة بالطرق الأخرى أما في حالة حجم عينة (200) أن طريقة المربعات الصغرى الاعتيادية قد حققت أفضل النتائج وتليها طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين.						LSRobMM
		100							
		200							Least Squares LSRobMM
	X-leverage Points	50	لنموذج بخمسة متغيرات توضيحية ولحجم عينة (50,100) وكل نسب التلويث للمتغيرات المستقلة أن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين قد حققت أفضل النتائج مقارنة بالطرق الأخرى أما في حالة حجم عينة (200) أن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين قد حققت أفضل النتائج وتليها طريقة المربعات الصغرى الاعتيادية.						LSRobMM
		100							
		200							LSRobMM Least Squares
			عندما يوجد في النموذج تسعه متغيرات مستقله m = 9						
Contamination with Outliers and Leverage Points	Y-outliers	50	لنموذج بتسعة متغيرات توضيحية ولكل أحجام العينات وكل نسب التلويث للمتغير المعتمد وكذلك للنموذج بدون شواذ أن طريقة المربعات الصغرى للمركبات الرئيسية والموزونة بمقدر MM الحصين قد حققت افضل النتائج مقارنة بالطرق الاخرى.						LSRobMM
		100							
		200							
	X-leverage Points	50	لنموذج بتسعة متغيرات توضيحية ولحجم عينة (50,200) وكل نسب التلويث للمتغيرات المستقلة أن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين قد حققت أفضل النتائج مقارنة بالطرق الأخرى أما في حالة حجم عينة (100) أن طريقة المربعات الصغرى الاعتيادية قد حققت أفضل النتائج وتليها طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بدالة Bisquare.						LSRobMM
		100							Least Squares LSPCRRobBisquare
		200							LSRobMM

وعليه فإن خلاصة تجارب المحاكاة ولـ 2970 محاولة قررنا بأن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين هي الأفضل أي أن نموذج انحدار المكونات الرئيسية بالطرق الحصينة قد تفوق على كل الطرق الأخرى.

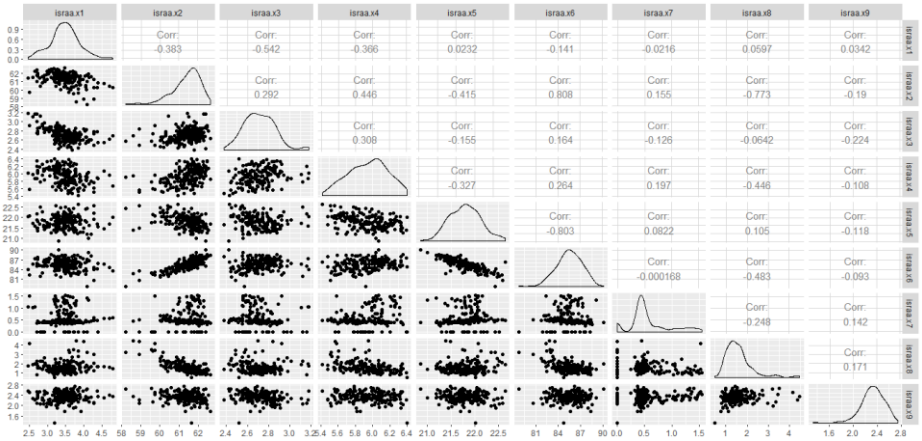
5- الجانب التطبيقي Application Part

لتطبيق الطريقة المقترحة تم جمع بيانات من معمل إسمنت بادوش للفترة من 2008- 2014، مع استبعاد أشهر الصيانة التي توقف المعمل فيها عن الانتاج. يتكون الاسمنت من بعض المواد الأساسية المتوافرة بصورة طبيعية من الحجر والرمل والحصى وبعض الإضافات الاخرى أثناء عملية التصنيع والتي تتضمن مواد تعمل على التغلب على بعض المشاكل الفنية ومواد لزيادة بياض الاسمنت. وقد تم تحديد المتغيرات الآتية:-

X_1 : تمثل أوكسيد المغنيسيوم Mgo، X_2 : تمثل أوكسيد الكالسيوم Cao، X_3 : تمثل أوكسيد الحديديك Fe2o3، X_4 : تمثل أوكسيد الالمنيوم Al2o3، X_5 : تمثل ثنائي أوكسيد السيليكون Sio2، X_6 : تمثل معامل الإشباع الجيري L.S.F، X_7 : تمثل مواد غير قابلة للذوبان In.R، X_8 : تمثل الفقدان بالحرق L.O. ، X_9 : تمثل ثالث أوكسيد الكبريت Sio3 ، **Y**: تمثل تمدد الإسمنت Autoclave.

(5-1) اختبار وجود مشكلة تعدد العلاقة الخطية

تم في البدأ التأكد من وجود مشكلة تعدد العلاقة الخطية بين المتغيرات المستقلة لبيانات معمل السمنت، وتم البدأ بمصفوفة الارتباط ورسوم الانتشار بين المتغيرات المستقلة وكما موضح في الشكل (1) أدناه.



الشكل (1): شكل يمثل مصفوفة الارتباط

نلاحظ من الشكل (1) أعلاه الذي يبين رسم مصفوفة الارتباط أن هنالك علاقة قوية للمتغير X_2 (أوكسيد الكالسيوم Cao) مع المتغيرين X_6 (معامل الإشباع الجيري L.S.F) و X_8 (الفقدان بالحرق L.O.)، وكذلك بين X_5 (ثنائي أوكسيد السيليكون Sio2) و X_6 وربما ينتج عن ذلك حصول مشكلة تعدد العلاقة الخطية. والجدول (5) أدناه يمثل المؤشرات العامة لوجود مشكلة تعدد العلاقة الخطية.

جدول(5): المؤشرات العامة لوجود مشكلة تعدد العلاقة الخطية

	MC Results	Detection
Determinant |X'X|:	0.0014	1
Farrar Chi-Square:	1204.2115	1
Red Indicator:	0.3350	0
Sum of Lambda Inverse:	82.8835	1
Theil's Method:	5.2734	1
Condition Number:	1721.5960	1

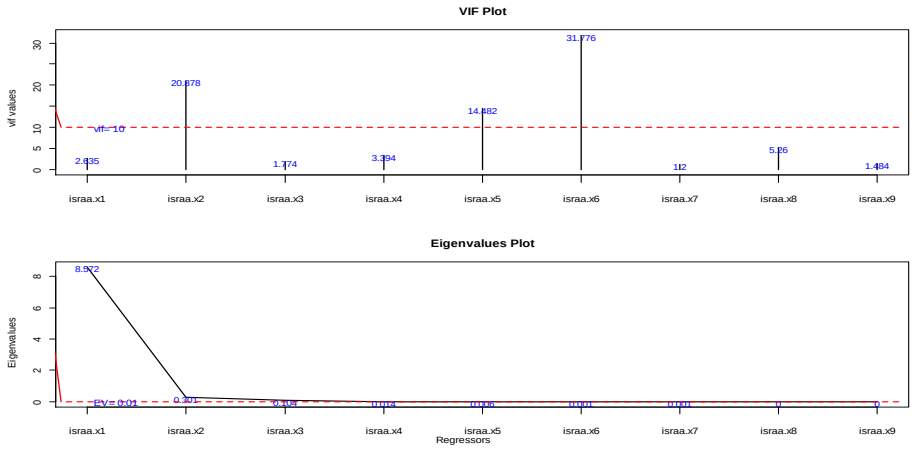
1 <-- COLLINEARITY is detected
0 <-- COLLINEARITY in not detected by the test

ونلاحظ من الجدول (5) أعلاه أن قيمة محدد مصفوفة المعلومات صغير جداً، وقريب من الصفر (0.0014) وكذلك فإن قيمة إحصاءة مربع كاي لفارار وكلاوبر (Farrar and Glauber) كبيرة جداً، كما أن قيمة العدد الشرطي Condition Number كبيرة جداً، كذلك فإن معيار Red Indicator غير مساوية للصفر، وأخيراً فإن قيمة مؤشر Theil أكبر من الواحد الصحيح بكثير. كل هذه الأمور تدل على وجود مشكلة تعدد العلاقة الخطية.

جدول (6): المؤشرات الفردية لوجود مشكلة تعدد العلاقة الخطية

	VIF	TOL	Wi	Fi	Leamer	CVIF
X1 : يمثل أوكسيد المغنيسيوم Mgo	2.6350	0.3795	36.3781	41.8085	0.6160	2.6188
X2 : يمثل أوكسيد الكالسيوم Cao	20.8781	0.0479	442.2871	508.3107	0.2189	20.7502
X3 : يمثل أوكسيد الحديديك Fe2o3	1.7740	0.5637	17.2204	19.7910	0.7508	1.7631
X4 : يمثل أوكسيد الألمنيوم Al2o3	3.3940	0.2946	53.2676	61.2192	0.5428	3.3733
X5 : يمثل ثنائي أوكسيد السيليكون Sio2	14.4822	0.0691	299.9799	344.7602	0.2628	14.3936
X6 : يمثل معامل الإشباع الجيري L.S.F	31.7765	0.0315	684.7761	786.9979	0.1774	31.5819
X7 : يمثل مواد غير قابلة للذوبان In.R	1.2002	0.8332	4.4552	5.1202	0.9128	1.1929
X8 : يمثل الفقدان بالحرق L.O.	5.2600	0.1901	94.7848	108.9340	0.4360	5.2278
X9 : يمثل ثالث أوكسيد الكبريت Sio3	1.4836	0.6740	10.7594	12.3656	0.8210	1.4745

ونلاحظ من الجدول (6) أعلاه وكما بين الشكل (1) السابق أن هنالك علاقة قوية للمتغير X_2 (أوكسيد الكالسيوم Cao) مع المتغيرين X_6 (معامل الإشباع الجيري L.S.F) و X_8 (الفقدان بالحرق L.O.)، وكذلك بين X_5 (ثنائي أوكسيد السيليكون Sio2) و X_6. هذا يظهر من خلال قيم معامل تضخم التباين التي زادت عن 10 وبالمقابل قيم ال TOL الصغيرة وكبر قيم Wi و Fi الكبيرة وكذلك قيم إحصاءة Leamer الصغيرة وكذا قيم CVIF والتي زادت عن العشرة وكل هذه أدلة على وجود مشكلة تعدد العلاقة الخطية، وقد تناول العديد من الباحثين مؤشرات الكشف عن وجود مشكلة تعدد العلاقة الخطية منهم (Asteriou and Hall, 2007), (Gujarati and Porter, 2009), (Farrar and Glauber, 1967), (Belsley et. al., 2004), (Chatterjee and Hadi, 2012), (Maddala, 1992), (Kovács et. al., 2005), (Kutner et. al., 2004), (Marquardt, 1970), (Curto and Pinto, 2011), (Greene, 2002), (Imdadullah et. al., 2016).



الشكل (2): شكل يبين رسم معامل تضخم التباين والجذور المميزة

والشكل (2) أعلاه يبين ما أكدته المؤشرات العددية ويتضح من رسمي VIF ورسم الجذور المميزة بأن هنالك ثلاثة متغيرات تعاني من مشكلة تعدد العلاقة الخطية.

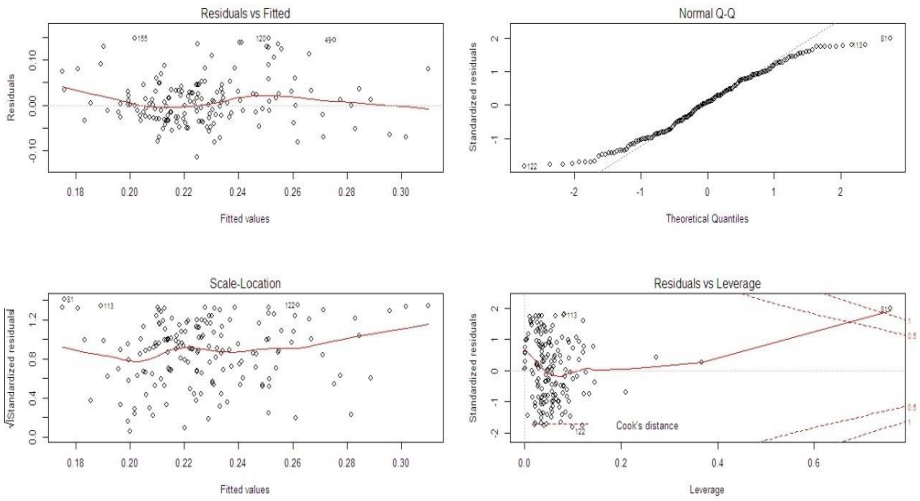
(2-5) انحدار المكونات الرئيسية

بعد التأكد من وجود مشكلة تعدد العلاقة الخطية تمت مقارنة أداء الطرق المقترحة على بيانات معمل الإسمنت بعد تحويل المتغيرات إلى الصيغة القياسية (لاختلاف وحدات قياسها) ومن ثم مقارنة كفاءة الطرق المقترحة في تقدير نموذج الانحدار لتمدد الإسمنت على متغيرات الدراسة التسعة. وقد كانت مقاييس الكفاءة كما هو مبين في الجدول أدناه والتي يتضح منها تفوق طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بمقدر MM الحصين.

جدول (7): مقاييس الكفاءة

	Residual standard error	R-Square	MSE	RMSE	MADE	MADPE
Least Squares	0.1063100	0.08037254	0.010697	0.103428	0.04302	20.2585
LSPCRRobHuber	0.1063100	0.08037254	0.010697	0.103428	0.04302	20.2585
LSPCRRobHampel	0.1063100	0.08037254	0.010697	0.103428	0.04302	20.2585
LSPCRRobBisquare	0.1055723	0.08099154	0.010698	0.103429	0.04272	20.5545
LSRobMM	**0.03301893**	**0.32952152**	**0.012328**	**0.111030**	**0.0333**	**15.5306**

ونلاحظ من رسم البواقي ضد القيم المقدرة في الشكل (3) أدناه لنموذج انحدار المتغير المعتمد y (تمدد الإسمنت) الموزون بمقدر MM (الرسوم للدوال الأخرى مطابقة لها) ضد المكونات الرئيسية المقابلة للمتغيرات المستقلة أنه لا توجد علامة لوجود ارتباط بين البواقي والقيم المقدرة ولا يوحي الشكل بوجود علاقة غير خطية. أما بالنسبة لرسم QQ-Plot نلاحظ وجود حوالي ثلاثة قيم شاذة (122, 113, 81) حيث نلاحظ انحراف النموذج عن التوزيع الطبيعي كما هو واضح في أطراف الرسم. وبالنظر إلى رسم جذر الأخطاء القياسية ضد القيم المقدرة أن النقاط متوزعة حول الخط بشكل منتظم الأمر الذي يدل على عدم وجود مشكلة عدم تجانس التباين بين الأخطاء. وأخيراً وبالنظر إلى رسم قيم الجذب Leverage (مسافات كوك Cook's Distance) ضد الأخطاء القياسية نجد أن الرسم قد أفرز قيمتين شاذتين في المتغير المعتمد (122, 113) وهما نفس القيمتين اللتين تم تشخيصهما في رسم QQ-Plot، ولا يبدو من الرسم وجود أية قيم جاذبة (Leverage Points) في البيانات.



الشكل (3): الرسوم التشخيصية لمشاكل نموذج الانحدار الخطي (LSRobMM)

6- الاستنتاجات:

على ضوء ما تم التوصل اليه في الجانبين التجريبي والتطبيقي يمكن استخلاص أهم الاستنتاجات الخاصة بالدراسة وكالاتي:

1- بقي الأسلوب المقترح لطريقة المربعات الصغرى للمكونات الرئيسية والموزونة بمقدر MM الحصين متفوقاً في حالة وجود مشكلتي الشواذ في المتغير المعتمد Y-outliers وتعدد العلاقة الخطية في البيانات في آنٍ واحد ما يدل على حصانته بالإتجاهين وبوجود تعدد العلاقة الخطية.

2- في حالة استخدام اسلوب المحاكاة ولكل النماذج وكل نسب التلويث وكل أحجام العينات، أظهر أسلوب MM الحصين المقترح (LSRobMM) تفوقاً واضحاً على نموذج المربعات الصغرى الاعتيادية، والنموذج الموزون بالدوال (Huber,Hampel,Bisquare) كون مقدر MM حصين ضد الشواذ في كل من المتغير المعتمد والمتغيرات المستقلة.

3- من خلال مناقشاتنا نستنتج أنه لـ 2970 محاولة مختلفة لسيناريوهات تجارب المحاكاة وتلويث متغير الاستجابة y، سجلت الطريقة المقترحة LSRobMM تفوقاً واضحاً على باقي الطرق وذلك من خلال الانخفاض الواضح في قيم معايير المقارنة (Least Squares, LSPCRRobHuber, LSPCRRobHampel, LSPCRRobBisquare)، كما أظهر اسلوب MM الحصين المقترح أفضلية على أسلوب المربعات الصغرى الاعتيادية في جميع أحجام العينة ونسب التلويث، وكذلك أظهر أفضلية على المقدرات الموزونة دوال الوزن (Huber،Hampel،Bisquare).

Reference

1. Al-Khatib, Ahmed Nazih Abdullah, (2010AD), "S**tudying the probability distribution of the random character factor by providing random prior information**," Master's thesis, College of Computer Science and Mathematics, University of Mosul- Iraq.
2. Al-Sheikh, Sawss, (2014 AD), "**Using the regression of the main components in the case of the existence of the problem of polylinearity, an applied study on the determinants of inflation in Algeria**," University of Adrar, Algeria.
3. Al-Obeidi, Nada Nizar Muhammad, (2015 AD) "**Proposed weighted methods for two-stage cluster immunity and estimation of regression models (clustered regression) by application to the data of a sample of thalassemia patients in Nineveh Governorate**," master's thesis, College of Computer Science and Mathematics, University of Mosul - Iraq.
4. Al-Mawla, Taqwa Abd al-Salam Taha, (2016AD), "**A comparison between the methods of regression of the principal components and partial least squares with the application on the Kirkuk Cement Factory**," Tikrit Journal of Pure Sciences, Issue-21, pp. (185-203).
5. Al-Nuaimi, Aswan Muhammad Tayyib Rashid, (2005AD), "**Choice of variables in character regression**," unpublished master's thesis, College of Computer Science and Mathematics, University of Mosul- Iraq.
6. Khattar, Gibran Abdul-Amir and Muhammad, Korkis Shahid and Ismail, Muhammad Salem, (2016AD), "**Estimating the parameters of multiple linear regression using robust methods (comparative study)**", Al Qadisiyah Journal of Computer Science and Mathematics, Iraq, Volume 8, Number 1.
7. Salem, Mohamed Abdel Wahhab Mohamed, (2018 AD), "**Methods of treating multiple linear overlap of aggregate investment data in Sudan (1980-2014 AD)**", Sudan University of Science and Technology - College of Graduate Studies.
8. Alexei, Botchkarev, (2018), "**Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio**", SSRN Electronic Journal, 10.2139/ssrn,3177507.
9. Almetwally, Ehab & Almongy, Hisham, (2018), "**Comparison Between M-estimation, S-estimation, And MM Estimation Methods of Robust Estimation with Application and Simulation"**.
10. Andersen, R., (2008), "**Modern methods for robust regression",** (No. 152), Sage.
11. Asteriou, D., and Hall, S. G., (2007), "**Applied econometrics: A modern approach using EViews and Microfit**", Palgrave Macmillan, New York, [p496].
12. Boiroju, N.K., and Reddy, M.K., (2012), "**A Graphical Method for Model Selection**", Pakistan Journal of Statistics & Operation Research, pp. 767-776.
13. Belsley, D. A., Kuh, E., and Welsch, R. E., (2004), "**Diagnostics: Identifying Influential Data and Sources of Collinearity**", John Wiley & Sons, New York.
14. Cooley, W.W. A., and Lohnes, P. R. A.,(1971), "**Multivariate Data Analysis**", JohnWiley & Sons, Australia, ISBN 9780471170600, [p496].
15. Curto, J. D., and Pinto, J. C., (2011), "**The corrected VIF (CVIF)**", Journal of Applied Statistics, 38(7):1499–1507.
16. Farrar, D. E., and Glauber, R. R., (1967), "**Multicollinearity in regression analysis: The problem revisted**", The Review of Economics and Statistics, 49:92–107, [p495, 496, 498].
17. Greene, W. H., (2002), "**Econometric Analysis**", Prentic-Hall, New Jersey, 5th edition, [p495, 497].
18. Gujarati, D. N., and Porter, D. C., (2008), "**Basic econometrics**", McGraw Hill, 5th edition, [p495, 496, 497, 498,503].
19. Heritier, Stephane & Cantoni, Eva & Copt, Samuel & Victoria-Feser, Maria-Pia., (2009), "**Robust Methods in Biostatistics**", 10.1002/978047074053.
20. Huber, P.J., (1964), "**Robust Estimation of a Location Parameter**", Annals of Mathematical Statistics.USA, 35:73- 101.
21. Imdadullah, M., Aslam, M., and Altaf, S., (2016), "**Mctest: An R Package for detection of collinearity among regressors**", R J., 8, 499–509.
    Available online: https://journal.r-project.org/archive/2016/RJ-2016-062/index.html
    (Accessed on 26 March 2020).
22. Kendall, M. G., (1957), "**A Course in Multivariate Analysis**", Charles Griffin & Company, London.
23. Kovács, P. Petres, T., and Tóth, (2005), "**A new measure of multicollinearity in linear regression models**", International Statistical Review / Revue Internationale de Statistique, 73(3):405–412.
24. Kutner M. H., Nachtsheim C. J., & Neter J., (2004), "**Applied Linear Regression Models**". McGraw Hill Irwin, 4th edition, [p495, 497].
25. Maddala, G.S., (1988), "**Introduction to Econometrics**", Macmillan, New York,[p496].
26. Maddala, G. S., (1992), "**Introduction to econometrics**", Macmillan, New York.

27. Makridakis, S., and Hibon, M., (1995), "**Evaluating accuracy (or error) Measures**", INSEAD Working Papers Series 95/18/TM. Fontainebleau, France.
28. Marquardt, D.W., (1970), "**Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation**", Technimetrics, 12(3):591–612.
29. Memmedli, M., and Ozdemir, O., (2009), "**A Comparison Study of Performance Measures and Length of Intervals in Fuzzy Time Series by Neural Networks**", Proceedings of the 8th Wseas International Conference on System Science and Simulation in Engineering.
30. Maronna, R. A., Martin, R. D., & Yohai, V. J., (2019), "**Robust statistics: theory and methods (with R)**", John Wiley & Sons.
31. Riazoshams, Hossein & Midi, Habshah & Ghilagaber, Gebrenegus., (2019), "**Robust Nonlinear Regression: with Applications using R**", 10.1002/9781119010463.
32. Rousseeuw , J. , Leroy , A. , (2005) ,**" Robust regression and outlier detection"**, John Wily and Sons , New York .
33. Sarwar, A., and Sharma, V., (2014), "**Comparative analysis of machine learning techniques in prognosis of type II diabetes**", AI & society, 29(1), 123-129.
34. Silvey, S. D., (1969), " **Multicollinearity and imprecise estimation**", Journal of the Royal Statistical Society, Series B (Methodological), 31(3):539–552, [p495, 497].
35. Stock, J. H, and Watson, M. W., (2010), "**Introduction to Econometrics**", Pearson Addison-Wesley, 3rd edition, [p496].
36. Susanti, Y., & Pratiwi, H., (2014), "**M estimation, S estimation, and MM estimation in robust regression**", International Journal of Pure and Applied Mathematics, 91(3), 349-360.
37. Theil, H., (1971), "**Principles of Econometrics**", John Wiley & Sons, New York, [p495, 497].
38. Willmott, C., and Matsuura, K., (2005), "**Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance**", Climate Research, 30(1), 79-82.
39. Woschnagg, E., and Cipan, J., (2004), "**Evaluating Forecast Accuracy**", UK Ö konometrische Prognose, University of Vienna, Department of Economics.
40. Yohai, V. J., (1987), "**High breakdown-point and high efficiency robust estimates for regression**", The Annals of Statistics, 642-656.

Employing Robust MM-estimators in Estimating Principal Component Regression Model - A Comparative Study

Esraa Najeeb Saeed Al-Saraf Bashar A. Al-Talib
College of Computer & Mathematical Science, University of Mosul, Mosul.

Abstract

This paper focuses on proposing the use of robust MM estimators in estimating the parameters of the principal component regression model, which is usually used in estimating the regression model when the explanatory variables are not independent. even in the presence of leverage points in the data and gives estimators with good efficiency, this estimator has been called the MM estimator, referring to the fact that more than one M estimator is used to obtain the final estimator as the estimation is done using the Iteratively Re-weighted Least Square (IRLS) method.

The classical and the proposed models will be applied to data obtained from the cement factory in Badoush. On this basis, a comparison between the known classical methods and the proposed method using MM estimators will be made through two experimental studies using simulation and application of cement factory data.

Key Words: Principal Component Regression, Leverage Points, Robust Methods, MM Estimators, Weighted Least Squares.