Iraqi Journal of Statistical Sciences, Vol.18, No. 1, 2021 , pp (1-13)

المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

أساليب حصينة موزونة لكشف والتعامل مع الشواذ في تقدير نموذج انحدار المكونات الرئيسية

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 8 آب 2020
تم القبول في 25 ايلول 2020
متاح على الإنترنت في 1 حزيران 2021

الكلمات الدالة:
إنحدار المكونات الرئيسية،
القيم الشاذة،
القيم الجاذبة،
المربعات الصغرى الموزونة،
تعدد العلاقة الخطية

المراسلة:
اسراء نجيب سعيد الصراف
basharaltalib@gmail.com

الخلاصة

يهدف البحث إلى إقتراح أسلوب لمعالجة مشكلة التداخل الخطي المتعدد بين المتغيرات التفسيرية ومشكلة وجود قيم شاذة في البيانات عن طريق استخدام اسلوب انحدار المكونات الرئيسية ، ومن ثم استخدام دوال أوزان حصينة لوزن دالة الهدف للتعامل مع وجود القيم الشاذة في البيانات ، ومن أجل التحقق من كفاءة المقدرات تم إجراء دراسة تجريبية من خلال أسلوب المحاكاة، كما تم تطبيق الطرق على بيانات حقيقية تم جمعها من ملفات معمل اسمنت بادوش في محافظة نينوى للفترة من (2008-2014) بتسعة متغيرات تفسيرية تمثل الخواص الكيميائية للإسمنت ومتغير تابع يمثل الخواص الفيزيائية للإسمنت (الصلابة) ، وقد تم اختبار فيما إذا كانت البيانات تعاني من مشكلة تعدد العلاقة الخطية ومن ثم تطبيق المربعات الصغرى باستخدام المكونات الرئيسية كمتغيرات مستقلة وتقدير النموذج ، وقد وجد أن المتغيرات تعاني من مشكلة تعدد العلاقة الخطية ، وتمت المعالجة عن طريق تطبيق انحدار المكونات الرئيسية الموزونة بأوزان حصينة وذلك لوجود قيم شاذة في البيانات بالإضافة لمشكلة التعدد الخطي.

DOI: https://doi.org/10.33899/iqjoss.2021.168371 , ©Authors, 2021, College of Computer and Mathematical Science, University of Mosul.
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

1- المقدمة Introduction

في حالة وجود مشكلة تعدد العلاقة الخطية في البيانات فإن تحليل الانحدار الخطي المتعدد يعطي مقدرات غير موثوقة لمعلمات الانحدار وتباين تلك المعلمات يمكن أن يكون كبيراً الأمر الذي يؤدي بنا إلى استخدام الطرق المتحيزة ومنها طريقة انحدار المكونات الرئيسية فضلاً عن بعض الحالات التي تنطوي على وجود القيم الشاذة والتي من الممكن معالجتها بإستخدام أوزان حصينة لكي تتم السيطرة على تأثير تلك القيم على معلمات النموذج المقدر .

وتعتبر طريقة المكونات الرئيسية أسلوب فعال في التعامل مع مشكلة التداخل الخطي بين المتغيرات التفسيرية ومعالجتها، وذلك لان المكونات الرئيسية دائماً ما تكون متعامدة (مستقلة)، كما أن المكونات الرئيسية قامت بدورها في اختزال عدد المتغيرات التفسيرية وهناك عدة اختبارات ومعايير معروفة تستخدم للكشف عن وجود مشكلة التداخل الخطي بين المتغيرات التفسيرية. ويجمع الاسلوب الكلاسيكي بين تحليل المكونات الرئيسية (Principal Component Analysis PCA) مع انحدار المربعات الصغرى. ومع ذلك، تعطي كلتا المرحلتين نتائج غير موثوقة عندما تحتوي مجموعة البيانات على قيم شاذة. ولذلك تم في هذا البحث اقتراح استخدام طريقة انحدار المكونات الرئيسية الحصينة والتي تتم على مرحلتين، يتم في الأولى تطبيق طريقة تحليل المكونات الرئيسية الاعتيادية على البيانات، ثم نقوم بإعادة تطبيق انحدار المكونات بإستخدام طريقة انحدار حصينة (Huber & Verboven,2003)(Huber & Verboven, 2003).

2- هدف البحث

يهدف هذا البحث إلى تناول مشكلتين في نماذج الانحدار الأولى هي وجود مشكلة تعدد العلاقة الخطية بين المتغيرات المستقلة والثانية مشكلة وجود الشواذ في البيانات سواءاً كانت في المتغير المعتمد أو في المتغيرات المستقلة ، وعليه هدف البحث إلى ايجاد طريقة تتعامل مع المشكلتين في آنٍ واحد بحيث تقلل من تأثير القيم الشاذة على النموذج وكذلك تزيل تأثير مشكلة تعدد العلاقة الخطية ، وعليه تم اللجوء إلى استخدام انحدار

المكونات الرئيسية الذي يتم فيه تحويل المتغيرات إلى مركبات أو مكونات رئيسية لاتعاني من مشكلة تعدد العلاقة الخطية وتكون مستقلة ثم استخدام أوزان حصينة لوزن النموذج واستخدام طريقة المربعات الصغرى الموزونة بمتغيرات هي عبارة عن المركبات أو المكونات الرئيسية للمتغيرات المستقلة وفي النهاية الوصول إلى نموذج نحصل من خلاله على مقدرات كفوءة تمتلك خاصيتي الكفاءة والحصانة.

3- الجانب النظري

3-1 طريقة انحدار المكونات الرئيسية Principal Components Regression Method

تم اقتراح انحدار المكونات الرئيسية لأول مرة بواسطة (Kendall,1957). حيث تم استخدام نتائج تحليل المكونات الرئيسية التي يتم إجراؤها على مقدرات نموذج الانحدار واستخدام المركبات الناتجة كمتغيرات جديدة . وبهذه الطريقة تكون المتغيرات المستقلة متعامدة وتضمن أن الحسابات أسهل وأكثر استقراراً (Jolliffe,1982). ويتم اللجوء إلى استخدام PCA في الانحدار الخطي لخدمة هدفين أساسيين. يتم تنفيذ الأول على مجموعات البيانات حيث يكون عدد المتغيرات المستقلة كبيراً وترتبط مع بعضها. لقد كانت طريقة انحدار المكونات الرئيسية اسلوباً لتقليل الأبعاد جنباً إلى جنب مع انحدار المربعات الصغرى الجزئية. أما الهدف الثاني من (PCR) فهو التخلص من تعدد العلاقة الخطية المتداخلة بين المتغيرات. ونظراً لأن كل مكون رئيسي هو متعامد، فقد تم استخدام PCR لمنع الأخطاء التي تتسبب بها المشكلة بين المتغيرات المستقلة المفترضة في الانحدار وعندما يتعلق الأمر باختيار عدد المكونات الرئيسية المناسبة، فإن الباحثين لم يجمعوا على رأي واحد واقترحوا عدة أساليب ومنها اختيار أفضل المكونات الرئيسية كما لو كانت متغيرات منتظمة. وبين باحثون آخرون أنه من الأفضل اختيار أول عدد محدد من المكونات الرئيسية التي تفسر أعلى التباين (Hadi&Ling,1998). وهذا يؤدي إلى رفض بعض المكونات الرئيسية التي تكون مساهمتها في تفسير التباين منخفضة ومع ذلك ، فقد تم انتقاد هذا النهج حيث يمكن للمكونات الرئيسية المرفوضة أن تكون في الواقع هي تلك التي ترتبط بالمتغير التابع وذلك إثر وجود مشكلة تعدد العلاقة الخطية مع مشكلة أخرى كوجود قيم شاذة في البيانات (Outliers) والتي تحرف النموذج عن مساره الطبيعي وهذا هو محور اهتمام هذا البحث.

تعتمد طريقة تحليل المكونات الرئيسية اسلوب تحويل المتغيرات التوضيحية الأصلية الى متغيرات جديدة تسمى " بالمكونات أو المركبات الرئيسية".
حيث ان كل مكون (مركب) رئيسي هو عبارة عن تركيبة خطية في المتغيرات التفسيرية الأصلية.
ويتم تحويل المتغيرات التفسيرية الى المكونات الرئيسية بالشكل الآتي:

$$\hat{y} = \hat{a}_{\circ} I + Xa\acute{a} + \underline{u} \quad (1)$$

حيث أن :

a: عبارة عن مصفوفة المتجهات المميزة المرافقة لمصفوفة الارتباط بين المتغيرات التوضيحية، فإذا عوضنا عن Xa بكمية ثابتة (pc) والتي تمثل مصفوفة ذات بعد (n×p) أعمدتها عبارة عن معاملات انحدار النموذج المحور فإن النموذج الناتج يأخذ الشكل الآتي:

$$y^{*}=\hat{a}_{\circ} I + pc^{*}\acute{a}^{*}+ \underline{u} \quad (2)$$

وعند تطبيق اسلوب المكونات الرئيسية يفضل تحويل المتغيرات الى متغيرات قياسية اذا كان هناك اختلاف في وحدات القياس ،حيث أن:

$$Z_{ij} = \frac{(x_{ij}-\bar{x}_j)}{\sigma_{ij}} \quad (3)$$

حيث أن: $\bar{x}_j$: يمثل الوسط الحسابي للمتغير

σ_{ij} : يمثل الإنحراف المعياري

ويعود تاريخ استخدام طريقة انحدار المكونات الرئيسية إلى أعمال كل من Beltrami في عام 1873 و Jordan في عام 1874 حيث قاما بشكل منفصل بوضع ما يسمى بـ Singular Value Decompostiton (SVD) مما مهد إلى تعريف تحليل المكونات الرئيسية PCA من قبل كل من Pearson في عام 1901 و Hotelling في عام 1933.

تعتبر طريقة المكونات الرئيسية واحدة من النماذج الخطية المتحيزة الواسعة الاستخدام لتخطي مشكلة تعدد العلاقة الخطية التي كثيراً ما يعاني منها نموذج الانحدار الخطي المتعدد . وتقوم طريقة انحدار المكونات (المركبات) الرئيسية على تحويل المتغيرات التفسيرية الأصلية المرتبطة دون حذف أي منها الى متغيرات جديدة متعامدة (أي مستقلة) تسمى بالمكونات الرئيسية ، وكل مركب رئيسي عبارة عن تركيب خطي في المتغيرات التفسيرية الأصلية(مستور وعبد الرحيم،2016). تقدم المكونات الرئيسية قدر كبير من المعلومات عن مشاهدات المتغيرات الأصلية مثل أنماط تجمعاتها وعلاقتها بالمتغيرات الأصلية ،وتقدم ايضاً معلومات عن الارتباطات بين المتغيرات الجديدة والقديمة والمجموعات أو التصنيفات التي تحتويها البيانات أو المتغيرات. وعادةً يتم ترتيب المكونات الرئيسية وفقاً لمقدار التباين بحيث تكون المركبة الأولى هي المركبة ذات التباين الأكبر، ومن ثم يتم اعتماد عدد قليل من المكونات التي يتوقع أن تفسر أكبر قدر ممكن من التباين ، ويتم اهمال المكونات ذات التأثير الأقل. وتعتبر

عملية ايجاد المكونات الرئيسية خطوة مهمة لإزالة أثر التعدد الخطي تمهيداً لاستخدام طريقة المربعات الصغرى الاعتيادية لتقدير معالم نموذج الانحدار الخطي الأصلية للمتغيرات التفسيرية (جبريل،2014). فإذا كانت $(X_1,X_2,\ldots,X_p)$ متغيرات تفسيرية ، فيمكن تعريف توليفة متعامدة منها وفقاً للمعادلة (4) التي تناظر المعادلة(1).

$$Z = XA \quad (4)$$

حيث تمثل Z مصفوفة المكونات الرئيسية من الرتبة $(n\times p)$، بينما مصفوفة A فهي عبارة عن مصفوفة متعامدة للمتجهات المميزة المعيارية المناظرة للجذور المميزة لمصفوفة معلومات النظام $(X'X)$ ورتبتها $(p\times p)$، عناصرها a_{ij} وأعمدتها A_j وهي تجعل المصفوفة $(X'X)$ مصفوفة قطرية ، وباعتبار $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_p$ قيم مميزة للمصفوفة $(X'X)$ فإن المتغير Z_j يتوزع بمتوسط يساوي الصفر وتباين λ_j.

وللتعبير عن Y كدالة في المكونات الرئيسية بدلاً من المتغيرات المستقلة $(X_1,X_2,\ldots,X_p)$ المرتبطة فيما بينها ، وبما أن A مصفوفة متعامدة حيث $A'A = I$ فيمكن اعتبار $X=ZA'$ بالنسبة الى معادلة نموذج الانحدار $Y=X\beta+\varepsilon$ فنحصل على نموذج الانحدار وفقاً للمعادلة التالية:

$$Y=ZA'\beta + \varepsilon \quad (5)$$

وعلى افتراض أن $A'\beta = \gamma$ فتصبح المعادلة لنموذج الانحدار وفقاً للمعادلة التالية:

$$Y=Z\gamma + \varepsilon \quad (6)$$

حيث γ تمثل متجه المعلمات $(\gamma_1,\gamma_2,\ldots,\gamma_p)$ المناظرة للمركبات الرئيسية $(Z_1,Z_2,\ldots,Z_p)$ التي يمكن تقديرها باستخدام طريقة المربعات الصغرى الاعتيادية وفقاً للمعادلة التالية:

$$\gamma = (Z'Z)^{-1}Z'Y = \Lambda^{-1}Z'Y \quad (7)$$

والمصفوفة Λ تعتبر مصفوفة قطرية من الرتبة $(p\times p)$ عناصرها عبارة عن الجذور المميزة للمصفوفة $(X'X)$. والتوقع لهذه المعلمات هو $E(\hat{\gamma})=\gamma$ وتباينها $Var(\hat{\gamma}) = \Lambda^{-1}\sigma^2$ ، عليه يمكن القول إن متجه المعلمات $\hat{\gamma}$ له توزيع طبيعي بمتوسط γ وتباين $\Lambda^{-1}\sigma^2$. وبالتالي فإن تباين أي معلمة ضمن متجه المعلمات $\hat{\gamma}$ يحسب وفقاً للصيغة:

$$Var\ (\gamma_j) = \frac{\sigma^2}{\lambda_j} \quad (8)$$

مع الأخذ في الاعتبار أن $\hat{\gamma}_i=Z_i'Y/\lambda_i$ ، ويتم التنبؤ بقيمة المتغير التابع Y وفقاً للمعادلة (9):

$$\hat{Y} = \sum_{j=1}^{p} Z_j\,\hat{\gamma}_j \quad (9)$$

ويكون التباين المخفض لتوفيق نموذج الانحدار باستخدام المركب الرئيسي Z_j يساوي المقدار $\lambda_j\,\gamma_j^2$ ، عليه فإن نسبة التباين المفسر في قيم متغير الاستجابة Y بواسطة المركب الرئيسي Z_j هي:

$$(\lambda_j\gamma_j^2\,/\,Y'Y)*100 \quad (10)$$

وتساوي هذه النسبة مربع معامل ارتباط متغير الاستجابة والمركب الرئيسي Z_j مع ضرب الناتج في مائة . وبناءاً عليه يكون مربع الخطأ لمعادلة الانحدار المقدرة هو:

$$MSE = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{j=1}^{p}\lambda_i\,\hat{\gamma}_j^2 \quad (11)$$

وللحصول على معلمات المتغيرات التفسيرية الأصلية لنموذج الانحدار، يستفاد من العلاقة بين المعلمات الأصلية $\hat{\beta}$ ومعلمات نموذج الانحدار $\hat{\gamma}$ الخاصة بانحدار المتغير Y على المكونات الرئيسية Z وفقاً لما يلي:

اذا كان

$$A'\hat{\beta} = \hat{\gamma} \quad (12)$$

و

$$A'A = I \quad (13)$$

فإن

$$\hat{\beta} = A\hat{\gamma} \quad (14)$$

وبما أن المعلمات $\hat{\beta}$ تتوزع طبيعياً بمتوسط $A\gamma$. فإن القيمة المتوقعة للمعلمة β_i تحسب وفقاً لما يلي:

$$E(\hat{\beta}_i) = \sum_{j=1}^{p} a_{ij}\ \gamma_j \qquad (15)$$

وبما أن المعلمات $\hat{\beta}$ لها التباين $\Lambda^{-1}A\ \sigma^2$. فإن تباين المعلمة β_i يحسب وفقاً للمعادلة التالية:

$$Var(\hat{\beta}_i) = \sigma^2 \sum_{j=1}^{p} {a_{ij}^2}/{\lambda_j} \qquad (16)$$

وعند استخدام الجذور المميزة لمصفوفة معاملات الارتباط بدلاً من مصفوفة التباين والتغاير كمدخلات في تحليل انحدار المكونات الرئيسية فإنه يجب استخدام $n\lambda_j$ بدلاً من λ_j. وتوضح العلاقة أن تباين معلمات نموذج الانحدار المقدرة $\hat{\beta}$ ايضاً تعتمد على الجذور المميزة للمصفوفة $(X'X)$ وبناءاً على ذلك فهي تتأثر بوجود الجذور المميزة الصغيرة التي ينتج عنها تضخم التباينات. ووفقاً لتعريف المكونات الرئيسية فإن الجذور المميزة الصغيرة التي تسهم في تضخيم تباين معلمات نموذج الانحدار دائماً تقابل المكونات الرئيسية الأخيرة للمصفوفة $(X'X)$،عليه يتطلب تخفيض التباين الكلي للمعلمات ، واستبعاد المكونات الرئيسية المقابلة لأصغر الجذور المميزة للمصفوفة $(X'X)$.

اقترح بعض الباحثين أمثال Chatterjee&Price,Jolliffee,Jeffers (جبريل،2014) أن يتم استبعاد المكونات الرئيسية التي تقابل الجذور المميزة التي تقل عن 70%. كما اقترح Morrison اختيار المكونات الرئيسية التي تفسر على الأقل 75% من التباين في قيم متغير الاستجابة. وهذه النسبة يمكن الحصول عليها بقسمة مجموع الجذور المميزة المقابلة لـ K من المكونات الرئيسية على مجموع الجذور المميزة عند استخدام مصفوفة التباينات والتغايرات للمتغيرات التفسيرية كمدخلات لتحليل المكونات الرئيسية وفقاً لـــ:

$$\left(\sum_{j=1}^{k} \lambda_j \ / \sum_{j=1}^{p} \lambda_j \right) * 100 \qquad (17)$$

اما عند استخدام مصفوفة معاملات الارتباط كمدخلات لتحليل المكونات الرئيسية فعندئذٍ يتم استخدام عدد المتغيرات التفسيرية P بدلاً من مجموع الجذور المميزة. ويتم بناء نموذج الانحدار لمتغير الاستجابة Y على المكونات الرئيسية المتبقية ، بعد استبعاد المركبات التي لا تحقق المعايير السابقة . بافتراض أن S من الجذور المميزة لها قيم كبيرة من بين P من الجذور المميزة للمصفوفة $(X'X)$ ، يكون هناك (P-S) من المكونات الرئيسية Z، ومن ثم يجري توفيق نموذج انحدار Y على المكونات الرئيسية المتبقية وبذلك تكون المعادلة التنبؤية كما يلي:

$$\hat{Y}_s = \sum_{j=1}^{s} Z_j\ \hat{\gamma}_j \qquad (18)$$

ويحسب مجموع مربعات الخطأ الخاص بنموذج الانحدار المقدر بعدد S من المكونات الرئيسية وفقاً لما يلي:

$$\hat{\sigma}_s^2 = MSE = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \sum_{j=1}^{s} \lambda_i\ \hat{\gamma}_j^2 \qquad (19)$$

ولحسن الحظ فإن خاصية التعامد لمقدرات المربعات الصغرى لـ γ سوف لن تختلف في حال استخدام جميع المكونات الرئيسية أو مجموعة جزئية منها، وتأسيساً على ذلك يتم تقدير معلمات نموذج الانحدار γ_s وفقاً لما يلي:

$$\hat{\gamma}_s = \Lambda_s^{-1}\ Z_s'\ Y \qquad (20)$$

حيث أن:

$$\Lambda_s = diag\ (\lambda_1, \lambda_2, \dots, \lambda_s) \qquad (21)$$

ويتم الحصول على متجه المعلمات $\hat{\gamma}_s$ بتجزئة المصفوفة وفقاً لما يلي:

$$A = [A_s : A_{p-s}] \qquad (22)$$

حيث أن:

$$A_s = [A_1, A_2, \dots, A_s] \qquad (23)$$

$$A_1 = [a_{11}, a_{12}, \dots, a_{1s}] \qquad (24)$$

ومصفوفة المكونات الرئيسية هي:

$$Z = [Z_s : Z_{p-s}] \qquad (25)$$

حيث أن:

$$Z_s = [Z_1, Z_2, \dots, Z_s] \qquad (26)$$

ومتجه المعلمات هو:

$$\gamma_s = [\gamma_s\ :\ \gamma_{p-s-1}] \qquad (27)$$

وعند الحصول على تقدير متجه المعلمات $\hat{\gamma}_s$ يمكن استخدامه في تقدير متجه معلمات نموذج الانحدار الأصلي $\hat{\hat{\beta}}$ وتحسب وفقاً لما يلي:

$$\hat{\hat{\beta}} = A_s \hat{\gamma}_s \qquad (28)$$

ويمكن الحصول على معلمات الانحدار للمتغيرات التفسيرية باستخدام معلمات الانحدار للمكونات الرئيسية وفقاً للمعادلة التالية:

$$\hat{\hat{\beta}}=\sum_{j=1}^{s} a_{ij}\ \hat{\gamma}_j \qquad , i = 1,2, \dots, p \qquad (29)$$

وتباين $\hat{\hat{\beta}}_i$ يقدر وفقاً لما يلي:

$$\text{Var}\ (\hat{\hat{\beta}}) = \sigma_s^2\ \sum_{j=1}^{s} \gamma_{ij}^2\ /\ \lambda_j \qquad (30)$$

ويعتبر المقدر $\hat{\hat{\beta}}$ مقدر متحيز ، ويحسب مقدار تحيزه وفقاً لما يلي:

$$\text{Bias} = E\ (\hat{\hat{\beta}}_i) - \beta_i \qquad (31)$$

$$= -\sum_{j=s+1}^{p} a_{ij}\, \gamma_j \qquad (32)$$

وبذلك فإن متوسط مربع الخطأ للمقدر $\hat{\hat{\beta}}$ يحسب وفقاً لما يلي:

$$\text{MSE}(\hat{\hat{\beta}}_i) = \text{var}\ (\hat{\hat{\beta}}_i) + \text{Bias of}\ (\hat{\hat{\beta}}_i)^2 \qquad (33)$$

$$= \sigma_s^2\ \sum_{j=1}^{s} a_{ij}^2\ /\ \lambda_j\ + (\ -\sum_{j=s+1}^{p} a_{ij}\, \gamma_j)^2 \qquad (34)$$

3-2 طريقة مقدرات (M)

قام الباحث (Huber,1973) بتوسيع نتائجه للتقدير الحصين من معلمة الموقع إلى حالة الانحدار الخطي باستخدام مقدرات M. وقد اكتسبت هذه التقديرات شهرة أكثر من بقية المقدرات الحصينة الأخرى لأنها أكثر مرونة وكذلك توفر إمكانية تعميمها مباشرة إلى الانحدار المتعدد . حيث أن طريقة مقدرات (M) تهدف إلى تصغير المقدار

$$\text{Min}\sum_{i=1}^{n} \rho(e_i) \qquad (35)$$

$$\text{Min}\sum_{i=1}^{n} \rho\left(\underline{y} - X\underline{\beta}\right) \qquad (36)$$

إذ تمثل ρ دالة بدلالة الأخطاء ولتصغير المعادلة (36) نشتقها جزئياً بالنسبة للمتجه $\hat{\beta}$ ومساواتها بالصفر وكما يلي:

$$\sum_{i=1}^{n} x_i\, \varphi\left(\underline{y}\ - X\underline{\hat{\beta}}\right) \qquad (37)$$

إذ تمثل φ المشتقة الجزئية للدالة (ρ) بالنسبة للمعلمات في المعادلة (37) وتمثل منظومة مكونة من (P) من المعادلات وتحل باستخدام أحدى الطرق العددية المعروفة أو طريقة المربعات الصغرى الموزونة (Weighted Least Squares Method) ولإيجاد مقدرات M التي تحقق المعادلة (36) وذلك باستخدام الصيغة التالية:

$$\hat{\beta}_M = (X'WX)^{-1}\, X'W\underline{y} \qquad (38)$$

إذ تمثل (W) مصفوفة الأوزان وهي مصفوفة قطرية($n \times n$)عناصرها القطرية معطاة بالصيغة الآتية:

$$W_i = \frac{\varphi(e_i)}{(e_i)} \qquad (39)$$

$$W_i = \frac{\left[\left(y_i - X_i\underline{\beta}_{\circ}\right)\right]}{\left(y_i - X_i\underline{\beta}_{\circ}\right)} \qquad (40)$$

إذ تمثل $\underline{\beta}_{\circ}$ القيم الابتدائية لمتجه معلمات النموذج ويتم استخدامها لتحديد الأوزان ويمكن استخدام مقدرات المربعات الصغرى كقيم ابتدائية ومن التكرار الأول نجد قيمة $\underline{\hat{\beta}_1}$ وفي التكرار الثاني نستخدم $\underline{\hat{\beta}_1}$ في إيجاد الأوزان لإيجاد $\underline{\hat{\beta}_2}$ وهكذا تستمر عملية التكرار حتى نحصل على مقياس التقارب (Convergence) المعرف بالصيغة الآتية:−

$$\text{Max}\left[\left|\hat{\beta}_j^{(n)} - \beta_j^{(n-1)}\right|\right] < \text{ʒ} \qquad (41)$$

إذ تمثل ʒ قيمة صغيرة جداً و(n) تمثل رقم التكرار أي أن الحل يتوقف عندما يصبح الفرق المطلق بين المعلمات المقدرة في المرحلة الحالية والمرحلة السابقة أصغر من القيمة المختارة (ʒ) أو يساويها ولجعل مقدرات M تمتلك خاصية (Scale Invariant) فان الدالة المطلوب تصغيرها هي :−

$$\text{Min}\sum_{i=1}^{n} \rho\left(y_i - X_i\underline{\beta}\right)/\hat{\sigma} \qquad (42)$$

ثم نشتقها بالنسبة للمتجه $\underline{\beta}$ ومساواتها بالصفر

$$\sum_{i=1}^{n} X_{ij}\Psi\left(y_i - X_i\underline{\beta}\right)/\hat{\sigma}=0 \qquad (43)$$

يمكن حل المعادلة أعلاه باستخدام المعادلة (38) حيث أن الأوزان يتم إيجادها وفق الصيغة الآتية:-

$$W_i = \frac{[\Psi(e_i/\hat{\sigma})]}{(e_i/\hat{\sigma})} \qquad (44)$$

$$W_i = \frac{\left[\Psi\left(\frac{y_i - X_i\underline{\beta}}{\hat{\sigma}}\right)\right]}{\left(\frac{y_i - X_i\underline{\beta}}{\hat{\sigma}}\right)} \qquad (45)$$

ولإيجاد $(\hat{\sigma})$ في المعادلة اعلاه والتي تمثل قيمة المقدر المعياري وان هذه القيمة تقدر مرة واحدة فقط باستخدام القيم الأولية قبل البدء بالتكرار وهناك عدة صيغ لتقديرها منها:

1) $\hat{\sigma} = 1.5\, med\, |e_i|$
2) $\hat{\sigma} = 2.1\, med\, |e_i|$
3) $\hat{\sigma} = 1.485[med\, |e_i - med\, e_i|]$
4) $\hat{\sigma} = \frac{[\text{med}|e_i|]}{0.6745}$

إذ تمثل (e_i) البواقي (Residuals) وmed يشير إلى الوسيط ولقد اقترح الباحثون عدداً من الدوال $\rho(.)$ أو مشتقاتها $\Psi(.)$ بحيث تجعل نتائج التقدير جيدة ولا تتأثر بوجود الشواذ وفيما يلي بعض الدوال المهمة لهذا النوع من المقدرات والمعرفة بدلالة الدالة $\Psi(e_i)$. وبإفتراض أن وسيط الأخطاء المطلقة (MAD) (Median Absolute Deviation) يأخذ الصيغة الآتية (الراوي،2017):-

$$MAD = median|ei - median\,(ei)|\,/\,0.6745 \qquad (46)$$

وقد عمد الباحثون (Montgomery, et. Al., 2001) في هذه الدراسة إلى عرض دوال الوزن الترجيحية لمقدرات M بأسلوب سهل وبأستخدام رموز سهلة لم يعهد أستخدامها في أدبيات الإحصاء الحصين لإشاعة إستخدامها من قبل الباحثين في المستقبل. وبأخذ الصيغة القياسية للبواقي (Standardized Residuals) باستخدام المعادلة (47) وكما يلي:

$$e_{is} = \frac{e_i}{MAD} \qquad (47)$$

وبأفتراض ثابت القطع ((Tunning Constant (c) الذي يجعل التباين المقدرMAD مقدر غير متحيز تقريباً لـ σ عندما يكون حجم العينة كبيراً والخطأ يتوزع طبيعياً (الطالب،2011).

1. دالة Hampel

$$\Psi_{Hampel}(e_{is}, c) = \begin{cases} 1 & if\ \ |e_{is}| \le a \\ \frac{a}{|e_{is}|} & if\ a < |e_{is}| \le b \\ \frac{a(c-|e_{is}|)}{|e_{is}|(c-b)} & b < |e_{is}| \le c \\ 0 & otherwise\ (i.e.\, |e_{is}| > c\,) \end{cases} \qquad (48)$$

حيث أن القيم الأفتراضية لثوابت القطع (Tuning Constatnts) تكون في برنامج S-Plus:

$a = 2 \quad b = 4 \quad \text{and } c = 8$

وهناك بعض المصادر تفترض قيم أخرى لثوابت القطع (Montgomery, et. Al., 2001) مثل:

$a = 1.7 \quad b = 3.4 \quad \text{and } c = 8.5$

2. دالة Huber (Huber,1964)

$$\Psi_{Huber}(e_{is}, c) = \begin{cases} 1 & \text{if } |e_{is}| \le c \\ \frac{c}{|e_{is}|} & \text{otherwise} \end{cases} \qquad (49)$$

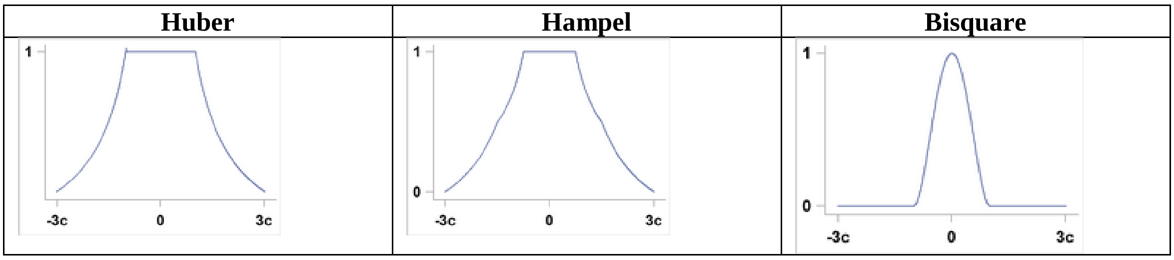
حيث أن c تأخذ القيمة الأفتراضية $c = 1.345$

3. دالة Bisquare والتي تسمى أحياناً بدالة الوزن التربيعي المزدوج (Tukey's Biweight)

$$\Psi_{\text{Bisquare}}(e_{is}, c) = \begin{cases} \left[1 - \left(\frac{e_{is}}{c}\right)^2\right]^2 & \text{if } |e_{is}| \le c \\ 0 & \text{otherwise} \end{cases} \qquad (50)$$

حيث أن c تأخذ القيمة الأفتراضية $c = 4.685$

الشكل (1) يبين الأشكال البيانية لدوال الوزن المشار إليها في أعلاه:



الشكل (1) : التمثيل البياني لدوال الوزن لمقدرات M المستخدمة

وأخيراً فإن ثابت القطع (Tunning Constant (C)) لكل دالة يستخدم لتعديل كفاءة المقدرات الناتجة لتوزيعات محددة بكفاءة تقريبية (95%) عندما تتبع الأخطاء التوزيع الطبيعي ، وان الاختبار الجيد لقيمة ثابت القطع يؤدي إلى زيادة حصانة المقدرات لان له تأثير كبير على حصانة المقدرات وان قيمته تتراوح مابين انحراف معياري واحد إلى انحرافين معياريين لقيم المشاهدات أو الأخطاء أي مثلا (S<H<2S) من ما تقدم بأن ثابت القطع يعدل للحصول على مقدرات جيدة (Al–Rawi,2017).

4- الجانب التجريبي

إن تحليل المكونات أو المركبات الرئيسية هو الأساس في أسلوب انحدار المكونات الرئيسية حيث أنه في تحليل المكونات الرئيسية الاعتيادي يتم ايجاد المكونات أو المركبات ثم ايجاد قيم التحميلات والجذور المميزة أما انحدار المكونات الرئيسية فتكون فيه المتغيرات المستقلة على شكل مكونات رئيسية نقوم بدراسة تاثيرها 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 اوزان حصينة على نموذج مقدر بطريقة المربعات الصغرى متغيراتها المستقلة عبارة عن المكونات الرئيسية للنموذج الاصلي. لمقارنة 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 وايضاً في حالات وجود 5% أو 6% و10% و 20% و 30% و 40% شواذ في البيانات في المتغير المعتمد Y وتم تطبيق إنحدار المربعات الصغرى الاعتيادية بعد تحويل المتغيرات الى المركبات الرئيسية ومن ثم تطبيق المربعات الصغرى الموزونة بدوال أوزان من مقدرات M ومنها Huber و Hampel و Bisquare كون هذه المقدرات تكون حصينة ضد الشواذ في قيم المتغير المعتمد (Y- Y Outliers). ولكون القيم الجاذبة (X-leverage) أكثر خطورة على النموذج من القيم الشاذة في المتغير المعتمد ولهذا فإنه من المتوقع أن وجودها في مشاهدات المتغيرات المستقلة سيؤثر على قيم حد الخطأ كونها قد تسحب الأنموذج بإتجاهها، وبناءاً على ذلك ولأجل تغطية كل الإحتمالات قمنا بتجريب تطبيق النماذج المقترحة 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 في المتغيرات المستقلة وكما هو مبين من الجداول (1-4). ولأجل المقارنة بين أداء كل طريقة من الطرق المستخدمة في حالات عدم وجود شواذ وحالتي وجود قيم شاذة في المتغيرين المعتمد والمتغيرات المستقلة تم إستخدام 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 Median Absolute Percentage Error MDAPE، 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 (2) : مقارنة كفاءة الطرق لنتائج المحاكاة لخمسة متغيرات ولعينة بحجم 100

عند وجود 40% شواذ في المتغير المعتمد وعينة بحجم n = 100 وخمسة متغيرات مستقلة m =5						
معيار المقارنة / الطريقة	Residual standard error	R-Square	MSE	RMSE	MADE	MDAPE
Least Squares	10.7801703	0.03307137	109.2393	10.45176	2.0125	111.3814
LSPCRRobHuber	2.9113579	0.04754012	110.2361	10.49934	1.4140	101.9624
LSPCRRobHampel	1.3081563	0.11396729	111.0857	10.53972	1.1610	100.7380
LSPCRRobBisquare	0.9820237	0.24543735	110.4490	10.50947	1.3831	100.1905
التفسير	نلاحظ أنه وبوجود 40% شواذ في المتغير المعتمد أن مقدر LSPCRRobBisquare قد حقق أفضل إنجاز تلاه مقدر المربعات الصغرى Least Squares من خلال اعلى قيمة لمعامل التحديد واقلها لمقاييس الكفاءة					

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

عند وجود 5% شواذ في المتغير المعتمد وعينة بحجم n = 200 وتسعة متغيرات مستقلة m=9						
معيار المقارنة / الطريقة	Residual standard error	R-Square	MSE	RMSE	MADE	MDAPE
Least Squares	3.6628278	0.04163720	12.74549	3.570083	0.8813	111.6284
LSPCRRobHuber	1.2901977	0.06022365	13.01326	3.607390	0.8111	99.5174
LSPCRRobHampel	1.0514110	0.06119842	13.16886	3.628893	0.7651	99.6844
LSPCRRobBisquare	0.8955774	0.07917774	13.07341	3.615717	0.7823	99.4572
التفسير	نلاحظ أنه وبوجود 5% شواذ في المتغير المعتمد أن مقدر LSPCRRobBisquare قد حقق أفضل إنجاز تلاه مقدر المربعات الصغرى Least Squares وذلك من خلال اعلى قيمة لمعامل التحديد واقلها لمقاييس الكفاءة					

جدول (4) : تفسير النتائج لكل تجربة المحاكاة التي بلغت 432 محاولة

نسب التلويث / N			0%	5% or 6%	10%	20%	30%	40%	الخلاصة (الغالبية)
			عندما يوجد في النموذج ثلاثة متغيرات مستقلةm = 3						
Contamination with Outliers and Leverage Points	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 بدالة Bisquare قد حققت أفضل النتائج مقارنة بالطرق الاخرى						LSPCRRobBisquare
		100							
		200							
	X-leverage Points	50	لنموذج بثلاثة متغيرات توضيحية ولكل أحجام العينات وكل نسب التلويث للمتغيرات المستقلة أن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية قد حققت أفضل النتائج مقارنة بالطرق الأخرى وتليها طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بدالة Bisquare						Least Squares
		100							
		200							LSPCRRobBisquare
			عندما يوجد في النموذج خمسة متغيرات مستقلة m = 5						
Contamination with Outliers and Leverage Points	Y-outliers	50	نموذج بخمسة متغيرات توضيحية ولكل أحجام العينات وكل نسب التلويث للمتغير المعتمد وكذلك للنموذج بدون شواذ أن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بدالة Bisquare قد حققت أفضل النتائج مقارنة بالطرق الاخرى						LSPCRRobBisquare
		100							
		200							
	X-leverage Points	50	لنموذج بخمسة متغيرات توضيحية ولكل أحجام العينات وكل نسب التلويث للمتغيرات المستقلة أن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية قد حققت أفضل النتائج مقارنة بالطرق الأخرى وتليها طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بدالة Bisquare						Least Squares
		100							LSPCRRobBisquare
		200							Least Squares
			عندما يوجد في النموذج تسعة متغيرات مستقلة 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 الاعتيادية للمركبات الرئيسية والموزونة بدالة Bisquare قد حققت أفضل النتائج مقارنة بالطرق الاخرى						LSPCRRobBisquare
		100							
		200							
	X-leverage Points	50	لنموذج بتسعة متغيرات توضيحية ولكل أحجام العينات وكل نسب التلويث للمتغيرات المستقلة أن طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية والموزونة بدالة Bisquare قد حققت أفضل النتائج مقارنة بالطرق الأخرى وتليها طريقة المربعات الصغرى الاعتيادية للمركبات الرئيسية						LSPCRRobBisquare
		100							Least Squares
		200							LSPCRRobBisquare

عند وج... m =				
معيار المقارنة / الطريقة	Residual standard error	R-Square	MSE	RMSE
Least Squares	5.3680415	0.1382662	26.51060	5.148845
LSPCRRobHuber	1.6737511	0.1057055	28.38303	5.327573
LSPCRRobHampel	1.2462578	0.1313538	29.13013	5.397233
LSPCRRobBisquare	0.9385016	0.2338350	28.21225	5.311521
التفسير	...واذ في المتغير المعتمد أن مقدر LSPCRRobBisquare قد حقق أفضل إنجاز تلاه مقدر المربعات الصغرى Least Squares وذلك من خلال اعلى قيمة لمعامل التحديد واقلها لمقاييس الكفاءة			

5- الجانب التطبيقي Application Part

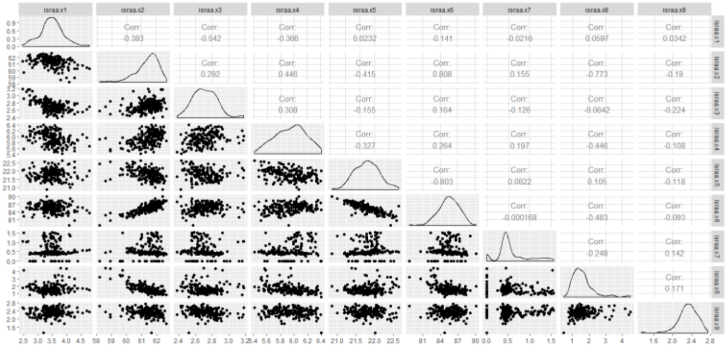
لتطبيق الطرق المقترحة تم جمع بيانات من معمل اسمنت بادوش للفترة من 2008- 2014، مع استبعاد أشهر الصيانة التي توقف المعمل فيها عن الانتاج. يتكون الاسمنت من بعض المواد الأساسية المتوافرة بصورة طبيعية من الحجر والرمل والحصى وبعض الإضافات الاخرى أثناء عملية التصنيع والتي تتضمن مواد تعمل على التغلب على بعض 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 إختبار 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.

الشكل (3): شكل يمثل مصفوفة الارتباط



نلاحظ من الشكل (3) 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.

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

Eigenvalues with INTERCEPT

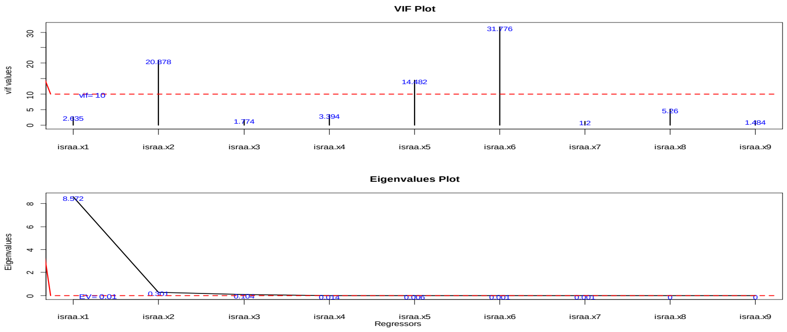
	Intercept	**israa.x1**	**israa.x2**	**israa.x3**	**israa.x4**	**israa.x5**	**israa.x6**	**israa.x7**	**israa.x8**	**israa.x9**
Eigenvalues:	9.5689	0.3026	0.1056	0.0142	0.0062	0.0013	0.0008	0.0005	0.000	0.000
Condition Indeces:	1.0000	5.6232	9.5199	25.9951	39.2908	86.1575	109.7136	143.0268	1175.945	1721.596

ونلاحظ من الجدول (5) أعلاه أن قيمة محدد مصفوفة المعلومات صغير جداً، وقريب من الصفر (0.0014) وكذلك فإن قيمة إحصاءة مربع كاي لفارار (Farrar and Glauber) كبيرة جداً، كما أن قيمة العدد الشرطي Condition Number كبيرة جدا، 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. وللتأكد من وجود المشكلة تم إجراء إختبار فارار – كلاوبر لمعرفة مصدر المشكلة.

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 الالمنيوم 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 (3) 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, 2008), (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, 2003), (Imdadullah et. al., 2016).



الشكل (4): شكل يبين رسم معامل تضخم التباين والجذور المميزة

والشكل (4) أعلاه يبين ما أكدته المؤشرات العددية ويتضح من رسمي VIF ورسم الجذور المميزة بان هنالك ثلاثة متغيرات تعاني من مشكلة تعدد العلاقة الخطية.

2-5 إنحدار المكونات الرئيسية

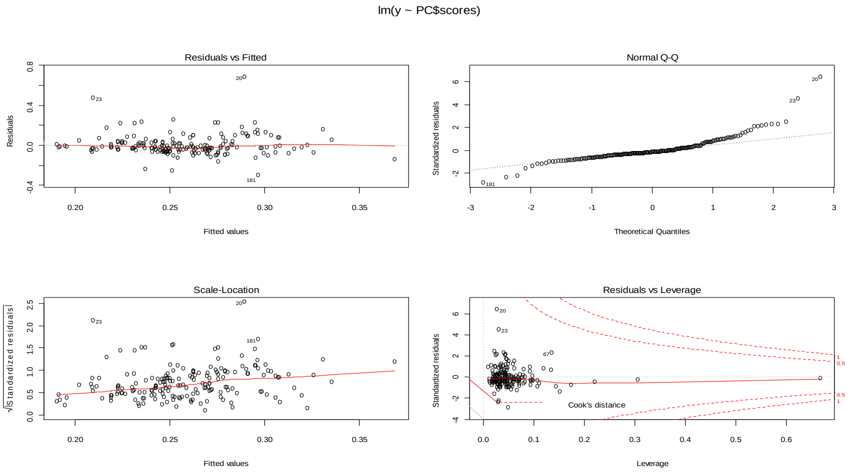
بعد التأكد من وجود مشكلة تعدد العلاقة الخطية تمت مقارنة أداء الطرق المقترحة على بيانات معمل الأسمنت بعد تحويل المتغيرات الى الصيغة القياسية (لإختلاف وحدات قياسها) ومن ثم مقارنة كفاءة الطرق المقترحة في تقدير نموذج الانحدار لتمدد الأسمنت على متغيرات الدراسة التسعة. وقد كانت مقاييس الكفاءة كما هو مبين في الجدول أدناه والتي يتضح منها تفوق طريقة المربعات الصغرى الموزونة بأوزان Bisquare من مقدرات M:

	Residual standard error	R-Square	Adjusted R-Square	MSE	RMSE	MADE
Least Squares	0.1063100	0.08037254	0.03361182	0.010697	0.103428	0.04302
LSPCRRobHuber	0.1063100	0.08037254	0.03361182	0.010697	0.103428	0.04302
LSPCRRobHampel	0.1063100	0.08037254	0.03361182	0.010697	0.103428	0.04302
LSPCRRobBisquare	**0.1055723**	**0.08099154**	**0.03426230**	0.010698	**0.103429**	**0.04272**

ونلاحظ من رسم البواقي ضد القيم المقدرة في الشكل (2) أدناه لنموذج انحدار المتغير المعتمد y (تمدد الإسمنت) الموزون بدالة وزن Bisquare من مقدرات M (الرسوم للدوال الأخرى مطابقة لها) ضد المكونات الرئيسية المقابلة للمتغيرات المستقلة أنه لاتوجد 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 (20, 23, 181) حيث

نلاحظ إنحراف النموذج عن التوزيع الطبيعي كما هو واضح في أطراف الرسم. وبالنظر إلى رسم جذر الاخطاء القياسية ضد القيم المقدرة أن النقاط متوزعة حول الخط بشكل منتظم الامر الذي يدل على عدم وجود مشكلة عدم تجانس التباين بين الأخطاء. وأخيراً وبالنظر الى رسم قيم الجذب Leverage (مسافات كوك Cook's Distance) ضد الأخطاء القياسية نجد أن الرسم قد أفرز قيمتين شاذتين في المتغير المعتمد (20, 23) وهما نفس القيمتين اللتان تم تشخيصهما في رسم QQ-Plot، ولايبدو من الرسم وجود أية قيم جاذبة (Leverage Points) في البيانات.

الشكل (2): الرسوم التشخيصية لمشاكل نموذج الانحدار الخطي LSPCRRobBisquare



6-الاستنتاجات

1. لم تتمكن مقدرات M من إعطاء صورة واضحة حول كفاءة المقدرات عند وجود قيم شاذة (قيم جاذبة أو مخلة) في بيانات المتغيرات المستقلة X-Leverage Points.
2. بقيت الأسلوب المقترح لطريقة المربعات الصغرى للمكونات الرئيسية LSPCRRobBisquare متفوقاً في حالة وجود مشكلتي الشواذ في المتغير المعتمد Y-outliers وتعدد العلاقة الخطية في البيانات في آنٍ واحد.
3. إن عدم ثبات سلوك المقدر المقترح LSPCRRobBisquare مقابل طريقة المربعات الصغرى الاعتيادية التي تفوقت في بعض الحالات يعد الى ان عملية التقدير قد تمت على مرحلتين، الأولى تم فيها التغلب على مشكلة تعدد العلاقة الخطية من خلال إيجاد المركبات الرئيسية المقابلة للمتغيرات المستقلة وهذا لصالح طريقة المربعات الصغرى، والثانية إستخدام أسلوب مقدرات M الحصين الذي تقتصر حصانته على وجود قيم شاذة في المتغير المعتمد، وبالمحصلة كانت مقدرات المربعات الصغرى الاعتيادية باستخدام المكونات الرئيسية كمتغيرات مستقلة بدلاً من المتغيرات الاصلية المرتبطة خطياً.

Reference

1. Al-Rawi, Rawiya Emad Karim, (2017AD), "Using fuzzy ordinal functions in the impartial estimation of the parameters of the simple linear regression model," master's thesis, College of Computer Science and Mathematics, University of Mosul-Iraq.
2. Student, Diaa Majeed, Student, Bashar Abdel Aziz and Student, Ali Diaa, (2011), "Using some strong statistical methods to determine the expected achievements in the jumping and jumping competition for men in the London and Rio de Janeiro Olympic Games (2012, 2016)", Al-Qadisiyah Conference , published in the conference proceedings, Al-Qadisiyah University, Iraq.
3. Adam Bremah Suleiman Mastour and Amal Al-Sir Al-Khader Abdel-Rahim, (2016 AD), "Treatment of the Linear Overlap Problem Using Principal Components Analysis (by application to fuel consumption in cars)", Sudan University of Science and Technology - College of Science - Department of Applied Statistics.
4. Jibril, Muhammad Suleiman Muhammad, (2014 AD), "Polylinearity, its causes and effects, and treatment with character gradient and main component gradient with application on hypothetical data," Sudan University of Science and Technology - College of Graduate Studies.
5. Asteriou, D., and Hall, S. G., (2007), "Applied econometrics: A modern approach using EViews and Microfit", Palgrave Macmillan, New York, [p496].
6. Boiroju, N.K., and Reddy, M.K., (2012), "A Graphical Method for Model Selection", Pakistan Journal of Statistics & Operation Research, pp. 767-776.

7. Belsley, D. A., Kuh, E., and Welsch, R. E., (2004), “Diagnostics: Identifying Influential Data and Sources of Collinearity”, John Wiley & Sons, New York.
8. Chatterjee, S., and Hadi, A. S., (2012), “Regression Analysis by Example", 4th. Ed., John Wiley and Sons, New York.
9. Curto, J. D., and Pinto, J. C., (2011), “The corrected VIF (CVIF)”, Journal of Applied Statistics, 38(7):1499–1507.
10. Farrar, D. E., and Glauber, R. R., (1967), “Multicollineanty in regression analysis: The problem revisted”, The Review of Economics and Statistics, 49:92–107, [p495, 496, 498].
11. Greene, W. H., (2003), “Econometric Analysis”, Prentic-Hall, New Jersey, 5th edition.
12. Hadi, A. S., and Ling .R. F., (1998), “Some cautionary notes on the use of principal components regression”, American statistician, 52, 15-19.
13. Huber, P.J., (1964), “Robust Estimation of a Location Parameter”, Annals of Mathematical Statistics.USA, 35:73- 101.
14. Huber, P. J., (1973), “Robust regression, Asymptotic, conjectures, and Monte Carlo”, Ann. Statist., Vol. 1, no.5, 799-821.
15. Hubert, M., and Verboven, K., (2003), “Robust PCR methods for Partial Least Squares Regression”, Journal of chemometrics 17,537-549.
16. Hubert, M., and Verboven, S., (2003), “A robust PCR methods of High – dimensional regressors”, Journal of chemometrics 17,438-452.
17. 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).
18. Jolliffe, I.T., (1982), “A note on the use of principal components in Regression”, Appl. Statist., 31, 300–303.
19. Kendall, M. G., (1957), “A Course in Multivariate Analysis”, Charles Griffin & Company, London.
20. 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.
21. Kutner, M.H., Nachtsheim, C.J., and Neter, J., (2004), “Applied Linear Regression Models”, McGraw Hill Irwin, 4th. Ed.
22. Maddala, G. S., (1992), “Introduction to econometrics”, Macmillan, New York.
23. Makridakis, S., and Hibon, M., (1995), “Evaluating accuracy (or error) Measures”, INSEAD Working Papers Series 95/18/TM. Fontainebleau, France.
24. Marquardt, D.W., (1970), “Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation”, Technimetrics, 12(3):591–612.
25. 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.
26. Montgomery, D. C., Peck ,E. A., and Vining, G. C., (2001), “Introduction to Linear Regression Analyses”, 3rd, edition, John Wiley & Sons, New York–USA.
27. Sarwar, A., and Sharma, V., (2014), “Comparative analysis of machine learning techniques in prognosis of type II diabetes”, AI & society, 29(1), 123-129.
28. 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.
29. Woschnagg, E., and Cipan, J., (2004), “Evaluating Forecast Accuracy”, UK Ö konometrische Prognose, University of Vienna,Department of Economics.

Iraqi Journal of Statistical Sciences, Vol.18, No. 1, 2021 , pp (1-13)

Solid weighted methods to detect and deal with anomalies in the estimation of the principal components regression model

Esraa Naguib Saeed Al-Sarraf Bashar Abdel Aziz Al-Talib

Abstract

This paper aims to propose an approach to deal with the problem of Multi-Collinearity between the explanatory variables and outliers in the data by using the method of Principal Component Regression, and then using a robust weighting functions for the objective function has been used to deal with the presence of outliers in the data, and in order to verify the efficiency of the estimators, an experimental study was conducted through the simulation approach, and the methods were also applied to real data collected from the files of Badoush Cement Factory in Nineveh Governorate for the period from (2008-2014) with nine explanatory variables representing the chemical properties of cement and a dependent variable representing the physical properties of cement (hardness). The data was tested whether it was suffer from multi-collinearity problem and then the least squares using principal components as an explanatory variables and the model was estimated, and it was found that the variables suffer from Multi-Collinearity problem, and the treatment was done by applying principal component regression weighed by robust weights due to the presence of outlying values in the data in addition to the collinearity problem.

Key Words: Principal Component Regression, outliers, Leverage Points, Weighted Least Squares, Multi-Collinearity