

مقارنة بين طريقتي انحدار المكونات الرئيسية والمربعات الصغرى الجزئية مع التطبيق على معمل إسمنت كركوك

ظافر رمضان مطر البدراني ، تقوى عبدالسلام طه المولى

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

thafer63@yahoo.com

Staralmula@yahoo.com

الملخص

تظهر في العديد من مجالات تطبيق تحليل الانحدار وجود حالة أكثر من متغير معتمد واحد يتأثر بمجموعة من المتغيرات التوضيحية وبنفس الوقت يكون عدد المشاهدات قليل نسبياً مقارنة بعدد المتغيرات ، وتظهر هنا مشكلة عدم توفر العديد من فروض تحليل الانحدار المتعدد إضافة إلى بروز مشكلة تعدد العلاقات الخطية بين المتغيرات التوضيحية بالإضافة إلى الارتباط بين المتغيرات التوضيحية والمتغيرات المعتمدة وانعكاسات ذلك على مقدرات الانحدار . وفي هذا البحث تم التعامل مع مشاكل من هذا النوع تتعلق بمتغيرات معمل إسمنت كركوك ، لقد تم استخدام طريقتي المكونات الرئيسية PC والمربعات الصغرى الجزئية PLS لمعالجة المشاكل أعلاه ، باعتبار أن الطريقة الأولى من الطرائق الشائعة في حل مشكلة تعدد العلاقات الخطية بين المتغيرات التوضيحية والثانية باعتبارها من الطرائق التي لها منهجية مختلفة في استخلاص المكونات المعتمدة على معالجة الارتباط الموجود فيما بين المتغيرات التوضيحية وكذلك بينها وبين المتغيرات المعتمدة . من خلال التحليل الإحصائي تم التوصل إلى أن طريقة المربعات الصغرى الجزئية PLS تمكنت من توفير أنموذج انحدار أمثل لجميع المتغيرات المعتمدة ، بالإضافة إلى ذلك تفوقها في القدرة على التنبؤ بالقيم المستقبلية لكل المتغيرات المعتمدة وأيضاً من حيث تخفيض البعد مقارنة بطريقة المكونات الرئيسية PC . تم استخدام البرنامج الإحصائي الجاهز (Minitab, Version, 16.1) في التحليل الإحصائي لبيانات البحث .

1. المقدمة

تعاني من المشاكل أعلاه خاصة بمعمل إسمنت كركوك ، حيث تمثلت المتغيرات التنبؤية بكل من الطاقة الكهربائية والنفط الأسود والأحجار والتراب مع وجود أكثر من متغير معتمد واحد والمتماثلة بكل من معامل الانبعاث الحراري الناتج من العملية الإنتاجية للإسمنت وكمية إنتاج الإسمنت وكمية إنتاج الكلنكر للفترة (1984-2013) .

2. مشكلة البحث

من أهم المشاكل التي تؤدي إلى حدوث انتهاك لواحد أو أكثر من فروض التحليل الإحصائي لطريقة OLS هي وجود مشكلة تعدد العلاقات الخطية وكذلك وجود أكثر من متغير معتمد واحد تعتمد وتتأثر بنفس المتغيرات التوضيحية وبالتالي وجود ارتباط بين المتغيرات المعتمدة والمتغيرات التوضيحية بالإضافة إلى الارتباط الموجود بين المتغيرات التوضيحية فيما بينها إضافة إلى مشكلة قلة عدد المشاهدات المتوفرة .

3. هدف البحث

يهدف البحث إلى التعرف على أوجه التشابه والاختلاف بين طريقتين لمعالجة مشكلة تعدد العلاقات الخطية فضلاً عن مشكلة وجود أكثر من متغير معتمد واحد ومشكلة قلة عدد المشاهدات ، وهما طريقة المكونات الرئيسية PC وطريقة المربعات الصغرى الجزئية PLS حيث تنشئ هاتين الطريقتين متغيرات جديدة تدعى بالمكونات ولكن بطرائق مختلفة توظف في تطبيق انحدار المربعات الصغرى الاعتيادية OLS، وفي هذا البحث يتم التعرف على طرائق إنشاء هذه المكونات

تحليل الانحدار عبارة عن وسيلة إحصائية يستخدم لتحليل العلاقة بين متغير تنبؤي واحد أو أكثر ومتغير تابع وبصورة عامة يستخدم لتحقيق هدفين ، الأول البحث عن دالة رياضية تقيد في معرفة كيفية ارتباط المتغيرات فيما بينها أما الهدف الثاني معرفة مدى دقة التنبؤ وقوة العلاقة بين هذه المتغيرات [1] ، ولتحليل انحدار المربعات الصغرى الاعتيادية OLS عدة فروض من الواجب توفرها ، وعدم توفر أحدها أو بعضها يقود إلى عدم الوثوق بصحة التقديرات الإحصائية التي يتم التوصل إليها والمشكلة الأكثر شيوعاً هي مشكلة تعدد العلاقات الخطية بين المتغيرات التوضيحية ، وبعض طرائق علاج هذه المشكلة يستوجب تحويل المتغيرات التنبؤية المرتبطة إلى متغيرات جديدة غير مرتبطة مستقلة خطياً تسمى عادةً بالمكونات ، وهناك طريقتان تقوم بهذه العملية ولكل منها أسلوب مختلف عن الأخرى ، الأولى طريقة المكونات الرئيسية PC حيث تعتبر هذه الطريقة من الطرائق الشائعة الاستخدام لمعالجة المشكلة أعلاه ، أما الثانية طريقة المربعات الصغرى الجزئية PLS التي تتميز عن الطريقة الأولى في إمكانيةها العالية على معالجة المشكلة أعلاه مع معالجة مشاكل أخرى في نفس الوقت مثل تأثير مشكلة صغر حجم عينة البحث وكذلك مشكلة وجود أكثر من متغير معتمد واحد في البحث حيث تعتمد فكرة احتساب المكونات لهذه الطريقة على مراعاة الارتباط الموجود بين المتغيرات التوضيحية والمتغيرات المعتمدة إضافة إلى الارتباط الموجود بين المتغيرات التوضيحية فيما بينها ، بينما طريقة PC تعتمد فكرة احتساب المكونات على معالجة الارتباط الموجود فقط بين المتغيرات التوضيحية فيما بينها [2] . وفي هذا البحث تم التطبيق على بيانات

الفرض السابع : أن تكون المتغيرات التنبؤية غير مرتبطة ببعضها البعض Multicollinearity No وفي الواقع يواجه الباحث هذا الفرض عندما يكون النموذج المدروس متضمناً أكثر من متغير تنبؤي واحد ، حيث يجب أن لا يكون بينهم أي تعدد خطي وبالتالي يمكن التعرف على أثر كل متغير تنبؤي على المتغير المعتمد بشكل منفصل . أما فيما يتعلق بخواص المتغير المعتمد أولها أن يكون التوزيع لهذا المتغير توزيعاً طبيعياً وتوقعه أو متوسطه يعطى في حالة الانحدار الخطي البسيط بالصيغة $\bar{Y} = E(Y_i) = \beta_0 + \beta_1 X_i$. أما تبين هذا المتغير فيعطى بالصيغة $\text{var}(Y_i) = E[Y_i - E(Y_i)]^2 = \sigma_e^2$. ويمكن وضع الخواص أعلاه بالشكل المختصر $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma_e^2)$. ويواجه العديد من الباحثين مشكلة انتهاك لواحد أو أكثر من فروض التحليل الإحصائي أعلاه وذلك بسبب توفر عدة مشاكل مجتمعة تؤدي إلى حدوث المشكلة أعلاه ، وأبرزها مشكلة تعدد العلاقات الخطية بين المتغيرات التوضيحية ومشكلة صغر حجم العينة المدروسة حيث أن اختيار حجم العينة المناسب لمشكلة البحث هي طريقة على الأغلب فعالة لتحقيق تقديرات تكون دقيقة وموثوق بها لاتخاذ قرارات حكيمة أو تحقيق الهدف العلمي [6] وفي بعض البحوث يتعدى الحصول على الحجم المناسب لعينة البحث كما في التجارب النووية وفي هذه الحالة تظهر مشكلة عدم توفر فرض التوزيع الطبيعي للمتغير المعتمد وكذلك ليوافق أنموذج الانحدار المقدر ويشير [7] إن شرط إجراء اختبار التوزيع الطبيعي للعينة مرتبط بحجم العينة ، إذ يعتبر شرطاً ضرورياً في حالة العينات الصغيرة أما في حالة العينات الكبيرة يمكن التخلي عنه وذلك لأن التوزيعات الاحتمالية تؤول إلى التوزيع الطبيعي في حالة العينات التي يزيد حجمها عن 30 مشاهدة ، وكذلك مشكلة وجود أكثر

من متغير معتمد واحد حيث في الواقع العملي قد يواجه الباحث عند دراسة ظاهرة معينة وجود أكثر من متغير معتمد واحد تعتمد وتتأثر بنفس المتغيرات التوضيحية وبالتالي وجود ارتباط بين المتغيرات المعتمدة والمتغيرات التوضيحية بالإضافة إلى الارتباط الموجود بين المتغيرات التوضيحية فيما بينها ، ومثال على ذلك ظاهرة التلوث البيئي [8] ، وعند وجود مثل هذه المشاكل تصبح الأخطاء القياسية للمعاملات متضخمة وبناءً على ذلك يكون التحليل التام للمعاملات مشكوك في صحته . وبصورة دقيقة ربما تملك المعاملات إشارات غير صحيحة وكذلك حدوث تغيرات ضخمة في حجم قيم المعاملات السابقة عند إضافة متغير جديد إلى الأنموذج ، بالإضافة إلى ذلك تصبح اختبارات t (t- tests) لاختبار معنوية المعاملات الخاصة بالمتغيرات التوضيحية في أنموذج الانحدار مشكوك في صحتها [9]. يمكن الحكم على الأنموذج المقدر إنه أنموذج أمثل ويمثل البيانات خير تمثيل ويمكن الاعتماد عليه في التنبؤ بالقيم المستقبلية للمتغير المعتمد في حالة توفر كافة فروض التحليل الإحصائي لطريقة OLS في ذلك الأنموذج [10] .

وانعكاسات ذلك على كفاءة نماذج الانحدار المستخلصة في تمثيل بيانات البحث وقدرتها على التنبؤ وعلى تخفيض البعد .

4. الجانب النظري

1-4 الانحدار ومشاكل التطبيق

تحليل الانحدار من أكثر الطرائق الإحصائية استعمالاً في مختلف العلوم لأنه يستخدم لوصف طبيعة العلاقة بين المتغيرات فيما إذا كانت طردية أو عكسية ، خطية أو غير خطية [1] . ويشير [3] إلى أن استعمالات تحليل الانحدار هي وصف البيانات Data Description وتقدير المعلمات Parameter- Estimation والتنبؤ Prediction والسيطرة Control . والانحدار يصنف إلى :

1- انحدار خطي Linear Regression قد يكون بسيط Simple Linear Regression ويضم متغير تنبؤي واحد فقط .

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

أو انحدار خطي متعدد Multiple Linear Regression ويضم عدة متغيرات تنبؤية ، ومعادلته بصورة عامة لـ m من المتغيرات التنبؤية تكون بالصيغة التالية :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i \quad (2)$$

2- انحدار غير خطي Linear Regression or Curvilinear Non Regression ويقسم أيضاً إلى قسمين ، انحدار غير خطي بسيط Simple Curvilinear Regression ويضم أيضاً متغير تنبؤي واحد فقط ، ومعادلته بصورة عامة تكون بالصيغة التالية :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \dots + \beta_m X_{i1}^m + \varepsilon_i \quad (3)$$

أو انحدار غير خطي متعدد Multiple Curvilinear Regression ويضم عدة متغيرات تنبؤية ومعادلته لمتغيرين تنبؤيين مثلاً هي :

$$Y_i = \beta_{00} + \beta_{10} X_{i1} + \beta_{01} X_{i2} + \beta_{11} X_{i1} X_{i2} + \beta_{20} X_{i1}^2 + \dots + \beta_{m0} X_{i1}^m + \beta_{m1} X_{i1}^m X_{i2} + \varepsilon_i \quad (4)$$

في العلاقة الخطية المدروسة هنالك خواص أو فروض تتعلق بالخطأ العشوائي ε_i وأخرى تتعلق بالمتغير المعتمد Y ، والفروض التي تتعلق بالخطأ العشوائي ε_i يمكن إجمالها كما يلي : [4 , 5]

الفرض الأول: ε_i متغير عشوائي حقيقي . ، الفرض الثاني :

$$E(\varepsilon_i) = 0$$

الفرض الثالث: $\text{var}(\varepsilon_i) = E[\varepsilon_i - E(\varepsilon_i)]^2 = \sigma_e^2$ ، الفرض

الرابع: ε_i يتوزع طبيعياً .

ويمكن وضع الفروض الأربعة السابقة بالشكل المختصر $N(0, \sigma_e^2)$.

$$\varepsilon_i \sim$$

الفرض الخامس:

$$\text{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \quad \forall i \neq j \quad (i, j = 1, 2, \dots, n)$$

الفرض السادس: $E(\varepsilon_i X_i) = 0$.

2-4 تشخيص مشكلة تعدد العلاقات الخطية :

تعرف مشكلة تعدد العلاقات الخطية على أنها درجة عالية من الارتباط أو الاعتماد الخطي Linear Dependency بين اثنين أو أكثر من المتغيرات التوضيحية في نموذج الانحدار المتعدد، وقد يكون السبب في حدوثها هو أن بعض المتغيرات التوضيحية ربما تقيس نفس المفاهيم أو الظاهرة [10]. بصيغة رياضية تحدث المشكلة أعلاه عندما تكون مصفوفة المتغيرات التوضيحية X ذات مرتبة غير كاملة Non Full Ranked حيث نقشل عندئذٍ طريقة المربعات الصغرى الاعتيادية OLS في توفيق أنموذج انحدار أمثل ليمثل بيانات الظاهرة المدروسة [2]. ولغرض تشخيص مشكلة تعدد العلاقات الخطية في أنموذج انحدار يحتوي على متغيرين توضيحيين فأكثر ، فإنه يتطلب حساب الآتي :

(a) القيم والمتجهات المميزة The Values and Eigen Vectors لمصفوفة المعلومات $X'X$:

تعد عملية حساب الجذور والمتجهات المميزة من طرائق الكشف عن تعدد العلاقات الخطية ، فإذا كانت قيمة أحد الجذور مساوية للصفر دل ذلك على وجود علاقة خطية تامة ، وبالعكس عندما تساوي الواحد الصحيح يدل هذا على انعدام أي علاقة خطية. أي إنه كلما اقتربت قيمة الجذر المميز من الصفر، دل ذلك على قوة العلاقة الموجودة بين المتغيرات التوضيحية . وكذلك فإن من ملاحظة المتجه المميز التابع له نستطيع أن نستدل على أهم المتغيرات المؤثرة فيه ، حيث تكون قيمتها كبيرة مقارنة بالقيم الأخرى الموجودة في المتجه المميز [11] .

(b) مصفوفة الارتباط Correlation Matrix للمتغيرات التوضيحية: يعد الارتباط البسيط من أبسط مقاييس الكشف عن التداخلات الخطية . فإذا كان $r(X_i, X_j) = \pm 1$ دل ذلك على وجود علاقة خطية تامة بين X_i و X_j ، حيث R هي مصفوفة الارتباط بين المتغيرات التوضيحية وقد أقتراح Gunst and Mason في عام 1980 فحص الارتباطات بين X_i و X_j التي يكون فيها $r(X_i, X_j) > 0.7$ فإذا أمكن التعبير عن الشكل الانتشاري بخط مستقيم معنى ذلك هناك علاقة خطية بين X_i و X_j ، وإذا كان معامل الارتباط بين X_i و X_j سالبا معنى ذلك هناك علاقة خطية تأخذ الشكل الآتي: [6]

$$X_i = -X_j \quad \text{or} \quad X_i + X_j = 0$$

أما إذا كان الارتباط البسيط بين X_i و X_j موجبا فإن العلاقة الخطية تأخذ الشكل الآتي :

$$X_i = X_j \quad \text{or} \quad X_i - X_j = 0$$

وفي حالة ارتباط أكثر من متغيرين بعلاقة خطية فإنه ليس من الضروري أن يكون $r(X_i, X_j)$ قريباً من الواحد الصحيح أو حتى كبير ولهذا السبب فإن مقياس الارتباط وحده غير كافٍ في الكشف عن التداخلات الخطية .

(c) العدد الشرطي Conditional Number والدليل الشرطي (CI) Conditional Index :

لقد طوّر المفهوم لمقياس العدد الشرطي من قبل Belsley, Kuh and Welsch في عام 1980 ليعطي مقياس الدليل الشرطي CI والموضحان في أدناه : [13 , 14]

1. العدد الشرطي Conditional Number ، ويأخذ الصيغة التالية :

$$\phi = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \dots (5)$$

حيث إن λ_{\max} و λ_{\min} تمثلان القيمة المميزة الأكبر والأصغر للمصفوفة $X'X$ على التوالي . لقد أقتراح هذا المقياس من قبل Belsley, Kuh and Welsch في عام 1980 . فإذا كانت القيمة لهذا المقياس كبيرة دل ذلك على وجود تداخل خطي بين المتغيرات التوضيحية ويعتبر هذا المقياس ضعيفاً ، ومما يجدر الإشارة إليه إن قيمة هذا المقياس تساوي الواحد الصحيح في حالة المصفوفة المتعامدة.

2. الدليل الشرطي (CI) Conditional Index ، ويأخذ الصيغة التالية :

$$CI_j = \sqrt{\frac{\lambda_{\max}}{\lambda_j}} \dots (6)$$

يستفاد من الدليل الشرطي في قياس درجة التعدد الخطي بالاعتماد على قيمة الجذر المميز j . حيث أقتراح كل من Belsley, Kuh and Welsch في عام 1980 أنه إذا كانت قيم الدليل الشرطي CI_j بحدود العشرة يشير ذلك على إن درجة التعدد الخطي ضعيفة ، أما إذا كانت $(30 \leq CI_j \leq 100)$ يشير ذلك على إن درجة التعدد الخطي متوسطة أما إذا كانت $(CI_j > 100)$ فهذا يعني وجود تعدد خطي عالٍ .

(d) نسبة التباين The Variance Proportion :

تحلل نسب التباين التغير في $\text{Var}(\hat{\beta})$ بالاعتماد على المصفوفة $X'X$ والتي يمكن كتابتها بالصيغة VDV' إذ أن V هي مصفوفة متعامدة Orthogonal Matrix أعمدتها المتجهات المميزة للمصفوفة $X'X$ وإن D هي مصفوفة قطرية Diagonal Matrix قطرها الرئيسي يمثل القيم المميزة للمصفوفة $X'X$ ، وعليه فإنه يمكن كتابة المصفوفة $\text{Var}(\hat{\beta})$ بالصيغة التالية :

$$\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \sigma^2 (VDV')^{-1} \dots (7)$$

حيث يمكن إيجاد كل عنصر من عناصر نسب التباين p_j لتباين المعلمة المقدرة من خلال استخدام الصيغة التالية :

$$p_j = \frac{\sigma^2 V_{jj}^2}{\lambda_j} \dots (8)$$

$$\text{Var}(\hat{\beta}_j)$$

حيث أن $\text{Var}(\hat{\beta}_j)$ هو التباين للمعلمة المقدرة $\hat{\beta}_j$ ، $(i = 0, 1, \dots, m)$ والتي يتم حسابها باستخدام الصيغة التالية :

إن تحليل المكونات الرئيسية PCA هو طريقة إحصائية متعددة المتغيرات تستخدم بصورة شائعة لتقليل عدد المتغيرات التنبؤية وحل مشكلة التعدد الخطي في نفس الوقت. وهذا التحليل يعود تاريخه إلى جهد العالم Pearson عام 1901 الذي له عدة تحاليل إحصائية منتشرة خلال عدة عقود من السنوات [20]. ويشير [21] إن عملية تخفيض البعد لفضاء بعد m من المتغيرات التنبؤية بواسطة PCA يتحقق بواسطة بناء مكونات رئيسية عبارة عن تركيبات خطية من المتغيرات التنبؤية الأصلية وبعدد m ومن ثم يتم اختيار عدد من المكونات أقل من m كمغيرات تنبؤية للمتغير المعتمد. ويمكن إعادة صياغة معادلة الانحدار بدلالة المكونات الرئيسية الممثلة لمعظم التباين الكلي باستعمال طريقة المربعات الصغرى الاعتيادية OLS عندئذ يتم الحصول على تقديرات لمعاملات انحدار المكونات الرئيسية وخالية من أثر التعدد الخطي [16]. ولحساب المكونات الرئيسية هنالك حالتان: [3, 21] الأولى: إذا كانت كل المتغيرات التوضيحية X_i لها نفس وحدات القياس ففي هذه الحالة تكون معاملات المكونات الرئيسية a_{ij} هي قيم المتجهات المميزة لمصفوفة التباين والتباين المشترك V للمتغيرات التوضيحية، حيث أن v_{ij} تباين المتغير التوضيحي X_i ، v_{ij} التباين المشترك بين المتغير X_i والمتغير X_j . ثم يتم إيجاد المعادلة المميزة للمصفوفة V بالصيغة:

$$|V - \lambda I| = 0 \dots (10)$$

حيث λ الجذر المميز، I مصفوفة الوحدة وبأبعاد $m * m$ أي بعدد المتغيرات التوضيحية X_i . حيث تم طرح λ من القيم القطرية للمصفوفة V ثم جعل قيمة محدها تساوي صفراً، وشكل المعادلة المميزة هو متعددة الحدود في λ من الدرجة m أي يكون شكلها: $\lambda^m + C_{m-1}\lambda^{m-1} + \dots + C_1\lambda + C_0 = 0$ وعند حل هذه المعادلة يتم الحصول على m من الجذور المميزة $\lambda_1, \lambda_2, \dots, \lambda_m$ وهذه الجذور ترتب بحيث λ_1 هو أكبر قيمة ويليه λ_2 وهكذا أي أن، $\lambda_1 > \lambda_2 > \dots > \lambda_m$ حيث جميع الجذور المميزة للمصفوفة V موجبة، ثم يتم إيجاد المتجهات المميزة حيث إنه لكل جذر مميز λ_j متجه مميز a_j خاص يقابله وذلك باستخدام الصيغة:

$$(V - \lambda I)a = 0 \dots (11)$$

حيث إن 0 هو متجه صفري، ثم بعد ذلك يتم اختيار a_j بحيث تساوي قيم المتجه المميز القياسي أي بحيث يكون $a_j' a_j = 1$ ، وعند حل المعادلة أعلاه يتم الحصول على m من المعادلات الآتية ويتم حل هذه المعادلات وإيجاد قيم المتجه المميز a_j المقابل للقيمة المميزة λ_j باستخدام القيد $a_1^2 + a_2^2 + \dots + a_m^2 = 1$ وبذلك يكون المتجه المميز a_j الناتج متجه مميز قياسي بحيث $a_j' a_j = 1$. وعندئذ تكون الصيغة للمكون الرئيسي PC_j كما يلي:

$$\text{Var}(\hat{\beta}_i) = \sigma^2 \sum_{j=1}^{m+1} \left(\frac{v_{ij}^2}{\lambda_j} \right) \dots (9)$$

حيث أن σ^2 تباين الخطأ، v_{ij}^2 العنصر i من قيم المتجه المميز j في المصفوفة V . يستفاد من نسب التباين في التعرف على أي من $\hat{\beta}_i$ تأثرت بال λ_j وذلك عندما تكون النسبة أكبر من 0.50 [14, 15]. إن إيجاد نسب تحليل التباين ومن خلال قيم الدليل الشرطي يمكن الاستفادة منهما معاً في بيان درجة التعدد الخطي وتشخيص المتغيرات التوضيحية المسببة لمشكلة التعدد الخطي، فإذا كانت قيم الدليل الشرطي بحدود (5-10) وإن اثنين أو أكثر من نسب تحليل التباين أكبر من 0.50 فإنه يدل على إن الارتباط ضعيف، أما إذا كانت قيم الدليل الشرطي ($10 < CI \leq 30$) وإن اثنين أو أكثر من نسب تحليل التباين أكبر من 0.50 فهذا يدل على إن الارتباط من المعتدل إلى العالي، أما إذا كانت ($CI > 100$) وإن اثنين أو أكثر من نسب تحليل التباين أكبر من 0.50 فهذا يدل على إن التعدد الخطي على درجة كبيرة [16].

عامل تضخم التباين (Variance Inflation Factor (VIF) : إن وجود الارتباط بين المتغيرات التوضيحية يؤدي إلى زيادة أو تضخم الأخطاء المعيارية لمعاملات الانحدار [17] لذا يقوم العامل VIF بقياس التضخم في التباين للمعاملات المقدرة لكل المتغيرات التوضيحية في النموذج ويعتمد هذا المقياس على فحص عناصر القطر الرئيسي للمصفوفة $(X'X)^{-1}$ حيث إن VIF_j ، ($j = 1, 2, \dots, m$) يساوي a_{jj} والتي تمثل $(1 - R_j^2)^{-1}$ حيث أن R_j^2 عبارة عن معامل التحديد لانحدار X_j على بقية المتغيرات التوضيحية، وإن a_{jj} تمثل عناصر القيم القطرية للمصفوفة $(X'X)^{-1}$. واقترح Gunst and Mason في عام 1980 بأنه إذا كان ($a_{jj} \geq 4$) فإن ذلك دليل كافٍ على كون المتغير X_j يعاني من تضخم في تباين معاملاته وإنه مرتبط ببقية المتغيرات التوضيحية [18]، وقد أشار بعض الباحثين أنه إذا كان ($a_{jj} = VIF_j \geq 10$) فإن هذا المقدار يكون سبباً كافياً لإهمال المتغير X_j من التحليل أو استخدام طريقة أخرى كبديل عن المربعات الصغرى الاعتيادية OLS في التقدير، ففي حالة كون المتغيرات التوضيحية متعامدة (مستقلة استقلال خطي تام) تكون قيمة R_j^2 مساوية للصفر وقيمة VIF_j تساوي الواحد الصحيح [19] وهذا يدل على عدم وجود إتفاق بين الباحثين في تحديد قيمة VIF التي تعتبر مؤشراً على وجود مشكلة تعدد العلاقات الخطية وهذا حافز للباحثين في استخدام أكثر من معيار (مؤشر) واحد في الحكم على وجود المشكلة أعلاه من عدمها.

3 - 4 تحليل المكونات الرئيسية Component Analysis : Principal

المعتمد ضد المكونات الرئيسية الناتجة من تحليل PC بينما المكونات الناتجة من تحليل PLS تكون محددة من قيم البيانات لكل من المتغيرات التوضيحية X's والمتغير المعتمد Y ولوجود متغير معتمد واحد يسمى التحليل في هذه الحالة univariate PLS وأيضاً لتحليل PLS إمكانية من تحليل مصفوفة من المتغيرات التنبؤية مع مصفوفة من المتغيرات المعتمدة لإيجاد مكونات متعامدة في الحالات التي يتطلب فيها دراسة أكثر من ظاهرة حياتية أو متغير معتمد في آن واحد بحيث جميع المتغيرات المعتمدة تعتمد على نفس المتغيرات التوضيحية X's ومرتبطة بها و يسمى التحليل في هذه الحالة Multivariate PLS ، إن طريقة PLSR تعمل على مراعاة الارتباط الموجود بين المتغير أو المتغيرات المعتمدة Y's والمتغيرات التنبؤية X's بالإضافة إلى الارتباط الموجود بين المتغيرات X's فيما بينها وإن ذلك يؤثر على البناء للمكونات الناتجة من التحليل وكما في PCR يتم استخدام OLSR للمتغير المعتمد أو لكل متغير معتمد ضد المكونات الناتجة من تحليل PLS [22] .

4-4-1 خوارزمية المربعات الصغرى الجزئية Partial Least Squares Algorithm :

على فرض وجود كلاً من المصفوفة X التي تمثل مصفوفة المتغيرات التوضيحية والمصفوفة Y التي تمثل مصفوفة المتغيرات المعتمدة في حالة وجود أكثر من متغير معتمد واحد أو المتجه Y الذي يمثل متجه المتغير المعتمد في حالة وجود متغير معتمد واحد ، هما بدلالة الصيغة القياسية. يشير [23] إنه لحساب أي مكون t_j ; $j = 1, 2, \dots, h$ بطريقة المربعات الصغرى الجزئية PLS تتبع الخطوات التالية :

1- أخذ قيم ابتدائية عشوائية للمتجه u_{start} وهي عبارة عن بعض القيم التي تؤخذ بطريقة عشوائية من مصفوفة المتغيرات المعتمدة Y القياسية في حالة وجود أكثر من متغير معتمد واحد ويجب أن يكون عدد هذه القيم المختارة يساوي n والذي يمثل عدد المشاهدات لجميع المتغيرات الداخلة في الدراسة قيد البحث وأيضاً يجب أن يكون المجموع لهذه القيم المختارة يساوي صفراً ، وفي حالة وجود متغير معتمد واحد فقط في التحليل المستخدم (حالة Univariate PLS) فإن u_{start} عند كل مكون جديد يتم حسابه هو نفسه المتغير Y القياسي .

2- إيجاد المتجه الأفقي w'_{old} باستخدام الصيغة $1 \times p$ والأبعاد لهذا المتجه هي $w'_{old} = u'_{start} X / u'_{start} u_{start}$ حيث p تمثل عدد المتغيرات التوضيحية أو التنبؤية في التجربة قيد الدراسة ، حيث يلاحظ عند هذه النقطة يتم التداخل بين المتغيرات التنبؤية X's والمتغير المعتمد Y أو المتغيرات المعتمدة Y's بهدف حل مشكلة الارتباط الموجود فيما بينهم .

$$PC_i = a_{1i}X_{1i} + a_{2i}X_{2i} + \dots + a_{mi}X_{mi} \dots (12)$$

الثانية: إذا كانت للمتغيرات التوضيحية X's وحدات قياس مختلفة فيستحسن في هذه الحالة تحويل X's إلى X*'s (متغيرات قياسية لها وسط حسابي يساوي صفر وتباين يساوي الواحد الصحيح) قبل إيجاد القيم a_{ij} وكما يلي :

$$X_{ij}^* = \frac{X_{ij} - \bar{X}_i}{\sqrt{v_{ii}}} \dots (13)$$

وفي هذه الحالة فإن معاملات المكونات الرئيسية a_{ij}^* هي قيم المتجهات المميزة لمصفوفة الارتباط R للمتغيرات X_{ij}^* ، حيث إن عناصر القطر الرئيسي تساوي الواحد الصحيح أي أن $r_{ii} = 1$ أما r_{ij} تساوي :

$$r_{ij} = \frac{\text{cov}(X_{ij}^*, X_{ij}^*)}{\sqrt{\text{var}(X_{ij}^*) \text{var}(X_{ij}^*)}} \dots (14)$$

تحت شرط القياسية فإن $\text{var}(X_{ij}^*) = \text{var}(X_{ij}^*) = 1$ وإن $r_{ij} = \text{cov}(X_{ij}^*, X_{ij}^*)$ ، ثم يتم إيجاد المعادلة المميزة للمصفوفة R باستخدام الصيغة الرياضية التالية :

$$|R - \lambda I| = 0 \dots (15)$$

حيث تم طرح λ^* من القيم القطرية للمصفوفة R والتي تساوي الواحد الصحيح ويجعل قيمة محددها تساوي صفراً يتم الحصول بذلك على m من الجذور المميزة وهذه الجذور أيضاً ترتب بحيث λ_1^* هو أكبر قيمة ويليه λ_2^* وهكذا ، أي إن $\lambda_1^* > \lambda_2^* > \dots > \lambda_m^*$ حيث إن جميع الجذور المميزة للمصفوفة R موجبة ، ثم يتم إيجاد المتجهات المميزة a_{ij}^* حيث لكل جذر مميز λ_j^* متجه مميز a_{ij}^* خاص بقلبه وذلك باستخدام الصيغة التالية :

$$(R - \lambda_j^* I)a_j^* = 0 \dots (16)$$

ثم بعد ذلك يتم اختيار a_{ij}^* بحيث تساوي قيم المتجه المميز القياسي أي بحيث يكون $a_j^* \cdot a_j^* = 1$ حيث يتم استخدام القيد $a_{1j}^2 + a_{2j}^2 + \dots + a_{mj}^2 = 1$ عند إيجاد قيم كل متجه مميز a_{ij}^* عندئذ يتم الحصول على متجهات مميزة قياسية بحيث $a_j^* \cdot a_j^* = 1$ ، وعندئذ تكون الصيغة للمكون الرئيسي PC_j^* كما يلي :

$$PC_j^* = a_{1j}^*X_{1j}^* + a_{2j}^*X_{2j}^* + \dots + a_{mj}^*X_{mj}^* \dots (17)$$

4-4 Partial Least Squares Analysis

إن طريقة انحدار المربعات الصغرى الجزئية PLSR تتشابه مع طريقة انحدار المكونات الرئيسية PCR التي تكون المكونات الرئيسية الخاصة بها تُصاغ من المتغيرات التوضيحية X's فقط ومن ثم استخدام انحدار المربعات الصغرى الاعتيادية OLSR للمتغير

البواقي للمصفوفة Y (حالة Multivariate PLS) أو إيجاد المتجه F_j للمتغير المعتمد الوحيد في التجربة قيد الدراسة (حالة Univariate PLS)، ويستخدم معامل الانحدار b_j أيضاً عند إيجاد القيم التنبؤية \hat{Y} .

10- يتم إيجاد مصفوفتي البواقي E_j و F_j لكل من المصفوفة X والمصفوفة أو المتجه Y على التوالي للمكون t_j كما يلي:

$$E_j = E_{j-1} - t_j p_j' \quad F_j = F_{j-1} - b_j t_j q_j'$$

$$X = E_0 \quad Y = F_0$$

11- عند النقطة أعلاه تنتهي عملية حساب المكون الأول t_1 ولغرض حساب المكون التالي t_{j+1} حيث $j = 1, 2, \dots, h$ يتم الذهاب إلى الخطوة الأولى من الخوارزمية وإجراء الآلية نفسها، حيث أن كلاً من المتجهات w_j و p_j و q_j وكذلك معامل الانحدار b_j يجب أن تخزن لغرض التنبؤ بالقيم المستقبلية للمتغير أو المتغيرات المعتمدة \hat{Y} .

هنالك بعض الملاحظات التي يجب إتباعها عند تطبيق الخوارزمية أعلاه وهي:

- 1) إذا كان القطاع الخاص بالمتغيرات المعتمدة يمتلك فقط متغير معتمد واحد عندئذٍ يمكن حذف الخطوة 7 فحص التقارب ولا يوجد ضرورة إلى تكرار إضافي.
- 2) بعد حساب المكون الأول فإن المصفوفة X في الخطوات 2 و 4 و 8 وكذلك المصفوفة أو المتجه Y في الخطوات 5 و 6 تستبدل بما ينسجم معها من مصفوفة البواقي E_j والمصفوفة أو المتجه F_j على التوالي.

5. الجانب التطبيقي

5-1 عينة البحث

تم التطبيق على بيانات مأخوذة من معمل إسمنت كركوك للمدة (1984-2013) حيث كانت المتغيرات المعتمدة: γ_1 معامل الانبعاث الحراري (مليار كيلو كالوري) حيث كالوري Calorie وحدة سرعة حرارية، γ_2 كمية إنتاج الإسمنت (طن)، γ_3 كمية إنتاج الكنكر (طن)، أما المتغيرات التنبؤية (التوضيحية) هي: x_1 الطاقة الكهربائية (كيلو واط في الساعة)، x_2 النفط الأسود (متر مكعب)، x_3 الأحجار (طن)، x_4 التراب (طن). وإن حجم العينة n لكل متغير من المتغيرات أعلاه يساوي 30 مشاهدة وهي بيانات سنوية للفترة (1984-2013) وقد تم استخدام 26 مشاهدة للفترة (1984-2009) لغرض النمذجة مع إبقاء البيانات للفترة (2010-2013) لغرض مقارنتها مع قيم التنبؤ التي يتم الحصول عليها باستخدام الأنموذج الأمثل المقدر بكل من طريقة انحدار المكونات الرئيسية PCR وطريقة انحدار المربعات الصغرى الجزئية PLSR وبالحالتيها العامة والخاصة. بما أن أحد فروض تحليل الانحدار للمتغير المعتمد

3- إيجاد المتجه الأفقي w_{new}' بصيغته القياسية Standardized من خلال الصيغة $w_{new}' = w_{old}' / \|w_{old}'\|$ ، والمتجهات w_j الناتجة من التحليل عبارة عن متجهات متعامدة بمعنى $cov(w_j, w_i) = 0$ وبذلك فإن عملية معالجة مشكلة الارتباط بين المتغير أو المتغيرات المعتمدة والمتغيرات التنبؤية قد تمت بنجاح، والأبعاد للمصفوفة W التي أعمدها تمثل المتجهات المتعامدة w_j هي $p \times h$ وإن $W'W = I$ حيث I تمثل مصفوفة الوحدة.

4- إيجاد المتجه العمودي t وهو عبارة عن عامل Factor ابتدائي مستخرج من المصفوفة X وذلك باستخدام الصيغة $t = Xw_{new}' / w_{new}' w_{new}'$ ، حيث يكون مجموع القيم لذلك المتجه العمودي t يساوي صفراً.

أما الخطوات التالية تكون بالنسبة إلى قطاع المتغير المعتمد Y أو المتغيرات المعتمدة Y 's:

5- إيجاد المتجه الأفقي q' من خلال الصيغة $q' = t'Y / t't$ وهو عبارة عن أحد تحميلات المصفوفة Y (Y Loadings) وفي حالة وجود متغير معتمد واحد تكون q' عبارة عن قيمة واحدة فقط.

6- إيجاد المتجه العمودي u وذلك باستخدام الصيغة $u = Yq / q'q$ ومجموع القيم لهذا المتجه يساوي صفراً.

7- هذه الخطوة عبارة عن عملية فحص التقارب Check Convergence، حيث يقارن المتجه t في الخطوة 4 مع كل من التكرارات السابقة (w_{old}', u, q') فإذا كان متساوياً إلى واحد منهم أو هنالك فرق ضئيل جداً عندئذٍ $w_{new}' = w_j'$ ، $t = t_j$ ، $q = q_j$ ، $u = u_j$ حيث $j = 1, 2, \dots, h$ ثم يتم الذهاب إلى الخطوة 8، عدا ذلك يتم الذهاب إلى الخطوة 2 ويكون المتجه u_{start} عبارة عن المتجه العمودي u الذي تم حسابه عند الخطوة 6 أعلاه وهكذا عند كل تكرار أثناء حساب المكون الواحد، لقد وضعت هذه المقارنة بهدف الحصول على مكونات t_j متعامدة وبذلك فإن تحليل PLS يعمل على معالجة مشكلة الارتباط الموجود بين المتغيرات التنبؤية والمتغير أو المتغيرات المعتمدة وذلك من خلال حساب العوامل w_j المتعامدة وبنفس الوقت تعالج مشكلة الارتباط الموجود فيما بين المتغيرات التنبؤية وذلك من خلال حساب العوامل t_j المتعامدة.

أما في الخطوة التالية يتم حساب تحميلات المصفوفة X (Loadings) كما يلي:

- 8- حساب المتجه الأفقي p_j' كما في الصيغة $p_j' = t_j' X / t_j' t_j$.
- 9- إيجاد معامل الانحدار b_j الناتج من انحدار u_j الذي يمثل المتغير المعتمد ضد t_j الذي يمثل المتغير التوضيحي حيث b_j عبارة عن قيمة واحدة فقط عند كل مكون t_j وبحسب بالصيغة: $b_j = u_j' t_j / t_j' t_j$ ، ويستخدم لغرض إيجاد المصفوفة F_j مصفوفة

حيث أن n عدد المشاهدات في العينة، σ^2 تباين البواقي Residual Variance، m تمثل عدد المعلمات المقدرة في معادلة الانحدار المتعدد باستثناء المعلمة β_0 ، ويستفاد من هذا المعيار في بيان أي من نماذج الانحدار المقدرة للمتغير المعتمد من خلال استخدام طرائق الانحدار PC، PLS، Univariate، Multivariate PLS هو الأمثل لذلك المتغير المعتمد لغرض الاعتماد عليه في تمثيل البيانات والتنبؤ لذلك المتغير المعتمد علماً أنه كلما كانت قيمة AIC للأتمودج المقدر قليلة كلما دل ذلك على أن هذا الأتمودج المقدر هو الأفضل والطريقة المستخدمة في تقديره هي الأكفأ في التقدير للبيانات قيد البحث .

2-5 نتائج معايير اكتشاف مشكلة التعدد الخطي وتفسيرها

1- Correlations Coefficients .

Y هو أن يكون التوزيع لذلك المتغير طبيعياً ، لذلك تم اختبار التوزيع الطبيعي لكل متغير معتمد باستخدام الاختبار الإحصائي Kolmogorov-Smirnov(KS) ووجد بأن كل منها لا يتبع التوزيع الطبيعي ولهذا السبب تم عمل تحويل Transformation للبيانات الخاصة بهذه المتغيرات وذلك من خلال استخدام تحويل الجذر التربيعي The- Square Root Transformation حيث أصبحت المتغيرات المعتمدة تتبع التوزيع الطبيعي والتي تم الرمز لها بعد إجراء التحويل عليها بالرمز $sq.Y_i$ [3 , 7] .

لقد تم استخدام معيار المقارنة معيار معلومات أكاكي Akaike's (Information Criterion AIC) حيث أُقترح هذا المعيار من قبل الباحث Akaike عام 1974 ويتم إيجاده باستخدام الصيغة : [24]

$$AIC \approx n(1 + \ln(2\pi)) + n \times \ln(\sigma^2) + 2m \dots (18)$$

Correlations: sq.Y1, sq.Y2, sq.Y3, X1, X2, X3, X4

	sq.Y1	sq.Y2	sq.Y3	X1	X2	X3
sq.Y1	0.922					
sq.Y2	0.000					
sq.Y3	0.975	0.968				
X1	-0.334	-0.382	-0.332			
X2	0.975	0.950	0.979	-0.314		
X3	0.949	0.961	0.984	-0.325	0.975	
X4	0.684	0.691	0.689	-0.448	0.671	0.672

Cell Contents: Pearson correlation
P-Value

الشكل رقم (1) : معاملات الارتباط بين المتغيرات التنبؤية والمعتمدة .

الجدول رقم (1): الدليل الشرطي C_{ij} و القيمة المميزة λ_j

Number	الدليل الشرطي C_{ij}	القيمة المميزة λ_j
1	1	1.71308E+13
2	5.174111954	6.39891E+11
3	82.31203807	2.52843E+09
4	7421.061667	3.11061E+05
5	1.100909628E+10	1.41343E-02

2- The Conditional Number $\phi = 1.100909628E+10$

3- The Conditional Index (C_{ij}) and The Eigenvalue

$$\lambda_j, j = 1, 2, \dots, m + 1 .$$

4- The Variance Proportion (p_{ij})

الجدول رقم (2) للمعاملات β_i ، $i = 0, 1, \dots, 4$ قيم نسب التباين p_{ij} 's

ت	Intercept	X ₁	X ₂	X ₃	X ₄
1	8.251041E-33	1.068254E-23	1.138575E-06	0.012903818	5.456931E-03
2	0.000000	2.952366E-24	2.659519E-06	0.052316533	0.968994355
3	1.615597E-26	2.324165E-17	0.91556092	0.915702814	0.017832609
4	1.227287E-17	1.682313E-08	0.084435505	0.01907709	7.717045E-03
5	1.000000	1.000000	0.000000	0.000000	0.000000

الجدول رقم (3)

المتغير التنبؤي	العامل VIF_j
X ₁ الطاقة الكهربائية	1.256
X ₂ النفط الاسود	20.810
X ₃ الاحجار	20.843
X ₄ التراب	2.066

5-The Variance Inflation Factor (VIF_j) .

الجدول التالي يبين قيم VIF_j للمعاملات المقدرة β_j ، $j = 1, 2, \dots, m$ ، الناتجة من تحليل انحدار OLS لكل متغير معتمد $sq.Y_1$ ، $sq.Y_2$ ، $sq.Y_3$ ، ضد المتغيرات التنبؤية الموضحة في الجدول أدناه :

6- The Eigen Vectors (V_j ' S)

V_1	V_2	V_3	V_4	V_5
-0.000001	-0.000000	-0.000017	-0.005197	0.999986
-0.000187	-0.000019	-0.003351	-0.999981	-0.005197
-0.091373	0.026990	-0.995445	0.003353	0.000000
-0.927711	0.361025	0.094944	-0.000152	-0.000000
-0.361945	-0.932166	0.007949	0.000058	-0.000000

الشكل رقم (2): المتجهات المميزة $X'X$ للمصفوفة V_j ' S

أن المتغيرين أعلاه يتسببان في حدوث مشكلة التعدد الخطي لأن لكل منهما نسبة تباين أكبر من النصف ودرجة التعدد الخطي متوسطة لأن قيمة Cl_3 تقع بين ($30 < Cl_3 < 100$) .
ونسف النتيجة يمكن الوصول إليها عند تحليل الارتباط بين المتغيرين التنبؤيين X_1 الطاقة الكهربائية و X_4 التراب وكذلك بين المتغير التنبؤي X_4 التراب مع بقية المتغيرات التنبؤية . من خلال ما سبق يمكن القول أو الحكم بأن هناك وجود لمشكلة التعدد الخطي وبدرجة عالية جداً وإن جميع المتغيرات التنبؤية متسببة في حدوث أو وجود المشكلة أعلاه . وكذلك من خلال ملاحظة مصفوفة الارتباط يتبين بأن هنالك ارتباط معنوي بين كل متغير معتمد من بين المتغيرات المعتمدة ، $sq.Y_1$ و $sq.Y_2$ و $sq.Y_3$ وبين المتغيرات التنبؤية X_2 و X_3 و X_4 حيث قيمة P-Value لكل معامل ارتباط من معاملات الارتباط أعلاه أقل من مستوى المعنوية α الذي يساوي 0.05 .

3-5 نتائج طريقة المكونات الرئيسية PC وتفسيرها

أدناه نتائج تحليل المكونات الرئيسية للمتغيرات التنبؤية الأربعة للفترة (1984-2009) :

من خلال النتائج السابقة لمعايير اكتشاف وجود مشكلة التعدد الخطي يلاحظ :
1- أن قيمة العدد الشرطي ϕ تساوي $1.100909628E+10$ وهذه القيمة كبيرة جداً ويُدل ذلك على إن مشكلة تعدد العلاقات الخطية بين المتغيرات التنبؤية موجودة وبدرجة عالية جداً .
2- هنالك ارتباط بين المتغيرين التنبؤيين X_2 النفط الاسود و X_3 الاحجار من خلال ملاحظة :
a. معامل الارتباط البسيط بين المتغيرين أعلاه من مصفوفة الارتباط يساوي 0.975 وقيمة P-Value لذلك الارتباط تساوي 0.000 دُل ذلك على معنوية علاقة الارتباط بين المتغيرين أعلاه عند $\alpha = 0.05$ والارتباط من النوع الموجب وبدرجة عالية .
b. قيم η^2 لكل من المتغيرين أعلاه تساوي 20.8 وهذه القيمة مرتفعة جداً وأكبر من القيمة 10 ويُدل ذلك على وجود ارتباط بين المتغيرين أعلاه .
c. لكل من المتغيرين أعلاه نسبة تباين p مرتفعة تساوي 0.915 وهذه القيمة أكبر من النصف 0.50 ومقترنة بالقيمة المميزة λ_3 والدليل الشرطي Cl_3 والذي يساوي 82.31203807 يُدل ذلك على

Principal Component Analysis: X1, X2, X3, X4

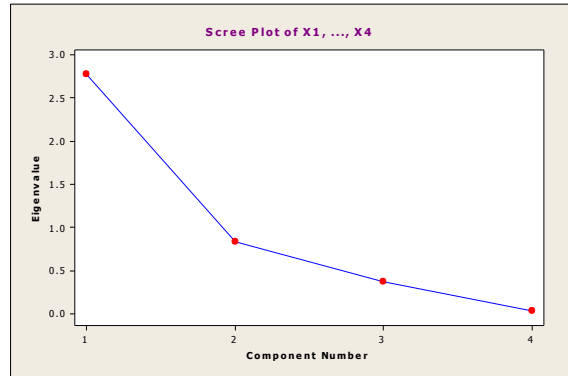
Eigenvalue	2.7721	0.8309	0.3724	0.0246
Proportion	0.693	0.208	0.093	0.006
Cumulative	0.693	0.901	0.994	1.000
Variable	PC1	PC2	PC3	PC4
X1	-0.331	0.892	0.308	-0.010
X2	0.561	0.317	-0.292	0.707
X3	0.562	0.305	-0.301	-0.707
X4	0.509	-0.106	0.854	-0.004

الشكل رقم (3) : نتائج تحليل PC للمتغيرات التنبؤية الأربعة الداخلة في البحث

منحنى حيث يتم إبقاء تلك المكونات الرئيسية التي تساهم في تكوين الشكل المنحني أعلى النقطة الفاصلة ومن ضمنها المكون التابع لتلك القيمة المميزة التي تمثل النقطة الفاصلة ويستبعد من التحليل المكونات الرئيسية أسفل تلك النقطة الفاصلة والتي تكون عبارة عن خط مستقيم [5] ، ويلاحظ من الشكل أدناه أن النقطة الفاصلة هي القيمة المميزة الثالثة وبذلك يتم اختيار المكونات الرئيسية الثلاثة الأولى كمتغيرات تنبؤية في تحليل الانحدار اللاحق :

يلاحظ من النتائج أعلاه إنه تم استخدام مصفوفة الارتباط في التحليل وذلك لاختلاف وحدات القياس لمتغيرات البحث وبذلك تم الحصول على مكونات رئيسية قياسية ومتعامدة . وتم اختيار المكونات الرئيسية الثلاثة الأولى لاستحواذها على أغلب التباين الموجود في المتغيرات التنبؤية اعتماداً على معيار Cattell الذي يعمل على أساس رسم القيم المميزة ضد عدد المكونات الرئيسية ومن ثم تحديد القيمة المميزة التي هي عبارة عن النقطة الفاصلة بين كون الشكل أعلى تلك النقطة

والأشكال التالية تبين نتائج تحليل انحدار OLS لكل متغير معتمد
ضد المكونات الرئيسة الثلاثة المختارة :



الشكل رقم (4) : الرسم البياني للقيم المميزة ضد عدد المكونات الرئيسة المحسوبة

Regression Analysis: sq.Y1 versus PC1, PC2, PC3

The regression equation is

$$\text{sq. Y1} = 22.3 + 4.85 \text{ PC1} + 2.39 \text{ PC2} - 2.09 \text{ PC3}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	22.3178	0.4527	49.30	0.000	
PC1	4.8523	0.2774	17.49	0.000	1.000
PC2	2.3940	0.5064	4.73	0.000	1.000
PC3	-2.0907	0.7565	-2.76	0.011	1.000

S = 2.30840 R-Sq = 93.9% R-Sq(adj) = 93.0%

PRESS = 164.315 R-Sq(pred) = 91.39%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	1790.62	596.87	112.01	0.000
Residual Error	22	117.23	5.33		
Total	25	1907.85			

There are no replicates.

Minitab cannot do the lack of fit test based on pure error.

Durbin-Watson statistic = 1.59437

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	19.415	0.555	(18.264, 20.567)	(14.491, 24.339)
2	13.023	0.728	(11.513, 14.533)	(8.003, 18.043)
3	22.965	0.776	(21.355, 24.575)	(17.915, 28.016)
4	33.867	0.826	(32.154, 35.581)	(28.783, 38.952)

Lack of fit test

Overall lack of fit test is significant at P = 0.021

الشكل رقم (5) : نتائج تحليل انحدار OLS للمتغير sq.Y₁ (معامل الانبعاث الحراري) ضد المكونات الرئيسة المختارة

Regression Analysis: sq.Y2 versus PC1, PC2, PC3

The regression equation is

$$\text{sq.Y2} = 685 + 161 \text{ PC1} + 62.2 \text{ PC2} - 72.4 \text{ PC3}$$

Predictor Coef SE Coef T P VIF

Constant 684.98 15.74 43.53 0.000

PC1 160.671 9.643 16.66 0.000 1.000

PC2 62.22 17.60 3.53 0.002 1.000

PC3 -72.43 26.30 -2.75 0.012 1.000

S = 80.2436 R-Sq = 93.1% R-Sq(adj) = 92.2%

PRESS = 198825 R-Sq(pred) = 90.34%

Analysis of Variance

Source DF SS MS F P

Regression 3 1917343 639114 99.26 0.000

Residual Error 22 141659 6439

Total 25 2059002

There are no replicates.

Minitab cannot do the lack of fit test based on pure error.

Durbin-Watson statistic = 1.31110

Predicted Values for New Observations

New Obs Fit SE Fit 95% CI 95% PI

1 581.8 19.3 (541.8, 621.8) (410.6, 753.0)

2 372.9 25.3 (320.5, 425.4) (198.4, 547.4)

3 726.7 27.0 (670.8, 782.7) (551.2, 902.3)

4 1058.4 28.7 (998.9, 1118.0) (881.7, 1235.2)

Lack of fit test

Overall lack of fit test is significant at P = 0.033

الشكل رقم (6) : نتائج تحليل انحدار OLS للمتغير sq.Y₂ (كمية إنتاج الإسمنت) ضد المكونات الرئيسية المختارة

Regression Analysis: sq.Y3 versus PC1, PC2, PC3

The regression equation is

$$\text{sq.Y3} = 658 + 170 \text{ PC1} + 87.4 \text{ PC2} - 77.3 \text{ PC3}$$

Predictor Coef SE Coef T P VIF

Constant 657.689 9.727 67.62 0.000

PC1 170.259 5.960 28.57 0.000 1.000

PC2 87.43 10.88 8.04 0.000 1.000

PC3 -77.31 16.25 -4.76 0.000 1.000

S = 49.5969 R-Sq = 97.6% R-Sq(adj) = 97.3%

PRESS = 76340.6 R-Sq(pred) = 96.65%

Analysis of Variance

Source DF SS MS F P

Regression 3 2222414 740805 301.16 0.000

Residual Error 22 54117 2460

Total 25 2276530

There are no replicates.

Minitab cannot do the lack of fit test based on pure error.

Durbin-Watson statistic = 1.81217

Predicted Values for New Observations

New Obs Fit SE Fit 95% CI 95% PI

1 557.58 11.93 (532.83, 582.33) (451.79, 663.37)

2 332.07 15.65 (299.62, 364.52) (224.22, 439.92)

3 677.42 16.68 (642.84, 712.01) (568.91, 785.94)

4 1063.68 17.75 (1026.87, 1100.49) (954.43, 1172.93)

Lack of fit test

Overall lack of fit test is significant at P = 0.010

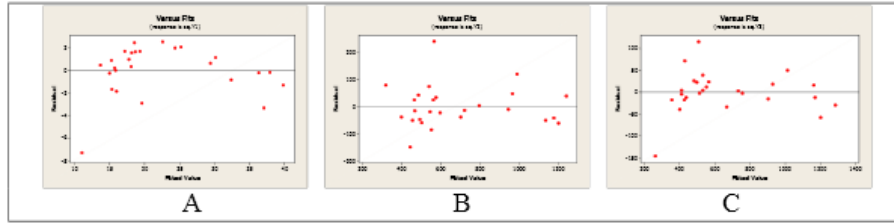
الشكل رقم (7) : نتائج تحليل انحدار OLS للمتغير sq.Y₃ (كمية إنتاج الكلنكر) ضد المكونات الرئيسية المختارة

المختبر الإحصائي T لجميع معاملات الانحدار المقدر ولجميع نماذج الانحدار المقدر وعند مستوى المعنوية α الذي يساوي 0.05، ويلاحظ أيضاً معنوية المختبر الإحصائي Lack of Fit حيث قيمة P-Value للمختبر الإحصائي أعلاه تساوي 0.021 ، 0.033 ، 0.010 للنماذج الثلاثة على التوالي ، وجميع هذه القيم أقل من مستوى المعنوية $\alpha = 0.05$ يدل ذلك على أن الأنموذج الخطي المقدر لكل

يلاحظ من النتائج أعلاه توفر جميع فروض التحليل الإحصائي لتحليل انحدار OLS ولكل أنموذج انحدار مقدر أعلاه من خلال، القدرة التفسيرية بناءً على معامل التحديد المصحح (adj) R - sq لكل أنموذج انحدار مقدر مرتفعة والقدرة التنبؤية اعتماداً على معامل التوقع (pred) R - sq لكل أنموذج انحدار مقدر مرتفعة ، وكذلك معنوية المختبر الإحصائي F لجميع نماذج الانحدار المقدر وأيضاً معنوية

ملاحظة عشوائية الشكل الانتشاري لرسم البواقي ضد القيم التنبؤية ولجميع النماذج المقدرة :

متغير معتمد ملائم ومطابق للبيانات ، كذلك وجد أن هنالك تجانس أو ثبات في تباين البواقي ولجميع نماذج الانحدار المقدرة وذلك من خلال



الشكل رقم (8) : الرسم الانتشاري للبواقي ضد القيم التنبؤية للمتغيرات المعتمدة ، A : $sq.Y_1$ (معامل الانبعاث الحراري) ، B : $sq.Y_2$ (كمية إنتاج الإسمنت) ، C : $sq.Y_3$ (كمية إنتاج الكلنكر) .

توفر لكافة فروض التحليل الإحصائي لتحليل انحدار OLS ولجميع نماذج الانحدار المقدرة .

4-5 نتائج طريقة المربعات الصغرى الجزئية PLS وتفسيرها :
المعتمدة وتفسيرها : تم في هذه الطريقة حساب مكونات من خلال تحليل مصفوفة المتغيرات المعتمدة القياسية مع مصفوفة المتغيرات التوضيحية القياسية في أن واحد باستخدام خوارزمية المربعات الصغرى الجزئية ، والمكونات التي أمكن حسابها هي أربعة مكونات قياسية ومتعامدة وتم اختيار المكون الأول والثاني كمتغيرات تنبؤية في تحليل انحدار OLS لكل متغير معتمد وتم الرمز لهما بالرمزين PCP_1 و PCP_2 على التوالي ، حيث بقية المكونات المحسوبة غير معنوية ووجودها يضعف نتائج التحليل الإحصائي OLS للأنموذج المقدر لكل متغير معتمد، والأشكال التالية تبين نتائج تحليل انحدار OLS لكل متغير معتمد ضد المكونات الرئيسية المختارة :

وأيضاً هنالك توزيع طبيعي للبواقي ولجميع نماذج الانحدار المقدرة كذلك يلاحظ أن هنالك استقلال ذاتي للبواقي الخاصة بالأنموذج المقدر للمتغير المعتمد $sq.Y_3$ حيث قيمة المخبتر الإحصائي DW تساوي 1.81 وهذه القيمة أكبر من القيمة $d_U = 1.66$ وأقل من القيمة 2 ، أما بخصوص بقية النماذج المقدرة لكل من المتغير المعتمد $sq.Y_2$ و $sq.Y_1$ فإن القيمة الإحصائية DW تقع في منطقة حرجة وهي $d_U < DW < d_L$ والقيمة d_L تساوي 1.05 والقيمة d_U قيمتها موضحة في أعلاه ، وهذا يعني أنه لا يمكن القول بوجود ارتباط ذاتي بين البواقي من عدمه ، وكذلك يلاحظ وجود استقلال خطي تام بين المتغيرات التنبؤية المتمثلة بالمكونات الرئيسية المتعامدة PC_1, PC_2, PC_3 حيث قيمة عامل تضخم التباين VIF لكل مكون في كل أنموذج انحدار مقدر تساوي الواحد الصحيح لا غير وبذلك تم التخلص من مشكلة تعدد العلاقات الخطية بالإضافة إلى ذلك هنالك

Regression Analysis: $sq.Y_1$ versus PCP_1, PCP_2

The regression equation is

$$sq.Y_1 = 22.3 + 4.99 PCP_1 + 2.13 PCP_2$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	22.3178	0.4549	49.07	0.000	
PCP1	4.9933	0.2809	17.78	0.000	1.000
PCP2	2.1288	0.5373	3.96	0.001	1.000

S = 2.31931 R-Sq = 93.5% R-Sq(adj) = 93.0%

PRESS = 161.255 R-Sq(pred) = 91.55%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1784.13	892.06	165.84	0.000
Residual Error	23	123.72	5.38		
Total	25	1907.85			

There are no replicates.

Minitab cannot do the lack of fit test based on pure error.

Durbin-Watson statistic = 1.59791

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	19.9207	0.657308	(18.5610, 21.2805)	(14.9339, 24.9076)
2	18.1575	0.761490	(16.5822, 19.7327)	(13.1076, 23.2073)
3	20.1061	0.791391	(18.4690, 21.7433)	(15.0367, 25.1756)
4	23.9375	0.465798	(22.9739, 24.9011)	(19.0439, 28.8312)

Lack of fit test

Overall lack of fit test is significant at P = 0.024

الشكل رقم (9) : نتائج تحليل انحدار OLS للمتغير $sq.Y_1$ (معامل الانبعاث الحراري) ضد مكونات MPLS المختارة

Regression Analysis: sq.Y2 versus PCP1, PCP2

The regression equation is

$$\text{sq.Y2} = 685 + 165 \text{ PCP1} + 54.1 \text{ PCP2}$$

Predictor Coef SE Coef T P VIF

Constant 684.98 16.18 42.34 0.000

PCP1 164.643 9.992 16.48 0.000 1.000

PCP2 54.09 19.11 2.83 0.009 1.000

S = 82.5024 R-Sq = 92.4% R-Sq(adj) = 91.7%

PRESS = 206724 R-Sq(pred) = 89.96%

Analysis of Variance

Source DF SS MS F P

Regression 2 1902449 951224 139.75 0.000

Residual Error 23 156553 6807

Total 25 2059002

There are no replicates.

Minitab cannot do the lack of fit test based on pure error.

Durbin-Watson statistic = 1.34370

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	593.469	23.3818	(545.100, 641.838)	(416.078, 770.860)
2	532.727	27.0878	(476.691, 588.762)	(353.094, 712.360)
3	631.462	28.1514	(573.227, 689.698)	(451.131, 811.794)
4	739.175	16.5694	(704.898, 773.451)	(565.098, 913.252)

Lack of fit test

Overall lack of fit test is significant at P = 0.016

الشكل رقم (10): نتائج تحليل انحدار OLS للمتغير sq.Y₂ (كمية إنتاج الإسمنت) ضد مكونات MPLS المختارة

Regression Analysis: sq.Y3 versus PCP1, PCP2

The regression equation is

$$\text{sq.Y3} = 658 + 175 \text{ PCP1} + 78.9 \text{ PCP2}$$

Predictor Coef SE Coef T P VIF

Constant 657.69 10.33 63.65 0.000

PCP1 175.380 6.381 27.48 0.000 1.000

PCP2 78.86 12.21 6.46 0.000 1.000

S = 52.6882 R-Sq = 97.2% R-Sq(adj) = 97.0%

PRESS = 85054.0 R-Sq(pred) = 96.26%

Analysis of Variance

Source DF SS MS F P

Regression 2 2212681 1106341 398.53 0.000

Residual Error 23 63849 2776

Total 25 2276530

There are no replicates.

Minitab cannot do the lack of fit test based on pure error.

Durbin-Watson statistic = 1.58754

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	576.664	14.9322	(545.775, 607.554)	(463.378, 689.951)
2	515.392	17.2989	(479.606, 551.177)	(400.674, 630.110)
3	575.086	17.9782	(537.895, 612.277)	(459.921, 690.250)
4	714.382	10.5816	(692.492, 736.272)	(603.212, 825.552)

Lack of fit test

Overall lack of fit test is significant at P = 0.007

الشكل رقم (11): نتائج تحليل انحدار OLS للمتغير sq.Y₃ (كمية إنتاج الكلنكر) ضد مكونات MPLS المختارة

بمعنى تم تطبيق التحليل أعلاه ثلاثة مرات أي بعدد المتغيرات المعتمدة الداخلة في البحث وفي كل مرة تم الحصول على مكونات قياسية خاصة بالمتغير المعتمد الذي استخدم في التحليل الإحصائي Univariate PLS معلماً أن المكونات الخاصة بكل متغير معتمد تختلف عن المكونات الخاصة بأي متغير معتمد آخر، والمكونات التي أمكن حسابها عند كل مرة تم فيها إجراء التحليل الإحصائي أعلاه هي أربعة مكونات قياسية ومتعامدة وتم اختيار فقط المكون الأول

يلاحظ من النتائج أعلاه توفر جميع فروض التحليل الإحصائي لتحليل انحدار OLS ولكل أنموذج انحدار مقدر أعلاه من خلال ، جميع المعايير الإحصائية السابقة الذكر في الصفحات (17-18) .
2-4-5 نتائج طريقة المربعات الصغرى الجزئية الأحادية المتغير المعتمد وتفسيرها

تم في هذه الطريقة حساب مكونات قياسية من خلال تحليل مصفوفة المتغيرات التنبؤية القياسية مع كل متغير معتمد قياسي على حدى،

التالية تبين نتائج تحليل انحدار OLS لكل متغير معتمد ضد المكونات الخاصة بذلك المتغير المعتمد :

Comp1 والمكون الثاني Comp2 الخاصة بكل متغير معتمد حيث بقية المكونات التي أمكن حسابها غير معنوية ووجودها يضعف نتائج التحليل الإحصائي OLS ولجميع المتغيرات المعتمدة ، والأشكال

Regression Analysis: sq.Y1 versus Comp1, Comp2						
The regression equation is						
sq.Y1 = 22.3 + 5.00 Comp1 + 2.09 Comp2						
Predictor	Coef	SE Coef	T	P	VIF	
Constant	22.3178	0.4545	49.10	0.000		
Comp1	5.0017	0.2809	17.81	0.000	1.000	
Comp2	2.0867	0.5353	3.90	0.001	1.000	
S = 2.31748 R-Sq = 93.5% R-Sq(adj) = 93.0%						
PRESS = 161.383 R-Sq(pred) = 91.54%						
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	2	1784.32	892.16	166.12	0.000	
Residual Error	23	123.53	5.37			
Total	25	1907.85				
There are no replicates.						
Minitab cannot do the lack of fit test based on pure error.						
Durbin-Watson statistic = 1.58383						
Predicted Values for New Observations						
New Obs	Fit	SE Fit	95% CI	95% PI		
1	19.9783	0.662285	(18.6083, 21.3483)	(14.9923, 24.9643)		
2	18.2246	0.767350	(16.6372, 19.8120)	(13.1746, 23.2747)		
3	20.0435	0.799081	(18.3905, 21.6965)	(14.9725, 25.1146)		
4	23.9781	0.465206	(23.0157, 24.9404)	(19.0883, 28.8678)		
Lack of fit test						
Overall lack of fit test is significant at P = 0.025						

الشكل رقم (12): نتائج تحليل انحدار OLS للمتغير sq.Y₁ (معامل الانبعاث الحراري) ضد مكونات UPLS المختارة

Regression Analysis: sq.Y2 versus Comp1, Comp2						
The regression equation is						
sq.Y2 = 685 + 164 Comp1 + 61.6 Comp2						
Predictor	Coef	SE Coef	T	P	VIF	
Constant	684.98	15.87	43.15	0.000		
Comp1	164.037	9.784	16.77	0.000	1.000	
Comp2	61.57	19.30	3.19	0.004	1.000	
S = 80.9453 R-Sq = 92.7% R-Sq(adj) = 92.0%						
PRESS = 195678 R-Sq(pred) = 90.50%						
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	2	1908302	954151	145.62	0.000	
Residual Error	23	150699	6552			
Total	25	2059002				
There are no replicates.						
Minitab cannot do the lack of fit test based on pure error.						
Durbin-Watson statistic = 1.33704						
Predicted Values for New Observations						
New Obs	Fit	SE Fit	95% CI	95% PI		
1	589.249	22.1415	(543.445, 635.052)	(415.649, 762.848)		
2	527.532	25.5700	(474.637, 580.428)	(351.928, 703.136)		
3	634.188	26.6275	(579.105, 689.271)	(457.913, 810.463)		
4	736.424	16.2958	(702.713, 770.134)	(565.616, 907.231)		
Lack of fit test						
Overall lack of fit test is significant at P = 0.013						

الشكل رقم (13): نتائج تحليل انحدار OLS للمتغير sq.Y₂ (كمية إنتاج الإسمنت) ضد مكونات UPLS المختارة

Regression Analysis: sq.Y3 versus Comp1, Comp2

The regression equation is
sq.Y3 = 658 + 176 Comp1 + 75.4 Comp2
Predictor Coef SE Coef T P VIF
Constant 657.69 10.47 62.84 0.000
Comp1 175.929 6.474 27.18 0.000 1.000
Comp2 75.37 12.27 6.14 0.000 1.000
S = 53.3686 R-Sq = 97.1% R-Sq(adj) = 96.9%
PRESS = 87777.4 R-Sq(pred) = 96.14%
Analysis of Variance
Source DF SS MS F P
Regression 2 2211022 1105511 388.14 0.000
Residual Error 23 65509 2848
Total 25 2276530
There are no replicates.

Minitab cannot do the lack of fit test based on pure error.

Durbin-Watson statistic = 1.57498

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	578.617	15.2661	(547.037, 610.198)	(463.788, 693.447)
2	517.888	17.7068	(481.259, 554.517)	(401.569, 634.208)
3	573.910	18.3464	(535.958, 611.863)	(457.167, 690.653)
4	715.030	10.7170	(692.861, 737.200)	(602.425, 827.636)

Lack of fit test

Overall lack of fit test is significant at P = 0.009

الشكل رقم (14) : نتائج تحليل انحدار OLS للمتغير sq.Y₃ (كمية إنتاج الكلنكر) ضد مكونات UPLS المختارة

(2013) وتم استخدام البقية من مشاهدات العينة لغرض النمذجة ، لقد تم حساب قيم المكونات للفترة (2010-2013) في كل طريقة من طرائق الانحدار أعلاه من خلال حساب قيم المكونات للفترة (1984-2013) ومن ثم حجز القيم الناتجة من المكونات المحسوبة أعلاه للفترة (2010-2013) كمجموعة حجز خارجاً Holdout Set ومن ثم إيجاد الأنموذج المقدر الناتج من تحليل البيانات للفترة (1984-2009) واستخدامه لغرض التنبؤ بالقيم المستقبلية لكل متغير معتمد للفترة (2010-2013) ومقارنتها بالقيم الأصلية المتوفرة للفترة الأخيرة وإيجاد متوسط مربعات الخطأ MSe لذلك التنبؤ لكل متغير معتمد وكل طريقة من الطرائق أعلاه ، والجدول التالي يبين القيم الأصلية والتنبؤية لكل متغير معتمد ولجميع الطرائق المستخدمة وللفترة أعلاه وكما يلي :

يلاحظ من النتائج أعلاه توفر جميع فروض التحليل الإحصائي لتحليل انحدار OLS ولكل أنموذج انحدار مقدر أعلاه من خلال ، جميع المعايير الإحصائية السابقة الذكر في الصفحات (17-18) .

5-5 التنبؤ

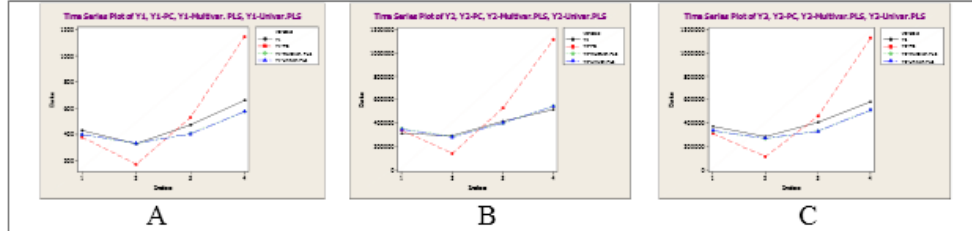
تم إيجاد القيم التنبؤية لقيم المتغيرات المعتمدة الداخلة في البحث للفترة (2010-2013) وذلك من خلال استخدام جميع نماذج الانحدار المقدر ولجميع المتغيرات المعتمدة الناتجة من استخدام كل من طريقة المكونات الرئيسية PC وطريقة المربعات الصغرى الجزئية متعددة المتغيرات MPLS وطريقة المربعات الصغرى الجزئية الأحادية المتغير UPLS ، لقد تم استخدام أسلوب تنبؤ العينة البعدية-Out of-Sample- Forecasting في عملية التنبؤ بالقيم المستقبلية [25] حيث تم حجز المجموعة الأخيرة من مشاهدات العينة (2010-

الجدول رقم (4) القيم الأصلية والتنبؤية لجميع المتغيرات المعتمدة ولجميع الطرائق المستخدمة للفترة (2010-2013)

كمية إنتاج الإسمنت γ_2				معامل الاتبعات الحراري γ_1				السنة
(UPLS)	(MPLS)	(PC)	γ_2	(UPLS)	(MPLS)	(PC)	γ_1	
$\hat{\gamma}_2$	$\hat{\gamma}_2$	$\hat{\gamma}_2$		$\hat{\gamma}_1$	$\hat{\gamma}_1$	$\hat{\gamma}_1$		
347214	352205	338491	309707	399	397	377	427	2010
278290	283798	139054	292572	332	330	170	328	2011
402194	398744	528093	415384	402	404	527	471	2012
542320	546380	1120211	516890	575	573	1147	658	2013
607852893	757403195	1.002739607E+11		3112.5	3154.5	67430.25		MSe التنبؤ
كمية إنتاج الكلنكر γ_3				السنة				
(UPLS)	(MPLS)	(PC)	γ_3					
$\hat{\gamma}_3$	$\hat{\gamma}_3$	$\hat{\gamma}_3$						
334798	332541	310895	371785	2010				
268208	265629	110270	282158	2011				
329373	330724	458898	407923	2012				
511268	510342	1131415	581290	2013				
3158955913	3201650921	8.462226072E+10		MSe التنبؤ				

طريقة المكونات الرئيسية PC أعطت قيم MSE جداً مرتفعة ولجميع المتغيرات المعتمدة ويدل ذلك على الكفاءة القليلة جداً لهذه الطريقة في القدرة على التنبؤ بالقيم المستقبلية لمتغير الاستجابة ، والشكل التالي يبين رسم القيم الأصلية والقيم التنبؤية ولجميع الطرائق أعلاه ولكل متغير معتمد من بين المتغيرات المعتمدة وللفترة (2010-2013) :

يلاحظ من النتائج أعلاه تفوق طريقة UPLS في القدرة على التنبؤ بالقيم المستقبلية لكل متغير معتمد من بين المتغيرات المعتمدة أعلاه حيث قيمة متوسط مربعات الخطأ للبواقي لكل متغير معتمد أقل قيمة عند الطريقة أعلاه مقارنة بقيم MSE لبقية الطرائق الأخرى ويدل ذلك على أن الطريقة أعلاه لها قدرة تنبؤية عالية ، أما طريقة MPLS أيضاً لها قدرة تنبؤية عالية ولكن تأتي بعد الطريقة UPLS ، أما



الشكل رقم (15): الرسم البياني للقيم الأصلية والتنبؤية باستخدام الأنموذج المقدر باستخدام جميع الطرائق أعلاه وللمتغيرات المعتمدة ، A: γ_1 (معامل الانبعاث الحراري) ، B: γ_2 (كمية إنتاج الإسمنت) ، C: γ_3 (كمية إنتاج الكلنكر).

بكونها الطريقة الأكفأ والأكثر قدرة على التنبؤ بالقيم المستقبلية للمتغير المعتمد ، ويمكن إجمال أهم معايير المقارنة بين الثلاثة أعلاه بهدف توضيح وتحديد الطريقة الأكفأ في معالجة مشاكل التطبيق وتوفيق أنموذج انحدار أمثل تتوفر فيه كافة فروض التحليل بالإضافة إلى امتلاكه قدرة تنبؤية عالية في التنبؤ بالقيم المستقبلية لكل متغير معتمد من بين المتغيرات المعتمدة الداخلة في البحث كما في الجدول التالي :

يلاحظ من الشكل أعلاه اقتراب القيم التنبؤية لكل من طريقة المربعات الصغرى الجزئية متعددة المتغيرات MPLS وطريقة المربعات الصغرى الجزئية الأحادية المتغير UPLS من القيم الأصلية ولكل متغير معتمد من بين المتغيرات المعتمدة ولكن قيمة MSE بطريقة UPLS هي الأقل وبذلك هي التي تملك القدرة التنبؤية الأعلى أما طريقة المكونات الرئيسية PC يلاحظ أن القيم التنبؤية لهذه الطريقة بعيدة جداً عن القيم الأصلية ولكل متغير معتمد ، وبذلك فإن القدرة التنبؤية لهذه الطريقة ضعيفة جداً مقارنة بالطريقة UPLS التي تميزت

الجدول رقم (5) معايير المقارنة بين PCR و MPLSR و UPLSR وللمتغيرات المعتمدة γ_1 معامل الانبعاث الحراري و γ_2 كمية إنتاج الإسمنت و γ_3 كمية إنتاج الكلنكر.

ت	معايير المقارنة	معامل الانبعاث الحراري Y_1			كمية إنتاج الإسمنت Y_2			كمية إنتاج الكلنكر Y_3		
		طريقة PC (3 مكونات)	طريقة PLS Multivar. (2 مكونات)	طريقة PLS Univar. (2 مكونات)	طريقة PC (3 مكونات)	طريقة PLS Multivar. (2 مكونات)	طريقة PLS Univar. (2 مكونات)	طريقة PC (3 مكونات)	طريقة PLS Multivar. (2 مكونات)	طريقة PLS Univar. (2 مكونات)
1	VIF لكل مكون	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
2	R-sq(adj)	93.0%	93.0%	93.0%	92.2%	91.7%	92.0%	97.3%	97.0%	
3	R-sq(pred)	91.39%	91.55%	91.54%	90.34%	89.96%	90.50%	96.65%	96.26%	
4	MSE النموذج	5.33	5.38	5.37	6439	6807	6552	2460	2776	
5	AIC (Ln)	123.292	121.534	121.486	307.808	307.253	306.260	282.790	283.932	
6	MSE التنبؤ	67430.25	3154.5	3112.5	1.002739607E+11	757403195	607852893	8.462226072E+10	3201650921	
7	P-Value قيمة المختبر KS للبوافي	P-Value= 0.033	P-Value= 0.026	P-Value= 0.013	P-Value= >0.150	P-Value= >0.150	P-Value= >0.150	P-Value= 0.122	P-Value= 0.094	
8	تجانس التباين للبوافي	متجانس	متجانس	متجانس	متجانس	متجانس	متجانس	متجانس	متجانس	
9	الاستقلال الذاتي للبوافي	لا يمكن التحديد بفشل الاختبار	مستقلة	مستقلة	لا يمكن التحديد بفشل الاختبار	لا يمكن التحديد بفشل الاختبار	لا يمكن التحديد بفشل الاختبار	مستقلة	مستقلة	
10	نتيجة اختبار حسن المطابقة للنموذج الخطي المقتر	مطابق للبيانات	مطابق للبيانات	مطابق للبيانات	مطابق للبيانات	مطابق للبيانات	مطابق للبيانات	مطابق للبيانات	مطابق للبيانات	

6. الاستنتاجات والتوصيات

إن أهم الاستنتاجات والتوصيات التي تم التوصل إليها من خلال الجانب التطبيقي هي :

1-6 الاستنتاجات :

1- من خلال المعايير الإحصائية المستخدمة في الكشف عن وجود ودرجة التعدد الخطي وما هي المتغيرات التنبؤية المسببة للمشكلة أعلاه وجد بأن هنالك وجود لهذه المشكلة في متغيرات معمل إسمنت كركوك وبدرجة مرتفعة وإن جميع المتغيرات التنبؤية تتسبب في حدوث المشكلة أعلاه وهي متغيرات من الواجب توفرها جميعها ولا يمكن الاستغناء عن أي منها في تقدير أنموذج انحدار يقيس التغيرات الحاصلة في كل متغير من المتغيرات المعتمدة الداخلة في البحث بالإضافة إلى ذلك وجود ارتباط لكل متغير معتمد مع المتغيرات التنبؤية عدا المتغير التنبؤي الطاقة الكهربائية وبسبب وجود مشكلة التعدد الخطي بين المتغيرات التوضيحية فيما بينها وكذلك الارتباط بين المتغيرات التوضيحية والمتغيرات المعتمدة بالإضافة إلى وجود مشكلة صغر حجم العينة الداخلة في البحث ، أدى ذلك إلى فشل طريقة OLS في توفيق أنموذج انحدار أمثل تتوفر فيه كافة فروض التحليل الإحصائي لهذه الطريقة ولجميع المتغيرات المعتمدة الداخلة في البحث.

2- النماذج المستخلصة بطريقة المربعات الصغرى الجزئية PLS وبحالتيها العامة وبالخاصة MPLS و UPLS على التوالي أكثر كفاءة من النماذج المستخلصة بطريقة المكونات الرئيسية PC من حيث :

(a) تخفيض البعد للبيانات ، حيث تمكنت طريقة PLS وبحالتيها العامة والخاصة من توفيق أنموذج انحدار تتوفر فيه كافة فروض التحليل الإحصائي ولجميع المتغيرات المعتمدة الداخلة في البحث بالاعتماد على إثبات فقط من المكونات القياسية المتعامدة بينما طريقة PC فإنها تمكنت من توفيق أنموذج انحدار تتوفر فيه كافة فروض التحليل الإحصائي ولجميع المتغيرات المعتمدة بالاعتماد على ثلاثة مكونات رئيسية .

(b) قيمة المعيار AIC باستخدام طريقة PLS وبحالتيها العامة والخاصة وللمتغيرين المعتمدين γ_1 (معامل الانبعاث الحراري) و γ_2 (كمية إنتاج الإسمنت) هي الأقل مقارنة بالقيمة لهذا المعيار وللمتغيرين أعلاه عند الطريقة PC وعلى الخصوص الحالة الخاصة UPLS ، أما بالنسبة إلى المتغير المعتمد γ_3 (كمية إنتاج الكلنكر) فإن أقل قيمة للمعيار AIC هي عند الطريقة PC .

3- تفوق طريقة المربعات الصغرى الجزئية PLS وبحالتيها العامة والخاصة على طريقة المكونات الرئيسية PC من حيث القدرة التنبؤية للنماذج المستخلصة، حيث كانت قيمة MSE التنبؤ أقل ما يمكن مع طريقة UPLS ثم طريقة MPLS وأخيراً طريقة PC .

4- بناءً على ما ورد من استنتاجات أعلاه تعد طريقة UPLS هي الأفضل في تمثيل البيانات وفي التنبؤ بالقيم المستقبلية للمتغير المعتمد.

5- عند وجود حالة يتطلب فيها دراسة أكثر من متغير معتمد واحد بحيث تعتمد على نفس المتغيرات التنبؤية بالإضافة إلى مشاكل التطبيق المذكورة في النقطة الأولى أعلاه فإن تطبيق طريقة UPLS بعدد من المرات مساوي لعدد المتغيرات المعتمدة وخاصة إن كان عددها كبير سيكلف وقتاً وجهداً ، ولغرض تسهيل العملية وتطبيقها لمرة واحدة فقط عند حساب المكونات المتعامدة بحيث تكون هذه المكونات نفسها عند كل متغير معتمد كمتغيرات تنبؤية في تحليل انحدار OLS عندها يفضل استخدام طريقة MPLS ولكن بقدرة تنبؤية أقل نسبياً من القدرة التنبؤية لطريقة UPLS .

2-6 التوصيات :

1- توصي الباحثة بتطبيق طريقة المربعات الصغرى الجزئية الأحادية المتغير UPLS في الدراسات التي تعاني متغيراتها التنبؤية والتي لا يمكن الاستغناء عن أي منها من خلال حذفها من التحليل الإحصائي من مشكلة تعدد العلاقات الخطية وخصوصاً عندما يكون الارتباط قوي جداً فيما بينها بالإضافة إلى الارتباط الموجود بين المتغير المعتمد وبين المتغيرات التنبؤية وكذلك مع وجود لمشكلة قلة عدد المشاهدات المتوفرة في التحليل الإحصائي ، وذلك لكون طريقة UPLS تقدم نماذج انحدار لها مطابقة جيدة للبيانات المدروسة على الرغم من وجود المشاكل أعلاه فضلاً عن ذلك قدرتها العالية في عملية تخفيض عدد المتغيرات التنبؤية في أنموذج الانحدار المقدر الأمثل الذي تتوفر فيه كافة فروض التحليل الإحصائي لانحدار OLS للمتغير المعتمد ضد المكونات المتعامدة الناتجة من تحليل UPLS وكذلك قدرة الأنموذج المقدر الناتج من هذه الطريقة على التنبؤ بالقيم المستقبلية وبكفاءة عالية جداً .

2- إمكانية استخدام طريقة المربعات الصغرى الجزئية الحصينة Robust PLS ومقارنة نتائجها مع الطرائق الإحصائية الأخرى .

3- توظيف الاساليب الذكائية كالشبكات العصبية في انحدار المربعات الصغرى الجزئية.

المصادر

- [1] Bluman , Allan G., "Elementary Statistics (A step by step Approach) ", Seventh Edition , Published by MC Graw – Hill , Obusiness Unit of The MC Graw – Hill Companies , New York , (2009) .
- [2] Maitra, Saikat and Yan, Jun, "Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression" , Casualty Actuarial Society , (2008), pp (79 – 90) .
- [3] الراوي ، خاشع محمود ، " المدخل إلى تحليل الانحدار" ، مديرية دار الكتب للطباعة والنشر، جامعة الموصل، العراق، (1987) .
- [4] Poole, Michael A., Farrell, Patrick N.O., "The Assumptions of the Linear Regression Model " , Received , 10 July (1970), pp (145-157) .
- [5] الحسنوي، أموري هادي كاظم والقيسي، باسم شلبية مسلم، "القياس الاقتصادي المتقدم- النظرية والتطبيق"، مطبعة الطيف، بغداد، العراق، (2002) .
- [6] Henry , Gary T., "Practical Sampling", Chapter Title , Sample Size, Publishing Company, SAGE Publications, Inc., City, Thousand oaks , 14 October (2013), [http://dx.doi.org/0.4135/9781412985451.n7.pp\(1-15\)](http://dx.doi.org/0.4135/9781412985451.n7.pp(1-15))
- [7] سليمان، أسامة ربيع أمين، " التحليل الإحصائي للبيانات باستخدام برنامج Minitab "، قسم الإحصاء والرياضة والتأمين، كلية التجارة ، السادات ، جامعة المنوفية ، جمهورية مصر العربية، (2007) .
- [8] البدراني، ظافر رمضان مطر، الحياي، عمر سالم وجاسم، عمر قصي، " التحليل الإحصائي للتلوث البيئي الناجم عن المولدات الأهلية في مدينة الموصل"، المجلة العراقية للعلوم الإحصائية، عدد خاص بوقائع المؤتمر العلمي السادس لكلية علوم الحاسوب والرياضيات ، العدد- 25، (2013)، ص. ص (62 – 80) .
- [9] Disatnik, David and Sivan, Liron, "The Multicollinearity illusion in Moderated Regression Analysis", Tel Aviv University, Carnegie Mellon University, USA., (2014), PP(1-12).
- [10] Jeeshim and Kucc, "Multicollinearity in Regression Models", No. 625, (2002), pp (1-8), <http://php.Indiana.edu/~kucc.625> .
- [11] وارطان، هاسميك أنترانيك، " تعدد العلاقات الخطية في نموذج الانحدار المتعدد - دراسة نظرية وتطبيقية "، رسالة ماجستير غير منشورة ، كلية الإدارة والاقتصاد، جامعة الموصل، العراق، (1989) .
- [12] سعيد، هيفاء عبدالجواد، " التقدير المتحيز في الانحدار - دراسة مقارنة " ، رسالة ماجستير غير منشورة ، كلية الإدارة والاقتصاد، جامعة الموصل ، العراق، (1996) .
- [13] Rawlings, John O., Pantula, Sastry G. and Dickey, David A., "Applied Regression Analysis", A research Tool, Second Edition, Berlin, New york, (1998) .
- [14] الصفاوي، صفاء يونس ويحيى، مزاحم محمد ،"استخدام انحدار الحرف والأنتروبي العظمى العامة في تحليل التلوث لمعمل إسمنت كركوك" ، المجلة العراقية للعلوم الإحصائية ، عدد خاص بوقائع المؤتمر العلمي الثاني للرياضيات- قسم الإحصاء والمعلوماتية، العدد-17، (2010)، ص. ص (239 – 260) .
- [15] Akdeniz, Fikri, "III-Conditioning and Multicollinearity", Department of Mathematics, Cukurova University, Adana, Turkey, Linear Algebra and its Applications, 321, (2000), pp(295-305), E-mail: akdeniz@mail.cu.edu.tr (F.akdeniz) .
- [16] يحيى، مزاحم محمد،" استخدام المكونات الرئيسية وانحدار الحرف في تقدير معادلة السعر العالمي للقمح للفترة من 1961-2002" ، مجلة تكريت للعلوم الإدارية والاقتصادية ، المجلد-1، العدد-1، (2005)، ص . ص (146 – 156) .
- [17] ويسلوسكي، جورج ، " الانحدار المتعدد وتحليل التباين "، ترجمة د. شلال حبيب الجبوري 1990، مطابع التعليم العالي ، موصل، العراق، (1976) .
- [18] الطالب ، محاسن صالح ، " استخدام أسلوب انحدار الحرف في التنبؤ لتقدير نماذج الصندوق الأسود مع التطبيق " ، المجلة العراقية للعلوم الإحصائية ، العدد - 22 ، (2012)، ص . ص (121 – 135) .
- [19] Chatterjee, Samprit and Hadi, Ali S., " Regression Analysis by Example", Fourth Edition, Copyright © John wiley & Sons, Inc., Printed in the United States of America, (2006) .
- [20] Travaglini, Guido, " Principal Components and Factor Analysis-Acomparative Study" , University of La Sapienza , Roma, Italy , Second Draft , MPRA Paper, No. 35486, Posted. 20, 13 October (2011), <http://Mpra.Ub.Uni-Muenchen.de/35486/> .
- [21] Nguyen, Danh V. and Rocke, David M., "on Partial Least Squares Dimension Reduction for Microarray-Based Classification: A simulation Study", University of California, Davis. USA, Computational Statistics and Data Analysis , (2004), pp (407 – 425), E-mail: Ucdnguyen@ucdavis.edu .
- [22] Garthwaite, Paul H., "An Interpretation of Partial Least Squares", Journal of The American Statistical Association, Theory and Method, Vol. 89, No. 425, (1994), pp (122-127) .
- [23] Geladi, Paul and Kowalski, Bruce, "Partial Least Squares Regression: A tutorial", University of Washington, Seattle, WA 98195 (U. S. A.), Printed in The Netherlands , (1986), pp (1-17) .
- [24] Cavanaugh, Joseph E., "The Akaike Information Criterion", Department of Statistics and Actuarial Science , The University of Iowa, (2012).
- [25] Makridakis, Spyros, Wheelwright, Steven C. and Hyndman, Rob J., "Forecasting " , Methods and Applications, 3rd Edition, New york, (1998)

A comparison Between Principal Component Regression and Partial Least Squares Regression Methods with application in The Kirkuk Cement Factory

Thafer Ramathan Muttar AL-Badrany, Takwa Abdulsalam Taha AL-Mola

Dep. of Statistics and Informatics, College of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq

thafer63@yahoo.com

Staralmula@yahoo.com

Abstract

Appear in Many Application Areas for Regression Analysis and Presence the case of More Than One Variable Dependent Affected by A variety of Explanatory Variable and at The Same Time The Number of Observation is Relatively Small Compared to The Number of Variables, and Show Here The Problem of not Provide Offers Several Hypotheses Multiple Regression Analysis, More Over Prominence Problem of Multicollinearity between the Explanatory Variables Beside The Correlation between The Explanatory Variables and The Dependent Variables and Reflexive that on the Regression Estimates. and in This Research was Dealing with Problems from this Type Related to Variables Kirkuk Cement Factory, used the Methods, Principal Component PC and Partial Least Squares PLS to Solve the Problems Above, The First Method is Considered as one of the Commonest Methods used in Solving the Problem of Multicollinearity between the Explanatory Variables and the Second Method is Considered as one of the Methods Which Dally Methodically Different in Deduction the Components Dependent on Curing The Correlation the Presence between The Explanatory Variables and The Dependent Variables. Through the Statistical Analysis, Orphan Conduction to The PLS Method it has Succeeded in Establishing the Optimal Regression Model for all Depended Variables, Besides Superiority This Method Whence Ability on the Prediction for Futuristic Values Apiece Dependent Variables and Also Whence Dimension Reduction. The (Minitab, Version, 16.1) is used in the Statistical Analysis for the Data of this Research .