مقارنة بين المكونات الرئيسة وطريقة المربعات الصغرى الجزئية لمعالجة مشكلة التعدد الخطي مع التطبيق على معمل السمنت

فرح عبد الغني يونس الصالح

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

(تاريخ الاستلام: ١٩ / ١٠ / ٢٠١٠ ---- **تاريخ القبول:** ١٦ / ٣ / ٢٠١١)

الملخص:

تعتبر مشكلة التعدد الخطي مشكلة خطيرة في مسائل تحليل الانحدار لذلك تم تسليط الضوء على هذه المشكلة بشكل اكثر دقة وايجاد الحلول لها ، وفي هذا البحث تم مقارنة المكونات الرئيسية pc ومقدرات المربعات الصغرى الجزئية pls والتي تعتبر احدى طرق معالجة مشكلة التعدد الخطي، لاختبار افضل طريقة بالاعتماد على معيار اقل مربعات الخطأ MSE كمعيار للمقارنة بين المقدرات

الجانب النظري

ان تحليل الأنحدار من الطرق الاحصائية الواسعة الأستخدام وان الهدف الأساسي منها هو السعي الدائم لمعرفة وتفسير العلاقات المختلفة بين الظواهر وذلك بتحديد العلاقات بين المتغيرات ، والذي يوضح العلاقة بين متغير واحد يدعى بالمتغير التابع (متغير الاستجابة (Response variable) ومتغير او متغيرات أخرى يعتقد أنها تسبب التباين في المتغير التابع وتسمى بالمتغيرات التوضيحية (التنبؤية) (predicted variables) كذلك يصف العلاقة بين المتغير المعتمد والمتغيرات التوضحية على هيئة نموذج رياضي وان دقة هذا النموذج الرياضي تعتمد على فروض التحليل حيث ان بعض هذه الفروض مرتبط بالعلاقة الدالية والبعض الاخر مرتبط بالمتغير العشوائي ، وهذه الفروض عدة واهمها مشكلة التعدد الخطي حيث أكددت العديد من البحوث والدراسات الاهتمام الكبير بمشكلة التعدد الخطي وايجاد الحل المناسب لها واهمها المكونات الرئيسية التي تعود الى العام ١٩٠١ عندما اقترحها karl person بوصفها وسيلة لحل مشكلة تعدد العلاقة الخطية واعتمدها طريقة استكشافية يمكن الاستفادة منها للتوصل الى تفسير وفهم العلاقة المتداخلة بين المتغيرات . ويعرف تحليل المكونات الرئيسية بانه اسلوب يهدف الى ايجاد عوامل factors اوتوليفات خطية تسمى بالمكونات الرئيسية قليلة مشتقة من المتغيرات الاصلية لتحل محلها بحيث تكون مؤهلة لتفسير معظم التباين الكلي للقيم الاصلية وتكون هذه المكونات الرئيسية متعامدة أي لايوجد ارتباط فيما بينها ويمكن كتابة المكونات الرئيسية كالاتي :

$$pc_i = a_{1i}\,x_1 + a_{2i}\,x_2 + \ldots\ldots\ldots + a_{mn}\,x_m$$

$$pc_i = \sum_{j=1}^{m} a_{ji}\,x_j \qquad (i, j = 1,2\ldots m)$$

حيث ان

Pc_i : تمثل المكون الرئيسي

a_{ij} : يمثل معامل j في المكون الرئيسي i الذي يمثل قيم المتجهات المميزة a_{ji} المرافقة للجذور المميزة λ_j للمصفوفة المستخدمة وباستخدام اسلوب المصفوفات فان Pc =X A وان المصفوفة A في الصيغة اعلاه اعمدتها تمثل المتجهات المميزة المرافقة للمصفوفة المستخدمة والمرتبة وفقاً لمقادير الجذور المميزة $\lambda_1 \rangle \lambda_2 \rangle \ldots\ldots \lambda_m$ وان كل عمود من الاعمدة المصفوفة A يمثل احدى المكونات الرئيسية وتعتمد فكرة احتساب المكونات الرئيسية على خصائص الجذور المميزة ومايرافقها من المتجهات المميزة لمصفوفة الارتباط او مصفوفة التباين والتباين المشترك للمتغيرات الاصلية x_s وحسب وحدات قياسها من حيث كونها متشابهة او مختلفة . حيث ان (s) تمثل جميع المتغيرات المدروسة من $x_1, x_2, \ldots\ldots x_n$ فاذا كانت المتغيرات x_s لها نفس وحدات القياس ففي هذه الحالة تكون معاملات المكونات الرئيسيةa_{ji} قيم المتجهات المميزة لمصفوفة التباين والتباين المشترك v للمتغيرات x_s المتمثلة بـ :

$$v = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mm} \end{bmatrix}$$

ولايجاد الجذور المميزة λ نقوم بطرح من القيم القطرية للمصفوفة v ثم نجعل قيمة محدتها تساوي صفر فنحصل على المعادلة المميزة للمصفوفة v أي

$$|v - \lambda I| = 0$$

وشكل المعادلة المميزة هو متعدد الحدود (polynomial) في λ من الدرجة m أي

$$\lambda^m + C_{m-1}\lambda^{m-1} + \ldots\ldots\ldots\ldots + C_1\lambda + C_0 = 0$$

وعند حل هذه المعادلة سنحصل على m من الجذور $\lambda_1 \lambda_2 \ldots\ldots \lambda_m$ وهذه الجذور ترتب بحيث λ_1 هو اكبر قيمة ويليه λ_2 وهكذا $\lambda_1 \rangle \lambda_2 \rangle \ldots\ldots \lambda_m$ وان لكل جذر مميز λ_j قيمة(a_j) مميز خاص يقابله ولايجاد هذه المتجهات تستخدم المعادلة

$$(v - \lambda I)a = 0$$

ونختار a_{ji} بحيث تساوي قيم المتجة المميز القياسي أي تكون $a'a = 1$ وعندئذ تدخل a_{ji} كمعادلة للمتغيرات x_j للمكونات الرئيسية اما عندما يكون للمتغيرات x'_s وحدات قياس مختلفة يستحسن تحويل x الى متغيرات قياسية لها وسط حسابي = صفر

ISSN:1813 –

وتباين =١ قبل إيجاد a_{ji} يتم اختيار المكونات الرئيسية المؤثرة تأثير معنوياً باختبار النسبة التجميعية للتباين المفسر لكل مكون كما ذكر (الراوي –١٩٨٧) ان عدد المكونات الرئيسية المختارة يكون بعدد الجذور المميزة الاكبر من واحد $(\lambda > 1)$ اما بالنسبة لطريقة المربعات الصغرى الجزئية (pls) فقد ظهرت في العديد من الدراسات فيها الكيمياء والعلوم الاجتماعية ويعد الباحث (wold;1966) اول من وجد هذه الطريقة واصبحت شائعة الاستخدام في الطب وخاصة في المعالجات السريرية حيث قلة عدد المشاهدات (عدد المرضى) مع كثرة عدد المتغيرات (x'_s) المتمثلة بأغراض المرض مع عدد المتغيرات المعتمدة (y'_s) والمتمثلة بالمستوى الصحي للمريض مع تحسن حالته الصحية ونتائج اخرى كما طورت هذه الطريقة من قبل (Friedman;1993) . تعتمد طريقة انحدار المربعات الصغرى الجزئية على مصفوفة التباين والتباين المشترك cov(x,y) اذ ان هذه الطريقة تسمح بتحديد العوامل بالمتغيرات الصماء (latent variables) والتي بدورها تكون افضل نموذج للمتغير المعتمد (y'_s) ويمكن وصفها بالشكل التالي : ليكن لدينا مدور مصفوفة المتغيرات التنبؤية (x'_s) مضرورية بمتجه عشوائي من U أي ان :

$$X'U = \begin{bmatrix} \sum x_{i1} U_i \\ \sum x_{i2} U_i \\ \vdots \\ \vdots \\ \sum x_{ip} U_i \end{bmatrix}$$

ومن خلالها يمكن الحصول على مصفوفة الـ β والتي توضح نوع وشكل العلاقة بين كل متغير معتمد مع كل متغير تنبؤكالاتي

	y_1 y_2 ..... y_j
x_1	β_{11} β_{12} ..... β_{1j}
x_2	β_{21} β_{22} ..... β_{2j}
$\vdots$	$\vdots$
x_k	β_{k1} β_{k2} ..... β_{kj}

ان طريقة المربعات الصغرى الجزئية عبارة عن تركيبة خطية للمربعات الصغرى لمصفوفة الارتباط والتباين المشترك بين المتغيرات التنبؤية والمتغيرات المعتمدة ، والجزء الذي تعتمد عليه طريقة المربعات الصغرى الجزئية في مصفوفة الارتباط والتباين هو جزء التقاطع Cross Block أي الارتباطات بين المتغيرات التنبؤية والمتغيرات المعتمدة . كما تقدم هذه الطريقة عوامل Factor scores على شكل مجموعات خطية بين المتغيرات التنبؤية الاصلية ، بذلك سوف لايكون هناك ارتباطات بين عوامل المتغيرات المستخدمة في النموذج الانحدار التنبؤي يمكن توضيح عمل طريقة المربعات الصغرى الجزئية من حيث حساب معلمات الانحدار ومصفوفة المتغيرات الصماء T والاعمدة المحملة P وكخطوة اولية يتم تحويل مصفوفة المتغيرات التنبؤية X ومصفوفة المتغيرات المعتمدة Y الى الصيغة القياسية بحيث تصبح E,F على التوالي وخطوات هذه الخوارزمية كما يلي (Abdi,2003) .

١ – اخذ قيم ابتدائية عشوائية للمتجه U

٢– ايجاد المتجه W من خلال العلاقة $W = E'U$

٣– ايجاد المتجة t وهو احد الاعمدة المصفوفة T للمتغيرات المستخلصة من المصفوفة X أي ان t_{old} =Ew

٤–تحويل المتجه t الى صيغة normalized بضربه ب $\frac{1}{\|t\|}$

$$t_{new} = \frac{1}{\|t\|} . t_{old}$$

$$\|t\| = \sqrt{t_1^2 + t_2^2 + t_3^2 + ... + t_n^2}$$

٥– ايجاد المتجه الموزون c للمصفوفة Y . $c = F't$

٦–ايجاد المتجه U وهو احد اعمدة المصفوفة U للمتغيرات المستخلصة للمصفوفة y أي ان

$$U_{old} = FC$$

٧–تحويل المتجه U الى صيغة normalized بضربه ب $\frac{1}{\|U\|}$

$$U_{NEW} = \frac{1}{\|U\|} . U_{old}$$

$$\|U\| = \sqrt{U_1^2 + U_2^2 + U_3^2 + ... + U_n^2}$$

٨– مقارنة قيم t_{new} في الخطوة (4) مع قيم U_{new} في الخطوة (7) فاذا كانت القيم متقاربة يتم حساب قيم β في حالة عدم التقارب يتم حساب قيمة المحمل P للمصفوفة X من خلال المعادلة التالية .

$$P = E't$$

بعد ذلك يتم حساب المصفوفتين الجديدين E_1,F_1 من طرح المتجه t من كل من المصفوفتين E,F وكما يلي :–

$$E_1 = E - tp'$$

$$F_1 = F - btc'$$

بالتعويض عن قيمة P بالمعادلة $P = E't$

وبالتعويض عن قيمة C بالمعادلة $C = F't$

$$E_1 = E - t(E't)'$$

$$= E - ttE$$

$$F = F - bt(F't)'$$

$$= F - btt'F$$

بعد ذلك تعاد الخوارزمية من جديد بنفس الخطوات ولكن باستخدام المصفوفتين الجديدتين F_1,E_1 مع العلم ان في كل عملية تكرار يتم مقارنة قيم المتجه (t) للدورة الاخيرة مع قيم المتجه (t) للدورة السابقة وتستمر عملية التكرار الى ان يتم الحصول على جميع اعمدة المصفوفة T (والتي عددها مساو الى رتبة المصفوفة x) ومن ثم يتم ايجاد مصفوفة المعلمات β وكما يلي

ISSN:1813 –

$$\beta = E'U(TEEU)^{-1}T'F$$

الجانب التطبيقي :

ان البيانات في هذا البحث تم الحصول عليها من كتاب (DRAPER & SMTTH , 1966) حيث تمثلت باربعة عناصر أساسية كمواد أولية لانتاج مادة السمنت والتي تمثل متغيرات مستقلة يحتوي كل متغير على (١٣) مشاهدة ومتغير معتمد y كمايلي :

$$y_i = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_4 + ei$$

وكما موضحة بالجدول الاتي :

جدول (١)

الومينات الكالسيوم الثلاثي	$X_1 : 3\ Ca\ O\ .\ AL_2O_3$
سليكات الكالسيوم الثلاثي	$X_2 : 3\ Ca\ O\ .\ S_iO_2$
الوهينات الكالسيوم الرباعي الحديد يكسي	$X_3 : 4\ Ca\ O\ .\ AL_2O_3\ .Fe_2\ O_3$
سليكات الكالسيوم الثنائي	$X_4 : 2\ Ca\ O\ .\ S_iO_2$

Y : الحرارة المنبعثة (سعرة / غم) بالنسبة المئوية لوزن مادة الذي يصنع منها مادة السمنت e_i : الخطأ العشوائي ويكون ذي توزيع طبيعي بوسط (٥) وتباين σ^2 اجري تحليل الانحدار بصورة عامة بين المتغير المعتمد y مع بقية المتغيرات المؤثرة الاربعة بأستخدام البرمجة الجاهزة minitab v_{15} وكانت نتائج التحليل بالشكل التالي :

وكانت نتائج التحليل بالشكل التالي :

$$y_i = 74.0 + 1.41x_1 + 0.390x_2 - 0.031x_3 - 0.266x_4$$

جدول (٢) قيم معلمات الانحدار B والـ VIF والمختبر الاحصائي T

predictor	coef	sEcoef	T	P	VIF
constant	73.97	72.39	1.02	0.337	-
X_1	1.4111	0.7695	1.83	0.104	38.5
X_2	0.3896	0.7478	0.52	0.616	254.4
X_3	-0.0314	0.7797	-0.04	0.969	46.9
X_4	-0.2656	0.7326	-0.36	0.726	282.5

S=2.527 R-sq=98.1 % R-sq (adj)=97.2%

وللتأكيد على وجود تداخل خطي بين المتغيرات التنبؤية فقد تم استخدام بعض اساليب للكشف عنها منها :

١-١ استخدام عامل تضخم التباين Variance Inflation (VIF)

من ملاحظة قيم عوامل تضخم التباين في الجدول اعلاه نجد انها مرتفعة جدا ولقد (marquardt , 1970) على وجد تعدد العلاقة الخطية بين المتغيرات التنبؤية في حالة كون قيم العوامل تضم التباين VIF اكبر من (٤) او (١٠) ومن خلال التحليل السابق يلاحظ ان هذه القيم جداً مرتفعة وهذا يدل على وجدود تداخل خطي كبير بين المتغيرات التنبؤية .

٢-١ ايجاد محدد المصفوفة $|x'x|$

اقترح هذا المعيار من قبل (Mason & Webster , 1975) اذ ينص على انه اذا كان ٠= $|x'x|$ فهذا يدل على وجود تداخل تام بين المتغيرات التنبؤية اما اذا كان١ = $|x'x|$ فهذا يدل على استقلالية المتغيرات التنبؤية فيما بينها . اما اذ كانت قيمة $|x'x|$ بين الصفر والواحد فهذا يدل على وجود تداخل خطي شبه تام ومن خلال القيم الذاتية للمصفوفة $(x'x)$ تم حساب محدد المصفوفة $(x'x)$ والذي كان يساوي

$$|x'x| = \prod_{i=1}^{P} Lj = 0.000014$$

وهذه القيم صغيرة جداً وقريبة من الصفر وهذا دليل اخر على وجود تداخل خطي كبير بين المتغيرات التنبؤية .

٣-١ استخدام عناصر مصفوفة الارتباط الواقعة خارج القطر :

اقترح هذا الاسلوب من قبل (Gunst & Mason,1980) حيثُ من خلال مصفوفة الارتباط $(x'x)$ والموضحة ادناه

$$(x'x) = \begin{bmatrix} 1.000 & 0.229 & -0.824 & -0.245 \\ 0.229 & 1.000 & -0.139 & -0.973 \\ -0.824 & -0.139 & 1.000 & 0.030 \\ -0.245 & -0.973 & 0.030 & 1.000 \end{bmatrix}$$

نلاحظ ان هنالك ارتباط تام بين جميع المتغيرات مما يدل وجود تداخل خطي كبير بين هذه المتغيرات والذي يؤدي الى عدم دقة نتائج التحليل . كذلك نلاحظ دليل اخر على التداخل بين المتغيرات التنبؤية هو قيمة R^2 أي تمثل (معامل التحديد) حيث ظهرت كبيرة جداً

٢- طريقة المكونات الرئيسة

لأجل معالجة مشكلة التعدد الخطي تم استخدام طريقة المكونات الرئيسية بالاعتماد على مصفوفة التباين والتباين المشترك لحساب معاملات المكونات الرئسية aij كالاتي :-

$$V = \begin{bmatrix} 34.603 & 20.923 & -31.051 & -24.167 \\ 20.923 & 242.141 & -13.878 & -253.417 \\ -31.051 & -13.878 & 41.026 & 3.167 \\ -241.167 & -253.417 & 3.167 & 280.167 \end{bmatrix}$$

ثم نجد الجذور المميزة وهي :

$$\lambda_1 = 517.80$$
$$\lambda_2 = 67.50$$
$$\lambda_3 = 12.41$$
$$\lambda_4 = 0.24$$

ومن الجذور المميزة نحسب مصفوفة المتجهات المميزة:

$$\begin{bmatrix} -0.068 & 0.646 & -0.567 & 0.506 \\ -0.679 & 0.020 & 0.544 & 0.493 \\ 0.029 & -0.755 & -0.404 & 0.516 \\ 0.731 & 0.108 & 0.468 & 0.484 \end{bmatrix}$$

$$pc_1 = -0.068x_1 - 0.679x_2 + 0.021x_3 + 0.731x_4$$
$$pc_2 = 0.646x_1 + 0.020x_2 - 0.755x_3 + 0.108x_4$$
$$pc_3 = -0.567x_1 + 0.544x_2 - 0.404x_3 + 0.468x_4$$
$$pc_4 = 0.506x_1 + 0.493x_2 + 0.516x_3 + 0.484x_4$$

وتكون المكونات الرئيسية هي :

حيث نختار الجذور المميزة الاكبر من واحد وهي الجذور (الاول والثاني والثالث) ثم نحسب معادلة الانحدار المتعدد وهي كالاتي :

$$y = 1.02557 + 0.01393pc1 - 0.42407pc2 - 0.63449pc3$$

والان نعوض عن كل من Pc_1 و Pc_2 و Pc_3 في معادلة الانحدار المتعدد اعلاه فنحصل على المعادلة الانحدار الاتية :

$$y = 1.02557 + 2.15701x_1 + 1.15367x_2 + 0.73568x_3 + 0.4823x_4$$

وان جدول تحليل التباين لـ y كالاتي

جدول (٣) تحليل التباين لانتاج السمنت بطريقة (pc)

souree	DF	SS	MS	F	P
Regession	3	2664.18	888.06	151.54	0.000
Residnal	9	52.74	5.86		
Error	12	2716.92			
Total					

٣- طريقة المربعات الصغرى الجزئية (Pls)

ان الطريقة الثانية لاجل معالجة مشكلة التعدد الخطي تم استخدام طريقة المربعات الصغرى الجزئية (pls) التي تم توضيحها بالجانب النظري وكانت النتائج كما يلي:-

جدول (٤) تحليل التباين لانتاج السمنت بطريقة (pls)

souree	DF	SS	MS	F	P
Regession	4	26665.83	666.458	104.36	0.000
Residnal Error	8	51.09	6.386		
Total	12	2716.92			

وللتعرف على مدى معنوية العلاقة الخطية المقترحة واختبار مدى تأثير المتغيرات التوضيحية على المتغير المعتمد . اختبرت الفرضية الخاصة نموذج الانحدار كالاتي :

$$H_0 : B_1 = B_2 = B_3 = B_4$$

$$H_1 : B_1 \neq B_2 \neq B_3 \neq B_4$$

اذ تبين من نتائج جدول (٣) رفض فرضية العدم وقبول الفرضية البديلة عند مستوى معنوية (α) ١% وهذا يعني ان هناك اختلافات معنوية بين المتغيرات التوضيحية على المتغير المعتمد ولايوجد مشكلة تعدد العلاقة الخطية وكذلك الحالة بالنسبة الى نتائج جدول (٤) رفض فرضية العدم وقبول الفرضية البديلة عند مستوى معنوية (α) ١% أي انه لايوجد مشكلة تعدد العلاقة الخطية .

الاستنتاجات

لقد تبين اهمية دراسة مشكلة تعدد العلاقة الخطية وتأثيرها على نتائج تحليل الانحدار

١ – تم اختبار فروض التحليل للبيانات واستنتجنا من خلال الاختبارات بأن البيانات تعاني من مشكلة التعدد الخطي حيث تم كشفه وفق اختبار تصخم التباين (VIF) ومصفوفة الارتباط حيث ظهرت ارتباط تام بين المتغيرات وكذلك عن طريق معامل الارتباط حيث ظهرت قيمة كبيرة جداً

٢ – في بحثنا هذا تم معالجة مشكلة تعدد العلاقة الخطية بطريقتين المكونات الرئيسية PC المربعات الصغرى الجزئية (PLS) حيث تبين من خلال التحليل الاحصائي Mse متوسط مربعات الخطأ ان طريقة المكونات الرئيسية اكثر كفاءة من طريقة المربعات الصغرى الجزئية تمتلك اقل قيمة لـ MSE .

٣- كما تبين من خلال التحليلات الاحصائية والمقارنة بين قيمة F المحسوبة مع F الجدولية ان هنالك اختلافات معنوية بين المتغيرات التوضيحية على المتغير المعتمد وكما اظهرت الدراسة انه لايوجد مشكلة تعدد خطي .

المصادر

١.الراوي ، خاشع محمود ، ١٩٨٧ ، " المدخل الى تحليل الانحدار " ، دار الكتب للطباعة والنشر جامعة الموصل .

1- Abdi , H . (2003). Partal least squares (pls) Regression .In Lewis – Beck , M . , A . Bryman , & T . Futing , editors , Encyclopedia of social oaks : Sage

2- Draper , N . R . and smtth , H . (1966) , ((Applied Regression Analysis)) , University of Wisconsin and Rensselaer polytechnic Institute .

3- Frank , I . E . and friedman , J . H . (1993) . ((Astatistical view of chemometrcs Regression Tools)) Technometrics , 35 , pp (109-148)

4- Marquardt , D . W . (1970) , ((Generalized Inverse , Ridge Regression , Biased Linear Estimation and Nonlinear Estimation)) , Technometrics , Vol . 12 , pp (591-612) .

5- Mason ,R.L.,Gunst,R.F. and Webster, I.T., (1975), "Regression Analysis and problems of Multicollinearity", comm. . In statistics .VOL.(4) No.(3) pp(277-292).

6- Saikat Maitra and Jun yan , (2008) , ((Principle component Analysis and Partial Least squares : Two Dimension Reduction Techiques for Regression ,)) casualty Actuarial society , pp (79-90).

٧- Wold , H . (1966) ((La Regression PLS . Paris : Technip Iterative Least squares)) , In P . R . Kishnaiaah (Ed) . Multivairte Analysis , New york, Academic Press . PP (391 -420).

Comparision between principal components and partial least square method in multicollinrarity

Farah Abdul – Ghana y. Al – Saleh
Dept. of statistic , Coll. Of comp.& math. Science, Mosul University, Mosul, Iraq
(**Received:** 19 / 10 / 2010 ---- **Accepted:** 16 / 3 / 2011)

Abstract: -
Multicollinearity is considered as one of the serions problem in regression analysis , so in this paper we give a big accuracy intention to this problem and to find its solutions .We compare between pc and pls , which is considered as one of important methods of multicollinraity MSE criterion is used for choosing the best estimation Method