مقارنة بين طرائق المربعات الصغرى الجزئية و المركبات الرئيسية باستعمال المحاكاة

م.د. رباب عبد الرضا صالح/ كلية الادارة والاقتصاد/ جامعة بغداد

المستخلص

تعد طريقة المركبات الرئيسة والمربعات الصغرى الجزئية من الطرائق المهمة في تحليل الانحدار حيث ان الاثنان تستعملان لتحويل مجموعه من المتغيرات ذات الارتباط العالي الى مجموعة من المتغيرات المستقلة الجديدة تعرف بالمركبات وتكون هذه المركبات خطية متعامدة مستقلة بعضها عن البعض الاخر باستعمال تحويلات خطية ويستعمل الاثنان ايضا في تخفيض الابعاد .

تم في هذا البحث استعمال طريقة المربعات الصغرى الجزئية باستعمال خوارزمية التكرار غير الخطي للمربعات الصغرى الجزئية Non-linear Iterative partial least squares NIPALS(PLS1) وطريقة انحدار المركبات الرئيسية بخوارزمية تجزئة القيم المفردة Singular value decomposition(SVD))).

اذ تم اجراء المقارنة للطريقتين المذكورتين آنفا من خلال تجارب المحاكاة عندما يتوزع الخطأ توزيعا طبيعيا لحجوم عينات وابعاد متغيرات مختلفة ، واتضح من خلال المقارنة ان طريقة المربعات الصغرى الجزئية افضل من طريقة المركبات الرئيسية في حالة كون عدد المشاهدات اكبر من عدد المتغيرات وكذلك في حالة كون عدد المتغيرات اكبر من عدد المشاهدات. .

المصطلحات الرئيسية للبحث/ انحدار المربعات الصغرى الجزئية - انحدار المركبات الرئيسية - المتغيرات الكامنة - تقليص الابعاد - التعدد الخطي .

مجلة العلوم
الاقتصادية والإدارية
المجلد ٢٢ العدد ٨٧
الصفحات ٥٠-٧١

١ -١ المقدمه

يعد أنحدار المركبات الرئيسة والمربعات الصغرى الجزئية اكثر الطرائق اهميه في الانحدار اذ ان هاتين الطريقتين تعالجان مشكلة التعدد الخطي في البيانات وكذلك يطبقان عندما يكون عدد المتغيرات اكثر من عدد المشاهدات .

ان تحليل المركبات الرئيسة أقترح من قبل الباحث Pearson عام (1901) وطور من قبل الباحث Hoteling عام (1933) وطبق من قبل الباحث Rao عام (1964) ثم Cooley و Morrison وغيرهم من الباحثين الذين طبقوا المركبات الرئيسة [6] .

اما طريقة PLS اول من طبقها الباحث Hermon Wold عندما تكون المتغيرات التوضيحية على درجة عالية من الارتباط وايضا عندما تكون المتغيرات اكثر من المشاهدات وطورت من نفس الباحث لتطبق خوارزمية NIPALS عام (١٩٧٣) وبعدها جاءت دراسات عديدة وقدمت خوارزميات اخرى نذكر منها خوارزمية PLSF المنسوبة الى Manne عام (1987)وخوارزمية non-orthogonalized scores المنسوبة الى Martens عام (1987)وخوارزمية SIMPLS المنسوبة الى De Jong عام (١٩٩٣) وخوارزمية improved kernel PLS المنسوبة الى Dayal عام (١٩٩٧) وخوارزمية Orthogonal projections to latent structures (O-PLS) المنسوبة Trygg andWold عام (2002) وغيرها من الخوارزميات [3].

وقد ذكرت عدة دراسات كان اساسها المقارنة بين طريقتي PLS وPCR ففي عام (٢٠٠٢) قدم الباحثان Yeniay Goktag& مقارنة بين طرائق الانحدار انحدار المركبات الرئيسة (PCR) المربعات الصغرى الجزئيه (PLS) ، انحدار الحرف(RR)، المربعات الصغرى الاعتياديه (OLS) وتوضحت من خلالها ان طريقة PLS تعطي اقل متوسط مربعات الخطا مع العدد نفسه من المركبات بالمقارنه مع PCR اي ان لها القابلية للتنبؤ للانموذج افضل مع عدد اقل من العوامل وفي عام (2008) قدم الباحثان Saikat& jun دراسة بينوا من خلالها ان الطريقتين لهما الهدف نفسه في الانحدار من تقليل الابعاد وحل مشكلة الارتباط العالي بين المتغيرات التوضيحية وبين انه طريقة PLS اكثر كفاءة من طريقة PCA لمعالجة تخفيض الابعاد وفي عام (٢٠١٠) قدم الباحثون Shaho واخرون مقارنة بين الطريقتين وضحوا من خلالها ان طريقة PLSR افضل للتنبؤ للاستجابة وان طريقة PCR تعطي اقل الاخطاء القياسية لمقدرات معاملات الانحدار وهذا ما اكده الباحث Ramzan في عام (2010) وغيرها من الدراسات التي اوضحت ان PLS لها قدرة تنبؤية أفضل مع عدد أقل من العوامل بالمقارنة مع PCR من خلال دراسات المحاكاة [15,14,11,5] وفي عام (٢٠١٢) قدم الباحث حسين دراسة تطبيقية قارن فيها طريقتي PCR و PLSمن خلال اخذ مكونين اثنين تم من خلالها بناء أنموذج تمدد الاسمنت على العوامل المؤثرة عليه اي ان الدراسة تناولت بيانات حقيقية واحدة فقط في حالة كون عدد المشاهدات اكبر من عدد المتغيرات باستعمال معيار المقارنة معامل التحديد واتضح من خلال المقارنة ان طريقة PLS كانت افضل من طريقة PCR لتلك البيانات [1] .

البحث الحالي تم فيه معالجة مشكلة التعدد الخطي و تقليص الابعاد في متعدد المتغيرات من خلال تطبيق طريقتي PCRو PLS والمقارنة بينهم من خلال تجارب محاكاة يتم فيها توليد بيانات في حالة الاخطاء العشوائية تتوزع توزيع طبيعي شملت جميع الحالات اي في حالة عدد المشاهدات اكبر من عدد المتغيرات ويوجد ارتباط عال بين المتغيرات التوضيحية وفي حالة عدد المتغيرات اكبر من عدد المشاهدات ولعدة احجام وابعاد مختلفة وتم من خلالها التغير في عدد المكونات من مكونين الى عشر مكونات وذلك لبيان الطريقة التي لها القدرة التنبؤية الافضل من خلال جذر متوسط مربعات الخطأ للنموذج وتم تعميم النتائج في ضوء هذه الدراسة.

١-٢ هدف البحث

سيتم في البحث معالجة مشكلة التعدد الخطي في حالة عدد المشاهدات اكبر من عدد المتغيرات وتخفيض الابعاد بين المتغيرات التوضيحية والتي تبرز في حالة كون عدد المتغيرات التوضيحية اكبر من عدد المشاهدات ويتم ذلك من خلال مقارنة طريقتي انحدار المركبات الرئيسة PCR والمربعات الصغرى الجزئية PLS مستعملين المحاكاة باختلاف الابعاد وحجوم العينات.

٣-١ مشكلة البحث

في الدراسات الخاصة بمتعدد المتغيرات وضمن الاساليب الاحصائية المستعملة في هذا التخصص كانت تأخذ حالة واحدة فقط وهي عندما يكون عدد المتغيرات اقل من عدد المشاهدات يمكن حل جميع التطبيقات الخاصة بمتعدد المتغيرات وبهذا تكون مصفوفة البيانات X ((full rank غير مفردة لكن في بعض الاحيان توجد مشاكل في البيانات مثل التعدد الخطي يتم استعمال عدة طرائق لمعالجتها مثل طريقة المركبات الرئيسية والمربعات الصغرى الجزئية وغيرها.

في اطار البحوث الحديثة فقد توجه عدد من الباحثين الى دراسة الحالة التي يكون فيها عدد المشاهدات اقل من عدد المتغيرات حيث ان مصفوفة X ستكون مفردة ولا يمكن تطبيق الطرائق المألوفة لدينا وعليه تطرق الباحثون الى هذه الحاله من خلال استعمال الطريقتين المذكورتين انفاً.

٢-٢ الجانب النظري

في الانحدار الخطي المتعدد MLR يكون حل المربعات الصغرى نسبة الى دالة الانحدار الاتيه: [6]

$$\underline{Y} = X\underline{\beta} + \varepsilon \qquad \dots \qquad (2-1)$$

حيث ان :-

$\underline{Y}$: موجه المتغير المعتمد ببعد nx1

X : مصفوفة المتغيرات التوضيحية ذات رتبة nxp

$\underline{\beta}$ موجه معلمات دالة الانحدار غير المعلومه ببعد px1

$\underline{\varepsilon}$: موجه الاخطاء العشوائية ببعد nx1

حيث ان تقدير المعلمات يعطي بالصيغة الاتيه:-

$$\underline{\beta} = (X'X)^{-1}X'y \qquad \cdots \qquad (2-2)$$

تكمن المشكلة عندما تكون المصفوفة Xمفردة عندما يكون عدد المتغيرات التوضيحية تزيد عن عدد المشاهدات عندما يكون هناك ارتباط عالي بين المتغيرات التوضيحية مما يقود الى محاولة معالجة مثل هكذا حالات بأستعمال عدة طرائق منها PLS , PCR حيث ان الفرق الرئيسي بين الطريقتين هو في تشكيل المركبات scores ففي طريقة المربعات الصغرى الجزئية تعتمد على ايجاد المعلومات بواسطة تعظيم مصفوفة التباين والتباين المشترك بين X و $\underline{Y}$ اما طريقة تحليل المركبات الرئيسة فتعتمد على ايجاد المعلومات بين المتغيرات التوضيحية X ولاتوجد معلومات حول متغير الاستجابة $\underline{Y}$ [9و13]

2-2-1 أنحدارالمربعات الصغرى الجزئية (PLSR)

توجد عدة خوارزميات فيما يتعلق بالمربعات الصغرى الجزئية أذ انها تعتمد على خطوتين اساسيتين الاولى هي ايجاد المتغيرات الكامنه latent variable بين X و $\underline{Y}$ من خلال تعظيم مصفوفة التباين والتباين المشترك والخطوة الثانية هي انحدار $\underline{Y}$ على المركبات t ومن الخوارزميات التي استعملت هي NIPALS (PLS1,PLS2) حيث ان PLS1 تستعمل عندما يكون متغير الاستجابة متجه اما PLS2 فتستعمل عندما يكون متغير الاستجابة مصفوفة وهي اولى الخوارزميات لحل مشكلة PLS ثم تليها خوارزمية Kernel حيث الاثنان تعطيان نفس النتائج لكن الفرق هو في كيفية حساب المركبات ففي الاول يتم بصورة تكرارية اما الخوارزمية الاخرى يتم عن طريق ايجاد المتجهات الذاتية ، وخوارزمية PLS1 تعطي نفس نتائج خوارزمية SIMPLS والاختلاف هو في كيفية استعمال الـ deflation (تفريغ البيانات) ففي الاولى يتم عن طريق X و $\underline{Y}$ اما في SIMPLS يكون لمصفوفة التباين والتباين المشترك .

نفرض لدينا المصفوفة $X_{n.p}$ والمتجه $\underline{Y}_{n.1}$ طريقة المربعات الصغرى الجزئية تعتمد على النموذج الثنائي بين X و $\underline{Y}$ وكالاتي [4و 9و 12] :

$$X = TP' + E \quad \dots \quad (2-3)$$

$$\underline{Y} = Uq' + f \quad \dots \quad (2-4)$$

حيث

T مصفوفة x-score ذات رتبة nxr

U مصفوفة Y - score ذات رتبة nxr

P مصفوفة x- loading ذات رتبة pxr

q متجه Y-loading ببعد 1xr

E مصفوفة x- residual ذات رتبة nxp

$\underline{F}$ متجه Y - residual ببعد nx1

المصفوفة P والمتجه $\underline{Q}$ له r من الاعمدة وهو محدد بما يأتي

(r< min (n ,p))

والعلاقة الداخلية التي تربط بين scores تعطى وكالاتي

$$U = T D + H \quad \dots \quad (2-5)$$

حيث D مصفوفة قطرية ذات رتبة rxr

H مصفوفة البواقي ذات رتبة nxr

الفكرة الاساسية في المربعات الصغرى الجزئية هو في كيفية ايجاد المتجه w من مجال X والمتجه c من المجال Y بحيث ان

Max COV(Xw Yc)

$$\text{with} \|t\| = \|Xw\| = 1 \quad and \quad \|u\| = \|Yc\| = 1 \quad \dots \quad (2-6)$$

حيث ان COV هو تقدير التباين المشترك

وان t,u هي اعمدة في المصفوفتين T,U ويتم تنفيذ التكرارات بطريقة متسلسلة وهذا يعني ان المتجهات scores يتم احتسابها الواحد بعد الاخر حتى يتم استخراج كافة المتجهات الى r تحت القيد عدم الارتباط بين المتجهات وتوجد طرائق عدة لحل المعادلة (2-6) منها خوارزمية Kernel وخوارزمية SIMPLS وخوارزمية NIPALS وغيرها من الخوارزميات وفي هذا البحث تم الاعتماد على خوارزمية NIPALS (PLS1) .

٢-٢-٢ خوارزمية NIPALS(PLS1)

فيما يأتي الخطوات الاساسية لخوارزمية NIPALS(PLS1) لحساب اول مركبة [4، 8]

١- في الخطوة الاولى يتم تهيئة U_1 عن طريق $\underline{Y}$ بحيث

$$U_1 = \underline{Y} \quad \dots \quad (2-7)$$

حيث ان U_1 متجه ببعد nx1

٢- حساب X- weight بواسطة انحدار OLS

$$\underline{W1} = X'U_1 / (U_1'U_1) \quad \dots \quad (2-8)$$

حيث ان $\underline{W1}$ متجه ببعد px1

٣- حيث ان $\underline{W1}$ يكون normalize بالشكل $\underline{W1} = \underline{W1} / \|W1\|$

4- اسقاط البيانات X على x-weight لحساب x-scores وكالاتي

$$t_1 = XW1 \quad \dots \quad (2-9)$$

حيث ان t1 متجه ببعد nx1

٥- حساب y-weight بواسطة انحدار OLS

$$c_1 = \underline{Y}' t_1 / (t'_1 t_1) \quad \dots \quad (2-10)$$

حيث c1 متجه ببعد 1x1

٦- حيث ان c_1 تكون normalize بالشكل $c_1 = c_1 / \|c1\|$

٧- اسقاطات بيانات y على y-weight لحساب y-scores

$$u_1^* = y c_1 \quad \dots \quad (2-11)$$

حيث u_1^* متجه ببعد nx1

٨- نحدد u_1^* بحيث تحقق مايلي

$$\Delta u = (u\Delta)'(u\Delta) \quad \dots \quad (2-12)$$

$$u\Delta = u_1^* - u_1$$

9- اذا كانت $\Delta u < \varepsilon$ وجدنا اول مركبة و نتوقف حيث ε قيمه صغيره عدا ذلك نذهب الى الخطوة الاولى

ونستعمل u_1^* اي ان $u_1 = u_1^*$ ونذهب للخطوة (٢)

10- ايجاد X- loading بواسطة انحدار OLS وكالاتي

$$p_1 = X' t_1 / (t'_1 t_1) \quad \dots \quad (2-13)$$

حيث ان p1 متجه ببعد px1

١١- ايجاد Y-loading بواسطة انحدار OLS

$$q = \underline{Y}' u_1 / (u'_1 u_1) \quad \dots \quad (2-14)$$

حيث ان q متجه ببعد 1x1

١٢- ايجاد التداخل الخطي للمعالم بواسطة انحدار OLS

$$d_1 = u'_1 t_1 / (t'_1 t_1) \quad \dots \quad (2-15)$$

حيث ان d1 متجه ببعد 1x1

١٣- عمل تفريغ deflate الى بيانات X

$$X_1 = X - t_1 p'_1 \quad \dots \quad (2-16)$$

14- عمل تفريغ deflate الى بيانات Y

$$Y_1 = Y - d_1 t_1 c'_1 \quad \dots \quad (2-17)$$

ونستمر بالخطوات من (١-١٤) عدة مرات وباستعمال البيانات المفرغه الى X وY حتى نحصل على كل المركبات المحددة ونستطيع ان نجد معاملات الانحدار بواسطة العلاقة الاتيه

١٥- ايجاد معاملات الانحدار

$$\beta = w(p'w)^{-1} c' \quad \dots \quad (2-18)$$

حيث ان W هي مصفوفة برتبة pxr

P مصفوفة برتبة pxr

C مصفوفه برتبة rxr

٢- ٣ انحدار المكونات الرئيسية

ان تحليل المكونات الرئيسة هي تقنية تقليدية لمتعدد المتغيرات تعتمد على مصفوفة التباين بين المتغيرات التوضيحية وفيها يتم تفسير اكبر جزء من التباين عن طريق ايجاد تراكيب خطية مستقلة بعضها عن البعض الاخر وكل مركبة رئيسة هي تركيبة خطية لكل المتغيرات فالمركبة الرئيسة الاولى تشرح اكثر التباين وهكذا لبقية المركبات هناك عدة خوازميات لايجاد المركبات الرئيسية وهي Mathematics of PCA ,Jacobi Rotation , NIPALS, Singular Value Decomposition
في هذا البحث تم تناول خوارزمية PCR باستعمال تجزئة القيم المفردة (Singular value (SVD)) decomposition) [13] .

٢-٣-١ خوارزمية PCR باستعمال SVD

في تحليل المركبات الرئيسة نحصل على اول مركبة رئيسة (PC) من خلال تجزئة المصفوفة X ذات الرتبة nxp الى ثلاث مصفوفات وكالاتي [11,10,7] :

$$X = T_0 S P' \quad \dots \quad (2-19)$$

نفرض ان t = min {n,p}
حيث ان :

T_0 : مصفوفة متعامدة ذات رتبة nxt وهي مصفوفة المركبات ويتم ايجادها من المتجه المميز لـ XX'

S : مصفوفة قطرية ذات رتبة txt وهي تساوي الجذور المربعة الى القيم المميزة لـX'X او XX' حيث ان S= diag $\{\lambda_1 > \lambda_2 \quad \dots \lambda_p > 0\}$

P : مصفوفة متعامدة ذات رتبة pxt وهي مصفوفة تحميل ويتم ايجادها من المتجه المميز الى X'X
حيث ان المركبات الرئيسة في التحليل هي T يمكن ان تكون .

$$\mathbf{T} = T_0 S \quad \dots \quad (2-20)$$

$$X = T_0 S P' + \varepsilon \quad \dots \quad (2-21)$$

اما معاملات الانحدار هي

$$\beta_{PCR} = P(T'T)^{-1} T'Y \quad \dots \quad (2-22)$$
$$= P S^{-1} T'_0 Y$$

ان خوارزمية SVD تعطي المتجه المميز والقيم المميزة التي تحتاج في تحليل المركبات الرئيسية.

٣-١ الجانب التجريبي

لتحقيق الهدف من البحث تمت المقارنة بين طريقتي المربعات الصغرى الجزئية والمركبات الرئيسية في حالة حد الخطأ يتوزع توزيع طبيعي وباستعمال اسلوب المحاكاة من خلال صياغة تجارب تكون فيها عدد المشاهدات اكبر من عدد المتغيرات n>p وتوجد بها مشكلة التعدد الخطي وحسب صيغة المعادلة (٢-٣) وكذلك عدد المتغيرات اكبر من عدد المشاهدات p>n وذلك بتغير في حجوم العينات وعدد المشاهدات وعدد المركبات تم استعمال معيار المقارنة جذر متوسط مربعات الخطا في التنبؤ $RMSE_h$ وقد تم كتابة البرنامج بلغة ماتلاب(version (R2011.

٣-٢ وصف تجارب المحاكاة

تم توليد االبيانات بحجوم عينات وابعاد مختلفة وكل تجربة تم تكرارها ١٠٠٠مرة وكانت حجوم العينات في حالة n>p كالاتي (n= ٢٥,٥٠,٧٥,١٠٠,١٥٠,٢٠٠) وعدد المتغيرات التوضيحية (P=10) .
اما في حالة n≤p كانت (n=10 ,P=25) ، (n=25 ,P=50) ، (n=50 p=50) ، (n=50 , p=75) (n=75 ,P=100) (n=100 ,P=150)

وتمت المقارنة بين جذر متوسط مربعات الخطأ للطريقتين المذكورة آنفاً وحسب المقياس الاتي [10]:

$$RMSE_h = 1/m \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}} \qquad \cdots(3-1)$$

m يشير الى عدد التكرارات

وبالاعتماد على نموذج معادلة الانحدار (١) حيث تم توليد حد الخطأ العشوائي بالتوزيع الطبيعي القياسي N(0,1) وتم توليد المتغيرات التوضيحية والمتغير المعتمد بالصيغة الاتية [2] :

$$\begin{aligned} E &= N(0,1) \\ x_1 &= N(0,1) \\ x_{p-1} &= N(0,0.1) + x_1 \\ y &= x_1 + \cdots + x_p + E \end{aligned} \qquad \cdots(3-2)$$

٤ - نتائج المحاكاة

نتائج المحاكاة في حالة عدد المشاهدات اكبر من عدد المتغيرات تم توليد البيانات بحسب الصيغة (٢-٣) بوجود مشكلة التعدد الخطي حيث تم التأكد من وجود هذه المشكلة من خلال اختبار معامل تصخم التباين Variance Inflation Factor (VIF) حيث اتضح ولكل حالات المقارنة وجود الارتباط بين المتغيرات التوضيحية ومن ثم وجود مشكلة التعدد الخطي حيث ان قيمة VIF ولكل المتغيرات في كل الحالات تزيد عن ١٠ وهذا مؤشر على وجود التعدد الخطي وفيما يلي تفسير النتائج من الجدول رقم (١) الى الجدول رقم(٦) نجد ان قيمة جذر متوسط مربعات الخطأ لطريقة PLS اقل من طريقة PCR ولكافة المركبات وانه كلما ازداد عدد المشاهدات يكون التقارب جدا قليل بازدياد عدد المركبات لطريقة PLS وهذا مؤشر على ان الطريقة المذكورة آنفاً تاخذ اقل عدد من المركبات مقارنة من طريقة PCR وان قيمة جذر متوسط مربعات الخطا يتساوى عند المركبة الاخيرة لكلا الطريقتين وهذا يعني ان الطريقتين تضمنت جميع المعلومات .

نتائج المحاكاة في حالة المتغيرات اكبر من عدد المشاهدات

من الجدول رقم (٧) الى الجدول رقم (١٢) نجد ان قيمة جذر متوسط مربعات الخطأ لطريقة PLS اقل من طريقة PCR وانها تتناقص بزيادة عدد المركبات للطريقتين المنذكورتين آنفاً وبزيادة عدد المتغيرات

جدول رقم (١) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=25) و (p=10)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0252	0.0239	0.0232	0.0235	0.0231	0.0232	0.0234	0.0229	0.0233
PCR	0.0298	0.0289	0.0282	0.0277	0.0264	0.0258	0.0252	0.0238	0.0233

جدول رقم (٢) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=50) و (p=10)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0280	0.0278	0.0278	0.0278	0.0278	0.0277	0.0279	0.0277	0.0276
PCR	0.0306	0.0304	0.0300	0.0296	0.0294	0.0289	0.0286	0.0281	0.0276

جدول رقم (٣) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=75) و (p=10)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0293	0.0290	0.0291	0.0291	0.0290	0.0291	0.0291	0.0290	0.0291
PCR	0.0311	0.0307	0.0306	0.0303	0.0300	0.0298	0.0296	0.0292	0.0291

جدول رقم (4) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=100) و (p=10)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0298	0.0297	0.0297	0.0298	0.0299	0.0298	0.0299	0.0298	0.0298
PCR	0.0311	0.0310	0.0307	0.0307	0.0306	0.0303	0.0302	0.0300	0.0298

جدول رقم (5) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=150) و (p=10)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0306	0.0304	0.0303	0.0303	0.0304	0.0305	0.0305	0.0304	0.0304
PCR	0.0315	0.0312	0.0310	0.0309	0.0309	0.0308	0.0307	0.0305	0.0304

جدول رقم (٦ يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=200) و (p=10)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0307	0.0307	0.0307	0.0307	0.0306	0.0307	0.0306	0.0307	0.0307
PCR	0.0314	0.0313	0.0312	0.0312	0.0310	0.0310	0.0308	0.0308	0.0307

جدول رقم (٧) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=10) و (p=25)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9
PLSR	0.0100	0.0047	0.0022	0.0010	0.0004	0.0001	0.0000	0.0000
PCR	0.0213	0.0189	0.0167	0.0143	0.0121	0.0093	0.0056	0.0000

جدول رقم (٨) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=25) و (p=50)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0096	0.0057	0.0036	0.0023	0.0015	0.0009	0.0006	0.0003	0.0002
PCR	0.0186	0.0180	0.0174	0.0165	0.0158	0.0152	0.0146	0.0139	0.0131

جدول رقم (٩) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=50) و (p=50)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0107	0.0082	0.0070	0.0061	0.0055	0.0050	0.0046	0.0043	0.0040
PCR	0.0168	0.0163	0.0161	0.0157	0.0155	0.0151	0.0148	0.0146	0.0143

جدول رقم (10) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=50) و (p=75)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0089	0.0060	0.0044	0.0033	0.0026	0.0020	0.0015	0.0012	0.0009
PCR	0.0163	0.0160	0.0156	0.0154	0.0150	0.0147	0.0143	0.0140	0.0137

جدول رقم (١١) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=75) و (p=100)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0083	0.0057	0.0044	0.0035	0.0028	0.0023	0.0019	0.0016	0.0013
PCR	0.0149	0.0148	0.0146	0.0143	0.0141	0.0138	0.0136	0.0135	0.0132

جدول رقم (١٢) يوضح مقارنة بين طريقتي PLS وPCR باستخدام المعيار $RMSE_h$ عند (n=100) و (p=150)

عدد المركبات / الطرائق	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10
PLSR	0.0075	0.0049	0.0035	0.0027	0.0021	0.0016	0.0013	0.0010	0.0008
PCR	0.0145	0.0143	0.0141	0.0140	0.0138	0.0136	0.0134	0.0133	0.0131

٥- الاستنتاجات

من نتائج تجارب المحاكاة تم التوصل الى الاستنتاجات الاتية:-

١- في حالة عدد المشاهدات اكبر من عدد المتغيرات ولمعيار المقارنة $RMSE_h$ ولجميع المركبات نجد ان PLS لها القدرة العالية للتنبؤ مع اقل عدد من المركبات بالمقارنة مع طريقة PCR وانها مساوية لها عند المركبة الاخيرة.

٢- في حالة عدد المتغيرات اكبر من المشاهدات ولمعيار المقارنة $RMSE_h$ ولجميع المركبات نجد ان طريقة PLS تعطي اقل الاخطاء في حالة التنبؤ لمتغير الاستجابة مقارنة مع طريقة PCR.

٦- التوصيات

بناء" على ماتم التوصل اليه من استنتاجات في الجانب التجريبي يمكن ادراج التوصيات الاتية التي يراها الباحث ضروريه

١- استعمال طريقة PLS في حالة عدد المشاهدات اكبر من المتغيرات وبالعكس من اجل التخلص من مشكلة التعدد الخطي ولتخفيض الابعاد

٢- اجراء مقارنات اخرى بين الطريقتين المذكورتين آنفاً في حالة وجود بيانات ملوثة او بيانات تحتوي على قيم شاذة.

المصادر

1 -حسين ، الهام عبد الكريم ٢٠١٢"مقارنة بين استعمال أنموذج المربعات الصغرى الجزئية PLSR وانحدار المكونات الرئيسية PCR في العوامل المؤثرة على تمدد الاسمنت " ، مجلة التربية والعلم المجلد ٢٥ العدد ٢.

٢- Adnan,N., (2006) "A comparative Study on Methods For Handling Multicollinearity Problems" J. Matematika, V. 22, N.2, PP. 109_119

٣- Andersson,M.,(2009)" A comparison of nine PLS1 algorithms'. *Chemometrics*; 23: 518–529

٤- Chong,I., JunT ,C., 2005 ''Performance of some variable selection methods when multicollinearity present"Chemometrics and Intelligent Laboratory Systems 78 , 103–112

٥- Engelen, S., Hubert, M., et al. (2004) ''Robust PCR and Robust PLSR: a comparative study " statistics for industry and technology,105-117

٦- Garcia,H., Filzmoser ,P., 2011 ''Multivariate Statistical Analysis using the R package chemo metrics ''

٧- Jorgensen,B., Goegebeur,Y., '' Module 6: Principal components analysis''

٨- Jorgensen,B., Goegebeur,Y., ''Module 7: Partial least squares regression I"

9- Maitra ,S.,(2008) " Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression"
Casualty Actuarial Society
10- Mevik, B.,Wehrens,R.,(2007)''The PLS Package: Principal Component and Partial Least Squares Regression in R''
11- Romzan,S., et'al. (2010) ''prediction method for Time- series regression models with multicollinearity'' world applied sciences Journal 11(4):443-450
12-Roon,P.,Zakizadeh,J.,Chartier,S .(2014)''Partial Least Squares tutorial for analyzing neuroimaging data"
13-Varmuza,K., Filzmoser ,P., 2008 ''Introduction to multivariate statistical analysis in chemometrics'' Taylor & Francis Group, LLC.
14- Wentzell,P.D., Montoto,L.V., , 2003 ''Comparison of principal components regression and partial least squares regression through generic simulations of complex mixturesChemometrics and Intelligent Laboratory Systems 65 , 257–279
١٥- Yeniay,O. , Gokta, A., 2002," A comparison of Partial Least Squares Regression with other prediction methods "Journal of Mathematics and Statistics , Volume 31, 99-111

Comparison of Partial Least Squares and Principal Components Methods by Simulation

Abstract

The methods of the Principal Components and Partial Least Squares can be regard very important methods in the regression analysis, where they are used to convert a set of highly correlated variables to a set of new independent variables, known components and those components are be linear and orthogonal independent from each other , the methods are used to reduce dimensions in regression analysis

In this paper , we use Partial Least Squares method with Non -linear Iterative partial least squares NIPALS(PLS1) algorithm and the principal components method with Singular Value Decomposition(SVD)algorithm , the simulation experiments are conduct to compare between their methods assuming that the error is normally distributed , several combination are supposed in simulation for both sample size, number of observation, dimension, and we find that the partial least squares method is better than the Principal Components method in two case, number of observation is greater than the number of variables(n>p) and the number of variables is greater than the number of observation (p>n).

Keywords/ Partial Least Squares Regression (PLSR); Principal Components Regression (PCR); latent variables, Dimension Reduction, Multicollinearity