مقارنة بين طريقة المربعات الصغرى الجزئية وخوارزمية تجزئة القيم المفردة لتقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلة التعدد الخطي بأستعمال المحاكاة

أ.د. محمود مهدي البياتي/كلية الادارة والاقتصاد/جامعة بغداد
الباحثة/ هديل حميد شاكر/كلية الادارة والاقتصاد/جامعة بغداد

تاريخ التقديم:2018/7/8
تاريخ القبول:2018/9/12

المستخلص

يعد أنموذج الانحدار اللوجستي من النماذج الاحصائية المهمة حيث يوضح العلاقة بين المتغير التابع ثنائي الاستجابة والمتغيرات التوضيحية (التفسيرية).
أن العدد الكبير لمتغيرات توضيحية تستعمل عادة لتوضيح الاستجابة ادى الى ظهور مشكلة التعدد الخطي(Multicollinearity) بين المتغيرات التوضيحية التي تجعل تقدير معلمات النموذج ليست دقيقة.
تم في هذا البحث استعمال طرائق لتقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلة التعدد الخطي وهذه الطرائق هي طريقة انحدار المربعات الصغرى الجزئية(PLSR) و خوارزمية تجزئة القيم المفردة(SVD)، اذ تم استخدام اسلوب المحاكاة للمقارنة بين طرائق التقدير من خلال متوسط مربعات الخطأ(MSE) للأنموذج.
واتضح من خلال المقارنة أن خوارزمية تجزئة القيم المفردة (SVD) هي الافضل في تقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلة التعدد الخطي.

المصطلحات الرئيسة للبحث/ الانحدار اللوجستي، البيانات الثنائية، المربعات الصغرى الجزئية، خوارزمية تجزئة القيم المفردة، مشكلة التعدد الخطي

مجلة العلوم
الاقتصادية والإدارية
العدد 109 المجلد 24
الصفحات 458ـ 471

*بحث مستل من رسالة ماجستير

1- المقدمة Introduction

يعرف الانحدار بشكل عام بأنه أحد الأساليب الاحصائية المهمة التي تستخدم بشكل واسع جدا لتحديد وتوضيح التأثيرات بين المتغيرات التوضيحية (التفسيرية) والمتغير التابع (متغير الاستجابة) وتستخدم ايضا للتنبؤ عن قيمة المتغير التابع بدلالة المتغيرات التوضيحية بعد أيجاد معادلة الانحدار الخطية [1].

يعد أنموذج الانحدار اللوجستي الاسلوب الاحصائي المستخدم لتوفيق العلاقة بين المتغير التابع ثنائي القيمة وعدة متغيرات توضيحية أيا كان نوعها، ويسمى الأنموذج في هذه الحالة أنموذج الانحدار اللوجستي الثنائي، ويعد من الأساليب الاحصائية المهمة التي يمكن أستخدامها في كثير من مجالات الحياة، مثلا في مجال الطب والجيولوجيا والبايولوجيا والزراعة.

تكمن اهمية تحليل الانحدار اللوجستي عند مقارنته بالأساليب الاحصائية الاخرى في أن الانحدار اللوجستي هو أداة أكثر قوة، لانه يقدم اختبارا لدلالة المعاملات، كما انه يعطي الباحث فكرة عن مقدار تأثير المتغير التوضيحي على متغير الاستجابة الثنائية، فضلا عن ذلك فأن الانحدار اللوجستي يرتب تأثير المتغيرات، مما يسمح للباحث بالاستنتاج أن متغيرا ما يعد أقوى من المتغير الاخر في فهم ظهور النتيجة المطلوبة[3].

تعد مشكلة التعدد الخطي أو الارتباط المتعدد(Multicollinearity) بين المتغيرات التوضيحية، واحدة من أهم وأكثر المشكلات التي تقف عقبه أمام الباحثين عند أستخدام الانحدار اللوجستي، وهي تنشأ عندما يتضمن نموذج الانحدار أكثر من متغير توضيحي وتكون هناك علاقة ارتباط قوية بين اثنين أو أكثر من هذه المتغيرات، أو بين جميع المتغيرات.

لأهمية مشكلة التعدد الخطي في أنموذج الانحدار اللوجستي، تم اقتراح طرائق من قبل الكثير من العلماء والباحثين لمعالجة هذه المشكلة وللحصول على مقدرات دقيقة في الأنموذج اللوجستي.

وهناك العديد من الدراسات والبحوث التي تناولت موضوع أنموذج الانحدار اللوجستي والمشاكل التي يتعرض لها، وطرائق تقدير معلمات هذا الأنموذج ففي عام (2001م) ناقش الباحثان (Tormod and Bjorn-Helge)[16] مشكلة التعدد الخطي في أنموذج الانحدار اللوجستي والتحليل المميز وحلها بأستخدام المربعات الصغرى الجزئية (PLS) والمركبات الرئيسية (PC)، اذ تم التوصل الى أن المربعات الصغرى الجزئية هي أفضل طريقة لحل مشكلة التعدد الخطي من المركبات الرئيسية وتوصل ايضا الى أن المربعات الصغرى الجزئية افضل من المركبات الرئيسة في ايجاد بعض المركبات المهمة في عملية تصنيف البيانات بأستعمال الانحدار اللوجستي والدالة المميزة وفي عام (2005م) قام (Anne and Korbinian)[7] بأستخدام طريقة المربعات الصغرى الجزئية وهي أسلوب احصائي مهم في تحليل البيانات للأبعاد العالية لبيانات الالاف من الجينات وكذلك لتقليص الابعاد حيث أن المرحلة الاولى، أستخدم التصنيف بالطريقة الكلاسيكية (الانحدار اللوجستي) بأستخدام المركبات الرئيسة للمربعات الصغرى الجزئية وفي العام نفسه (2005م) درس الباحثان (LI shen and Eng) [13] طريقة المربعات الصغرى الجزئية(PLS) وطريقة تجزئة القيم المفردة (SVD) مع الانحدار اللوجستي الجزائي (Penalized Logistic Regression) اذ تم تحديد مجموعة فرعية من (16) جينة لتصنيف سرطان الدم الحاد حيث أن خطأ الاختبار على هذه المجموعة الفرعية من الجينات هو صفر تجريبيا اذ تم التوصل الى أن هذه الطرائق تعطي نتائج اكثر دقة وفي عام (2009م) قام الباحثان (Francesca and Fabio) [12] بأستخدام طريقة تجزئة القيم المفردة (SVD) في حالة أنموذج الانحدار اللوجستي لتحديد ميزه في علم التصنيف حيث توصل الباحثان الى أن طريقة تجزئة القيم المفردة (SVD) تعطي نتائج أكثر دقة.

2- مشكلة البحث Problem of the Research

في حالة كون متغير الاستجابة من النوع الثنائي أي ثنائي الاستجابة (0,1) مع متغير توضيحي واحد فأنه بالامكان استخدام أنموذج الانحدار اللوجستي في تحليل هذه البيانات، ولكن في ظل وجود متغيرات توضيحية عديدة 7 فأكثر فأنه لايمكن اعتماد النموذج اللوجستي بصيغته العادية وذلك بسبب ظهور مشاكل في المتغيرات التوضيحية، ومن هذه المشاكل هي مشكلة التعدد الخطي (Multicollinearity) لذلك يتم اللجوء الى استعمال طرائق تلائم واقع هذه البيانات.

3- هدف البحث Object of Research

يهدف هذا البحث الى معالجة مشكلة التعدد الخطي التي من الممكن أن تظهر بين المتغيرات التوضيحية من خلال أستخدام طريقة انحدار المربعات الصغرى الجزئية (Partial Least Square Regression: PLSR) وخوارزمية تجزئة القيم المفردة (Singular Value Decomposition: SVD) في الأنموذج اللوجستي للحصول على مقدرات اكثر دقه من طريقة المربعات الصغرى الاعتيادية في الأنموذج اللوجستي، وبأستخدام أسلوب المحاكاة يتم المقارنة بين طرائق التقدير من خلال معيار متوسط مربعات الخطأ MSE.

4- الجانب النظري:

1-4 أنموذج الانحدار اللوجستي: Logistic regression model

يبنى أنموذج الانحدار اللوجستي على فرض أساسي هو أن المتغير التابع (Y) ثنائي الاستجابة يأخذ أحدى القيمتين (0,1) أما النجاح (Success) بأحتمال (π_i) أو الفشل (Failure) بأحتمال $(1-\pi_i)$ لذلك يكون المتغير (y_i) يتوزع بحسب توزيع برنولي)$\mathrm{Ber}(\pi_i$ [10] .

أي أن
$$y_i \sim Ber(\pi_i)$$
i=1,2,…,n

ومن ثم فأن دالة الكثافة الاحتمالية تكون وفق الصيغة الاتية :

$$P(Y_i = y_i) = \pi_i^{y_i} (1-\pi_i)^{1-y_i} \qquad \dots(1)$$

y_i=0,1

أذ أن :

y_i متغير تابع ثنائي الاستجابة (0,1)

π_i أحتمال حدوث الاستجابة عندما $y_{i=1}$

$1-\pi_i$أحتمال عدم حدوث الاستجابة عندما $y_{i=0}$

لذلك فأن توقع المتغيرy_i يمثل أحتمال حدوث الاستجابة (π_i) وكالاتي:

$$E(y_i) = pr(y=1) = \pi_i$$

أما تباين المتغير y_i بحسب توزيع برنولي كالاتي:

$$V(y_i) = \pi_i(1-\pi_i)$$

ليكن $X_1, X_2, \dots, X_p$ مجموعة من المتغيرات التوضيحية ولتكن n تمثل عدد المشاهدات لهذه المتغيرات التي تكون المصفوفة X [5] .

$$X = (X_{ij})_{n\times p} \qquad \dots\dots\dots(2)$$

أذ أن:

i=1,2,…,n ، (n) تمثل حجم العينة.

j=1,2,…,p ، (p) تمثل عدد المتغيرات التوضيحية.

فأذا كان $y_i = [y_1, y_2, \dots, y_n]$ عينة عشوائية من المتغير ثنائي الاستجابة وأن $y_i \in \{0,1\}$ ومن ثم فأن أنموذج الانحدار اللوجستي يكتب بالصيغة الاتية:

$$y_i = \pi_i + \varepsilon_i \quad .....(3)$$

اذ أن π_i تمثل دالة الانحدار اللوجستي (احتمال الاستجابة)

$$\pi i = \frac{\exp\{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}\}}{1 + \exp\{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}\}} \quad .....(4)$$

أو يمكن كتابته بالصيغة الاتية:

$$\pi i = \frac{\exp\{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\}}{1 + \exp\{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\}} \quad .....(5)$$

$$(1 - \pi i) = \frac{1}{1 + \exp\{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\}} \quad .....(6)$$

حيث $\beta_0, \beta_1, \ldots, \beta_p$ هي معلمات للنموذج، وأن ε_i يمثل حد الخطأ العشوائي بمتوسط صفر وتباين $\pi_i(1-\pi_i)$

نلاحظ من المعادلة (5) أن شكل العلاقة بين المتغيرات التوضيحية (X_{ij}) واحتمال الاستجابة π_i لا يمكن أن يكون خطيا وانما تأخذ شكلا منحنيا.

لقد اقترح (Berkson) عام 1944م بأنه يمكن تحويل دالة الانحدار اللوجستي الى دالة خطية وبحسب الصيغة الاتية[9]:

$$\frac{\pi_i}{(1-\pi_i)} = \exp\{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\} \quad ....(7)$$

وبأخذ اللوغاريتم الطبيعي لكلا الطرفين نحصل على:

$$Z_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \quad ...(8)$$

أذ أن

(Z_i) تمثل العلاقة الخطية الناتجة من اخذ اللوغاريتم الطبيعي ل $\left(\frac{\pi_i}{1-\pi_i}\right)$ والذي يتبع التوزيع الطبيعي بمتوسط $(X_i\beta)$ وتباين $[n_i\pi_i(1-\pi_i)]^{-1}$ اي أن[9]:

$$Z_i \sim N\left((X_i\beta), [n_i\pi_i(1-\pi_i)]^{-1}\right) \quad .....(9)$$

2-4 طرائق تقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلة التعدد الخطي:

1-2-4 طريقة انحدار المربعات الصغرى الجزئية: Partial Least Square Regression(PLSR)

أن طريقة المربعات الصغرى الجزئية 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 ،

نفرض لدينا المصفوفة $Xn \times p$ والمتجه $\underline{Y}n \times 1$ المربعات الصغرى الجزئية تعتمد على النموذج الثنائي بين X و **$\underline{Y}$** وكالاتي [11] [15]:

$$X = TP' + E \quad ... .(10)$$

$$\underline{Y} = Uq' + f \quad ....(11)$$

اذ

T : مصفوفة درجات- مصفوفة x-score ذات رتبة nxr حيث ان T هي مصفوفة متماثلة symmetric

U : مصفوفة درجات-Y-score ذات رتبة nxr

P' : مصفوفة تحميلات-x-loading ذات رتبة pxr

q': متجه تحميلات-Y-loading ببعد 1xr

E: مصفوفة البواقي-x-residual ذات رتبة nxp

F : متجه البواقي-Y-residual بعد nx1

المصفوفة P' والمتجه q' له r من الأعمدة وهو محدد بما يأتي:

$(r < \min\ (n, p))$

اذ أن:

:P عدد المتغيرات

:n عدد المشاهدات

:r عدد المركبات

والعلاقة الداخلية التي تربط بين scores تعطى وكالاتي

.....(12)

$U = TD + H$

حيث D مصفوفة قطرية ذات رتبة rxr

H مصفوفة البواقي ذات رتبة nxr

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

Max COV(Xw,Yc)

$with\|t\|=\|Xw\|=1 \quad and\|u\| = \|Yc\| = 1 \quad ....(13)$

اذ 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 (13) منها خوارزمية Kernel وخوارزمية SIMPLS وخوارزمية NIPALS وغيرها من الخوارزميات وفي هذا البحث تم الاعتماد على خوارزمية التكرار غير الخطي للمربعات الصغرى الجزئية Non-linear Iterative partial least squares NIPALS(PLS1)

4-2-1-1 خوارزمية NIPALS(PLS1)

فيما يأتي الخطوات الاساسية لخوارزمية NIPALS(PLS1) لحساب اول مركبة[2] [4]

1- في الخطوة الاولى يتم تهيئة U_1 عن طريق **Y** بحيث

U_1= **Y**......(14)

اذ ان U_1 متجه ببعد nx1

2- حساب X-weight

$\underline{W1} = X'U_1/(U'_1U_1)$ .....(15)

اذ ان $\underline{W_1}$ متجه ببعد px1

اذ أن $\underline{W_1}$ يكون normalize بالشكل $\underline{W1}$=**W1**/||W1||

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

t_1=XW1.....(16)

اذ ان t1 متجه ببعد nx1

4-حساب y-weight

$C_1=\underline{\acute{Y}}t_1/(t'_1t_1)$.....(17)

اذ c1 متجه ببعد 1x1

اذ أن c1 تكون normalize بالشكل $C_1 = C_1/||C_1||$

5- اسقاطات بيانات Y على Y-weight لحساب y-scores

.....(18) U^*_1= Y C_1

حيث U^*_1 متجه ببعد nx1

6- نحدد U^*_1 بحيث تحقق ما يأتي

$$\Delta U = (U\Delta)'(U\Delta) \quad ....(19)$$

$U\Delta = U^*_1 - U_1$

اذا كانت $\Delta U < \varepsilon$ وجدنا اول مركبة ونتوقف حيث ε قيمه صغيره عدا ذلك نذهب الى الخطوة الاولى ونستعمل U^*_1 بدل U_1 ونستمر بالخطوات.

7-ايجاد X-loading

$$P_1 = X't_1/(t'_1t_1) \quad ....(20)$$

اذ ان P_1 متجه ببعد px1

8- ايجاد Y-loading

$$q = \underline{Y}'U_1/(U'_1U_1) \quad ....(21)$$

حيث ان q متجه ببعد 1x1

9- ايجاد التداخل الخطي للمعالم بواسطة انحدار OLS

$$d_1 = U'_1t_1/(t'_1t_1) \quad .....(22)$$

اذ ان d1 متجه ببعد 1x1

10- عمل تفريغ deflate الى بيانات X وبيانات Y

$$X_1 = X - t_1P'_1 \quad ....(23)$$

$$Y_1 = Y - d_1t_1C'_1 \quad ....(24)$$

ونستمر بالخطوات من (1-10) عدة مرات وبأستعمال البيانات المفرغة الى X وY حتى نحصل على كل المركبات المحددة ونستطيع ان نجد معاملات الانحدار بواسطة العلاقة الاتيه

11- ايجاد معاملات الانحدار

$$\beta = w(p'w)^{-1}c' \quad .....(25)$$

اذ ان W هي مصفوفة برتبة pxr

P مصفوفة برتبة pxr

C مصفوفة برتبة rxr

2-2-4 خوارزمية تجزئة القيم المفردة: Singular Value Decomposition (SVD)

تقوم هذه الخوارزمية بأختصار عدد الابعاد الى اقل مايمكن بحيث تكون الابعاد المتبقية لها اكبر تباين (بمعنى اخر ضغط البيانات عن طريق ازالة المتكرر)، عادة تستخدم كعملية سابقة لعمليات تحليل وتنقيب البيانات ، يمكن النظر الى تجزئة القيمة المفردة (SVD) من ثلاث نقاط متوافقة مع بعضها البعض، فمن جهه يمكننا أن نرى ذلك كوسيلة لتحويل المتغيرات المترابطة الى مجموعة من تلك غير المترابطة التي تعرض أفضل العلاقات المختلفة بين عناصر البيانات الاصلية، في نفس الوقت SVD هو طريقة لتحديد وترتيب الابعاد لنقاط البيانات التي تظهر اكبر التباين، اذ انه بمجرد تحديد اكبر تباين فأنه من الممكن العثور على افضل تقريب من نقاط البيانات الاصلية بأستخدام ابعاد اقل، ومن ثم يمكن عدها SVD كوسيلة لخفض البيانات، تستخدم خوارزمية تجزئة القيم المفردة (SVD) للتخلص من مشكلة التعدد الخطي التام بين المتغيرات التوضيحية.[8]

تستخدم خوارزمية تجزئة القيم المفردة (SVD) لأيجاد المركبات الرئيسة اذ انها تعطي المتجه المميز والقيم المميزه التي تحتاج في تحليل المركبات الرئيسية ، اذ انه في تحليل المركبات الرئيسة (PCA) نحصل على اول مركبة رئيسة (PC) بأستعمال SVD من خلال تجزئة المصفوفة X ذات الرتبة $n \times p$ الى ثلاث مصفوفات وكالاتي [8] [14] :

$$X = T_0\ S\ P' \qquad ......(26)$$

نفرض أن

$$t = min\{n, p\}$$

اذ أن:

T_0 : مصفوفة متعامدة ذات رتبة $n \times t$ وهي مصفوفة المركبات ويتم إيجادها من المتجه المميز لـ XX'

S : مصفوفة قطرية ذات رتبة $t \times t$ وهي تساوي الجذور المربعة الى القيم المميزة لـ $X'X$ أو XX' اذ أن $S = diag\{\theta_1 > \theta_2 \dots \theta_p > 0\}$

P : مصفوفة متعامدة ذات رتبة $p \times t$ وهي مصفوفة تحميل ويتم ايجادها من المتجه المميز الى $X'X$

اذ أن المركبات الرئيسة في التحليل هي T يمكن أن تكون

$$T = T_0\ S \qquad ....(27)$$

$$X = T_0\ S\ P + \varepsilon \qquad ....(28)$$

أما معاملات الانحدار هي

$$\beta = P(T'T)^{-1}\ T'y \qquad .... (29)$$

$$= P\ S^{-1}T'_0$$

5- الجانب التجريبي:

1-5 مراحل تطبيق تجربة المحاكاة

Stages of the application of simulation experiment

لقد تضمنت تجارب المحاكاة كتابة عدد من البرامج بلغة (MATLAB 2017)، اذ أن الأنموذج الذي تم الاعتماد عليه يكون وفق الصيغة الاتية:

$$y_i = \pi_i + \varepsilon_i \qquad ....(30)$$

اذ يتم وصف مراحل تجربة المحاكاة من خلال الخطوات الاتية:

1- تعيين القيم الافتراضية للمعلمات وهذه المرحلة من اهم المراحل التي يعتمد عليها، حيث تم تحديد القيم الافتراضية للمعلمات كقيم اوليه من دراسات سابقه، اذ أختيرت قيم المعلمات والنموذج المفترض كما مبين في ادناه:

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
0.33	0.41	-0.95	-0.17	-0.29	1.38	-0.61	-0.94

2- توليد المتغيرات التوضيحية من خلال استعمال اسلوب مونت-كارلو(Mont-Carlo) في المحاكاة حيث يتم توليد سبعة متغيرات توضيحية وفق التوزيع الطبيعي القياسي وحدوث مشكلة التعدد الخطي بحسب الصيغة الاتية [6]

$$X_{ij} = (1 - P^2)^{\frac{1}{2}} u_{ij} + P u_{i(p+1)} \quad , \quad i=1,2,\dots,n \ , \ j=1,2,\dots,p \qquad ....(31)$$

u_{ij} : الاعداد العشوائية المولدة والتي تتبع التوزيع الطبيعي القياسي.

$u_{i(p+1)}$: يمثل قيم العمود الأخير من أعمدة المتغيرات المولدة.

j : يمثل عدد المتغيرات المرتبطة ، مع العلم انه $j' < P$.

i : يمثل عدد المشاهدات.

P : يمثل قيمة الارتباط بين المتغيرات التوضيحية .

3- توليد قيم متغير الخطأ العشوائي في أنموذج الانحدار اللوجستي تبعا لتوزيع برنولي.
4- حساب المتغير التابع (y) ثنائي الاستجابة الذي يتوزع توزيع برنولي، وفق طريقة توليد التحويل المعكوس(invers transformation) بالاعتماد على دالة الانحدار اللوجستي (π_i) وحد الخطأ العشوائي
5- من أهم العوامل الأخرى التي يتم اختيارها والمؤثرة هي كالاتي:

- **اختيار اربعة أحجام للعينات المفترضة وهي(200,100,50,25)**
- **اختيار القيم الافتراضية لمعاملات الارتباط وهي(0.99,0.90,0.80)**

6- تقدير معلمات أنموذج الانحدار اللوجستي وفق طرائق التقدير التي تم عرضها في الجانب النظري وهي كالاتي:

1- طريقة انحدار المربعات الصغرى الجزئية (PLSR).
2- خوارزمية تجزئة القيم المفردة (SVD).

7- أما في هذه المرحلة، تتم المقارنة بين طرائق التقدير المدروسة بالاعتماد على المقياس الاحصائي متوسط مربعات الخطأ (MSE) للأنموذج.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - \pi_i)^2 \ldots\ldots(32)$$

واخيرا سوف يتم تكرار تجربة المحاكاة (1000) مره.

2-5 تحليل نتائج المحاكاة

جدول رقم (1) يبين تقديرات المعلمات و قيم متوسط مربعات الخطأ لأنموذج الانحدار اللوجستي في حالة $n = 25$ وبارتباطات مختلفة

coefficients		ethods	
		PLS-LR	SVD-LR
$\rho = 0.80$	$\hat{\beta}_0$	0.6223	0.6037
	$\hat{\beta}_1$	-1.5584	-0.1316
	$\hat{\beta}_2$	-2.9660	-0.3779
	$\hat{\beta}_3$	1.6159	0.4421
	$\hat{\beta}_4$	-0.0956	-0.4892
	$\hat{\beta}_5$	1.4277	0.4850
	$\hat{\beta}_6$	0.5762	-0.0021
	$\hat{\beta}_7$	-1.7556	-0.3079
MSE		0.0430	0.2272
$\rho = 0.90$	$\hat{\beta}_0$	0.9379	0.6027
	$\hat{\beta}_1$	-1.2520	-0.1542
	$\hat{\beta}_2$	-4.0195	-0.4661
	$\hat{\beta}_3$	3.8164	0.6021
	$\hat{\beta}_4$	-2.1037	-0.6292
	$\hat{\beta}_5$	4.4086	0.6707
	$\hat{\beta}_6$	1.4833	0.0011
	$\hat{\beta}_7$	-1.6693	-0.3835

MSE		**0.5315**	**0.2270**
$\rho\ = 0.99$	$\hat{\beta}_0$	**1.9452**	**0.5764**
	$\hat{\beta}_1$	**1.2372**	**-0.7521**
	$\hat{\beta}_2$	**1.6299**	**0.3048**
	$\hat{\beta}_3$	**0.2982**	**1.0379**
	$\hat{\beta}_4$	**-2.5430**	**-2.0744**
	$\hat{\beta}_5$	**1.1551**	**1.4358**
	$\hat{\beta}_6$	**-1.3145**	**-0.1619**
	$\hat{\beta}_7$	**-0.7477**	**-0.2738**
MSE		**2.6251**	**0.1292**

جدول رقم (2) يبين تقديرات المعلمات و قيم متوسط مربعات الخطأ لأنموذج الانحدار اللوجستي في حالة $n = 50$ وبارتباطات مختلفة.

coefficients		**Methods**	
		PLS-LR	**SVD-LR**
$\rho\ = 0.80$	$\hat{\beta}_0$	**0.8437**	**0.6347**
	$\hat{\beta}_1$	**0.1583**	**0.1286**
	$\hat{\beta}_2$	**-0.6667**	**-0.0492**
	$\hat{\beta}_3$	**-2.0207**	**-0.1581**
	$\hat{\beta}_4$	**-0.7534**	**-0.1441**
	$\hat{\beta}_5$	**3.8577**	**0.4544**
	$\hat{\beta}_6$	**0.6635**	**-0.0710**
	$\hat{\beta}_7$	**-1.4759**	**-0.4231**
MSE		**0.3158**	**0.3256**
$\rho\ = 0.90$	$\hat{\beta}_0$	**1.1388**	**0.6656**
	$\hat{\beta}_1$	**2.9498**	**0.2217**
	$\hat{\beta}_2$	**0.5989**	**0.0182**
	$\hat{\beta}_3$	**-0.6850**	**-0.0349**
	$\hat{\beta}_4$	**-1.4520**	**-0.3393**
	$\hat{\beta}_5$	**1.1056**	**0.4066**
	$\hat{\beta}_6$	**0.4241**	**-0.0501**
	$\hat{\beta}_7$	**-0.8912**	**-0.4399**

MSE		0.9166	0.3797
$\rho = 0.99$	$\hat{\beta}_0$	1.6533	0.6643
	$\hat{\beta}_1$	2.3920	0.7841
	$\hat{\beta}_2$	0.3864	0.4513
	$\hat{\beta}_3$	1.1341	-0.6195
	$\hat{\beta}_4$	-1.0094	-0.8437
	$\hat{\beta}_5$	-1.4502	0.1186
	$\hat{\beta}_6$	-0.2631	-0.0264
	$\hat{\beta}_7$	-0.0776	-0.1912
MSE		3.1336	0.3723

جدول رقم (3) يبين تقديرات المعلمات و قيم متوسط مربعات الخطأ لأنموذج الانحدار اللوجستي في حالة $n = 100$ وبارتباطات مختلفة.

coefficients		Methods	
		PLS-LR	SVD-LR
$\rho = 0.80$	$\hat{\beta}_0$	1.1774	0.5379
	$\hat{\beta}_1$	0.6195	0.1808
	$\hat{\beta}_2$	1.1958	-0.1083
	$\hat{\beta}_3$	0.5448	0.0466
	$\hat{\beta}_4$	-1.5659	-0.2114
	$\hat{\beta}_5$	2.9040	0.2641
	$\hat{\beta}_6$	-1.0970	-0.2212
	$\hat{\beta}_7$	-1.6154	-0.2053
MSE		2.2721	0.6820
$\rho = 0.90$	$\hat{\beta}_0$	1.5461	0.5557
	$\hat{\beta}_1$	-0.0853	0.0781
	$\hat{\beta}_2$	1.8931	-0.0843
	$\hat{\beta}_3$	-0.2563	0.0397
	$\hat{\beta}_4$	-0.7778	-0.2314
	$\hat{\beta}_5$	1.8156	0.2201
	$\hat{\beta}_6$	-0.0810	-0.1365
	$\hat{\beta}_7$	-1.1259	-0.1795
MSE		5.7854	0.6695
$\rho = 0.99$	$\hat{\beta}_0$	1.5222	0.6193
	$\hat{\beta}_1$	-0.6439	-0.1130
	$\hat{\beta}_2$	2.6572	0.1032
	$\hat{\beta}_3$	0.3443	0.1135

	$\hat{\beta}_4$	-0.3744	-0.3512
	$\hat{\beta}_5$	0.3990	0.1550
	$\hat{\beta}_6$	-0.3114	-0.1460
	$\hat{\beta}_7$	-0.3966	0.0097
MSE		5.3224	0.8018

جدول رقم (4) يبين تقديرات المعلمات و قيم متوسط مربعات الخطأ لأنموذج الانحدار اللوجستي في حالة $n = 200$ وبارتباطات مختلفة

coefficients		Methods	
		PLS-LR	SVD-LR
$\rho = 0.80$	$\hat{\beta}_0$	1.1008	0.5934
	$\hat{\beta}_1$	1.1164	0.0686
	$\hat{\beta}_2$	-2.0406	-0.2736
	$\hat{\beta}_3$	-0.4392	0.0273
	$\hat{\beta}_4$	-0.8210	0.0195
	$\hat{\beta}_5$	3.0267	0.3835
	$\hat{\beta}_6$	0.0160	-0.2345
	$\hat{\beta}_7$	0.2301	-0.1521
MSE		2.8633	1.1444
$\rho = 0.90$	$\hat{\beta}_0$	1.3609	0.5957
	$\hat{\beta}_1$	1.0980	0.0443
	$\hat{\beta}_2$	-1.5818	-0.4175
	$\hat{\beta}_3$	-0.4076	0.0305
	$\hat{\beta}_4$	-0.9296	0.0432
	$\hat{\beta}_5$	1.8816	0.4421
	$\hat{\beta}_6$	0.8017	-0.2134
	$\hat{\beta}_7$	0.7385	-0.1546
MSE		6.5187	1.0897
$\rho = 0.99$	$\hat{\beta}_0$	1.6389	0.5875
	$\hat{\beta}_1$	1.4889	-0.1597
	$\hat{\beta}_2$	-0.9998	-0.8841
	$\hat{\beta}_3$	-0.5393	0.2081
	$\hat{\beta}_4$	-1.3706	-0.2092
	$\hat{\beta}_5$	1.2753	0.7104
	$\hat{\beta}_6$	0.7280	-0.1635
	$\hat{\beta}_7$	0.5995	0.2207
MSE		11.9241	0.9535

من خلال النتائج المبينة في الجدول(1) و (2) و (3) و(4) نلاحظ الاتي:
1- عند حجوم العينات (25) و (50) وعندما يكون معامل الارتباط (P=0.80) نلاحظ أن طريقة المربعات الصغرى الجزئية (PLS) تمتلك اقل (MSE) بينما عندما يكون معامل الارتباط (P=0.90 , 0.99) نلاحظ أن خوارزمية تجزئة القيم المفردة (SVD) تمتلك اقل (MSE).
2- عند حجوم العينات (100) و (200) ولجميع قيم معاملات الارتباط (P=0.80 , 0.90 , 0.99) نلاحظ بأن خوارزمية تجزئة القيم المفردة (SVD) افضل من طريقة المربعات الصغرى الجزئية (PLS) في معالجة مشكلة التعدد الخطي لأنموذج الانحدار اللوجستي وذلك لانها تمتلك اقل (MSE).

6- الاستنتاجات والتوصيات

1-6 الاستنتاجات

1- أثبتت خوارزمية تجزئة القيم المفردة (SVD) كفاءتها في تقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلة التعدد الخطي بالنسبة لقيم معاملات الارتباط (0.80 , 0.90 , 0.99) ولحجوم العينات الكبيرة وذلك لانها حققت اقل (MSE) للأنموذج وكذلك أثبتت خوارزمية (SVD) بأنها الافضل في حالة حجوم العينات المتوسطة والصغيرة ولقيم معاملات الارتباط (0.90 , 0.99) وذلك لانها حققت اقل (MSE) للأنموذج .
2- أثبتت طريقة المربعات الصغرى الجزئية (PLS) بأنها الافضل في حالة حجوم العينات المتوسطة والصغيرة ولقيمة معامل الارتباط (0.80) لانها حققت اقل (MSE) للأنموذج.

2-6 التوصيات

1- استعمال خوارزمية تجزئة القيم المفردة (SVD) في تقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلة التعدد الخطي وبأختلاف احجام العينات لما تبديه من كفاءة ومرونة في التطبيق.
2- استعمال طريقة المربعات الصغرى الجزئية (PLS) في تقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلة التعدد الخطي في حالة حجوم العينات المتوسطة والصغيرة.
3- استخدام خوارزميات اخرى لطريقة المربعات الصغرى الجزئية (PLS) غير خوارزمية (NIPALS(PLS1)) المستعملة في البحث مثل خوارزمية (SIMPLS) وخوارزمية (Kernel) في تقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلة التعدد الخطي.

7-المصادر

1- البياتي ، محمود مهدي حسن . (2012) ، " تطبيق عملي لتحليل البيانات الاحصائية " ، الجزيرة للطبع والنشر/جامعة بغداد.
2- صالح ، رباب عبد الرضا . (2016) ، " مقارنة بين طرائق المربعات الصغرى الجزئية والمركبات الرئيسية بأستعمال المحاكاة " ، مجلة العلوم الاقتصادية والادارية ، المجلد 22 ، العدد 87 .
3- عباس ، علي خضير . (2012) ، " استخدام نموذج الانحدار اللوجستي في التنبؤ بالدوال ذات المتغيرات الاقتصادية التابعة النوعية " ، مجلة جامعة كركوك للعلوم الادارية والاقتصادية ، المجلد 2 ، العدد 2.
4- Abdi , Hervi , (2010). " Partial Least Squares Regression and Projection on Latent Structure Regression (PLS Regression) " , John Wiley & Sons, 1-10
5- Aguilera , A.M., and Escabias , M and M.J. Valderrama , (2006). " Using principal components for estimating Logistic regression with high dimensional Multicollinearity data ", Computational statistics Data Analysis 50 , 1905-1924.
6- Al-Hassan , Yazid m., (2008). " A Monte Carlo Evaluation of Some Ridge Estimators " , J.J. Appl. Sci: Natural Sciences Series 10(2): 101-110.
7- Anne Laure Boulesteix , Korbinian Strimmer , (2005). "Partial Least Squares: a Versatile Tool for the analysis of high-dimensional genomic data " , Briefings in Bioinformatics , oxford University , P.32-44.
8- Baker , K., (2005) . " Singular Value Decomposition Tutorial ".

9- Berkson , J.,(1944). "Application of the Logistic Function to Bioassay",JASA Vol.39,pp.357-365.

10- Cook , D., Dixon , P., Duckworth , W.M., Kaiser , M.S.,Koehler, K., Meeker , W.Q and Stephenson , W.R., (2001). " Binary Response and Logistic Regression Analysis ", University NSF/ILI Project Beyond Traditional statistical Methods , grant.

11- Chong , I., Jun T,C., (2005). "Performance of some variable selection methods when multicollinearity Present ", Chemometrics and Intelligent Laboratory systems 78 , 103-112.

12- Francesca , Fallucchi and Fabio , Massimo Zanzotto , (2009) , " Singular Value Decomposition for Feature Selection in Taxonomy Learning " , Via del Politecnico 00133 Rome , Italy.

13- LI Shen , Eng , Chong , Tan , (2005). " PLS and SVD Based Penalized Logistic Regression for cancer Classification Using Microarray Data " , School of Computer Engineering , Nanyang Technological University , Singapore , pp 219- 228.

14- Mevik , B., Wehrens , R., (2007). " The PLS Package: Principal component and partial Least squares Regression in R ".

15- Roon , P., Zakizadeh , J., Chartier , S., (2014). " Partial Least Squares tutorial analyzing neuroimaging data ".

16- Tormod Naes and Bjorn-Helge Mevik , (2001). " Understanding The Collinearity Problem in Regression and Discriminant Analysis " , Journal of Chemo Metrics , p.413-426.

Comparison of the method of partial least squares and the algorithm of singular values decomposion to estimate the parameters of the logistic regression model in the case of the problem of linear multiplicity by using the simulation

Abstract

The logistic regression model is an important statistical model showing the relationship between the binary variable and the explanatory variables. The large number of explanations that are usually used to illustrate the response led to the emergence of the problem of linear multiplicity between the explanatory variables that make estimating the parameters of the model not accurate.

The methods used to estimate the parameters of the logistic regression model in the case of the linear multiplication problem.

These methods are the method of regression of the partial least squares and the algorithm of singular value decomposion.

The simulation method was used to compare estimation methods through the mean error squares of the model.

It has been shown through the comparison that the algorithm of singular value decomposion is best in estimating the parameters of the logistic regression model in the case of the problem of linear multiplicity.

Keywords **Logistic regression, binary data, partial least square, algorithm singular value decomposition, multicollinearity.**