التحليل المميز والانحدار اللوجستي بأستعمال المربعات الصغرى الجزئية (دراسة تجريبية (محاكاة))

أ.م.د. رباب عبد الرضا صالح البكري/جامعة بغداد/ كلية الإدارة والاقتصاد
الباحث / محمد شاكر محمود العزي/ جامعة بغداد/ كلية الإدارة والاقتصاد

تاريخ التقديم:2018/2/2
تاريخ القبول:2018/7/24

المستخلص

يعد اسلوبي الانحدار اللوجستي الثنائي regression logistic Binary والدالة المميزة الخطية Linear discriminant function من اهم الاساليب الاحصائية المستخدمة في التصنيف والتنبؤ، عندما تكون البيانات من النوع الثنائي (0،1) فانه لا يمكن استخدام الانحدار الاعتيادي فلذلك نلجأ الى الانحدار اللوجستي الثنائي والدالة المميزة الخطية في حالة وجود مجموعتين او اكثر، وفي حالة وجود مشكلة التعدد الخطي Multicollinearity بين البيانات (ان البيانات يوجد فيها ارتباطات عالية بين المتغيرات) اصبح عدم الامكان في استخدام الانحدار اللوجستي والدالة المميزة الخطية، ولحل هذه المشكلة توجد عدة طرائق منها طريقة انحدار المربعات الصغرى الجزئية Partial least square regression لحل مشكلة التعدد الخطي.

وقد جرى في هذه البحث المقارنة بين الانحدار اللوجستي الثنائيregression logistic binary والدالة المميزة الخطية linear discriminant function عن طريق خطأ التصنيف. حيث تم توليد بيانات بمتغير استجابة (Y) نوع ثنائي (0,1) تحتوي على مشكلة التعدد الخطي وبحجوم عينات (50-100-150-250-400) ومتغيرات (5-10-20). حيث تمت معالجة مشكلة التعدد الخطي بأستعمال طريقة المربعات الصغرى الجزئية Partial least square.

وتوصل البحث الى ان الدالة المميزة الخطية linear discriminant function هي أفضل في تصنيف البيانات من الانحدار اللوجستي الثنائيbinary logistic regression ، اذ صنفت الدالة المميزة البيانات بشكل صحيح وأكثر دقة من الانحدار اللوجستي الثنائي.

المصطلحات الرئيسة للبحث: الدالة المميزة الخطية – الانحدار اللوجستي الثنائي – المربعات الصغرى الجزئية – مشكلة التعدد الخطي – نسبة التصنيف – محاكاة.

مجلة العلوم الاقتصادية والإدارية
العدد 106 المجلد 24
الصفحات 407-419

*البحث مستل من رسالة ماجستير

1- المقدمة Introduction

ان اسلوب الانحدار اللوجستي من الاساليب المهمة التي تدخل في تحليل البيانات التي تكون فيها بيانات المتغير المعتمد (Y) يكون فيه البيانات ثنائية (0،1)، ويكون الهدف الاساس من هذه الطريقة هو ايجاد أفضل نموذج يصف الحالة بين المتغير المعتمد والمتغير التوضيحي (المتغيرات التوضيحية). وان طريقة التحليل المميز هي من التحليلات الاحصائية التي تهتم بمتعدد المتغيرات حيث يتم بهذا التحليل استعمال مجموعة من المتغيرات للتميز بين مجموعتين او أكثر بواسطة عدة دوال تميزية. سيتم اعتماد هذين الاسلوبين للعمل في هذا البحث، وللتخلص من مشكلة التعدد الخطي سيتم استعمال طريقة المربعات الصغرى الجزئية.

2- مشكلة البحث problem of research

لغرض استعمال بيانات من النوع ثنائي الاستجابة والتي تعاني من وجود مشكلة التعدد الخطي من اجل تكوين أنموذج احتمالي والذي يتم على اساسه تصنيف البيانات، حيث ثم استعمال اسلوبين من اساليب التصنيف لغرض ايجاد أفضل انموذج احتمالي بأقل خطأ تصنيف ممكن.

3- هدف البحث objective of research

يتضمن هذا البحث المقارنة بين اسلوبين من اساليب التصنيف وهما الانحدار اللوجستي الثنائي والدالة المميزة الخطية بوجود مشكلة التعدد الخطي ولمعرفة مدى قابليتهم لتصنيف البيانات بأقل احتمال خطأ للتصنيف. وذلك باستعمال تجارب محاكاة عندما يتوزع الخطأ توزيع طبيعي لحجوم عينات وابعاد مختلفة.

4- الجانب النظري

من الأساليب الإحصائية المهمة في متعدد المتغيرات والتي تستعمل في تحليل وتقويم العلاقات بين مجموعة من المتغيرات يكون فيه المتغير التابع (متقطع)، ليس دائما يكون فيها المتغير التابع مستمرا وذلك لغرض الحصول على انموذج رياضي يوضح العلاقة بين مثل هكذا بيانات تستعمل اسلوبي التحليل المميز والانحدار اللوجستي. وان هناك بعض الافتراضات الخاصة للأسلوبين أعلاه منها وان تكون المتغيرات التوضيحية مستقلة ولا يوجد أي ارتباط بينها ولمعالجة مثل هكذا حالات تستعمل عدة طرائق منها طريقة المركبات الرئيسية وطريقة المربعات الصغرى الجزئية وسيتم في هذا البحث التركيز على طريقة المربعات الصغرى الجزئية.

1-4 المربعات الصغرى الجزئية[1] [5] [6] Partial least square

تعد طريقة المربعات الصغرى الجزئية أكثر الطرائق أهمية في الانحدار فهي تستعمل لتقليص عدد المتغيرات التوضيحية المرتبطة في الانموذج الى مركبات غير مرتبطة (خطية، متعامدة)، او عندما يكون عدد المتغيرات التوضيحية أكثر من عدد المشاهدات في التجربة. وطريقة المربعات الصغرى الجزئية مشابهة لطريقة المركبات الرئيسية وأسلوب انحدار الحرف لمعالجة مشكلة التعدد الخطي ولكنها تختلف في الحسابات فخوارزمية (Partial least square) تأخذ بنظر الاعتبار التباين المشترك ما بين متغير (متغيرات) الاستجابة والمتغيرات التوضيحية، اما طريقة المركبات الرئيسية (PCA) تأخذ بنظر الاعتبار التباين بين المتغيرات التوضيحية فقط. يقوم بتحويل المتغيرات التوضيحية المرتبطة الى مركبات رئيسية تختلف في الحسابات فخوارزمية الحل بطريقة PLS طريقة تكرارية عند استعمالها تنتج سلسلة من النماذج ويتوقف الحل التكراري عندما نصل الى العدد الكلي من المركبات في الانموذج او عندما تكون البواقي مساوية للصفر.

ففي حال تساوي عدد المركبات مع عدد المتغيرات التوضيحية فان النتائج ستكون متطابقة مع طريقة المربعات الصغرى. ولأجل تحديد عدد المركبات التي تصغر من خطأ التنبؤ نستعمل طريقة العبور الشرعي Cross-validation وبعد تحديد عدد المركبات تقدر معالم انموذج الانحدار لكل متغير.

وان طريقة PLS اول من طبقها الباحث Wold عندما يكون هناك ارتباط عالي ما بين المتغيرات التوضيحية او عندما يكون عدد المتغيرات التوضيحية تفوق عدد المشاهدات. وتوجد عدة خوارزميات وأكثرها تداولا خوارزمية NIPALS عام 1973 للباحث Wold وخوارزمية SIMPLS 1993 المنسوبة للعالم De Jong وخوارزمية KERNEL المنسوبة للعالم DAYAL وخوارزمية orthogonal projection to latent structures عام 2002 المنسوبة الى TRYGG، خوارزمية NIPALS او تسمى (PLS1,PLS2) حيث ان خوارزمية PLS1 تستعمل عندما يكون متغير الاستجابة متجه اما خوارزمية PLS2 تستعمل عندما يكون متغير الاستجابة مصفوفة، وخوارزمية PLS1 تعطي نفس نتائج خوارزمية SIMPLS. وسنركز في هذا البحث على خوارزمية SIMPLS.

2-4 **خوارزمية** SIMPLS Algorithm SIMPLS[1][5][6]

1- إيجاد مصفوفة التباين والتباين المشترك للبيانات

$S_{xy} = \acute{X}Y$ (X and Y are enteral) .....(1)

$\tilde{X}_i = X_i - \bar{X}$.....(2)

$\tilde{Y}_i = Y_i - \bar{Y}$.....(3)

2- إعادة تكرار الخطوات من (2-1) الى (2-6) لكل من h=1,2,3,.....,k

2-1 إيجاد اول متجه مميز للمصفوفة $S_{xy}{}^{*}S_{xy}$ ول h^{th} من متجهات الاوزان (r_h) لطريقة المربعات الصغرى الجزئية (X-weight) مصفوفة الاوزان.

2-2 إيجاد h^{th} من المركبات الخطية (*X-score*) مصفوفة الدرجات

$t_h = \tilde{X}r_h$ and normalized .....(4)

$t_h = t_h / \|t_h\|$.....(5)

2-3 حساب h^{th} تحميلات X – loading بانحدار X على t_h

$\hat{P}_h = \acute{X}t_h$.....(6)

2-4 خزن المتجهات $t_h r_h$, p_h , في المصفوفات

$R_h=\{r_1,.....r_h\}$.....(7)
$T_h=\{t_1,.....t_h\}$.....(8)
$P_h=\{P_1,.....P_h\}$.....(9)

وبالتتابع

2-5 اذا كان *h= h+1* فيتم حساب المعادلة:

$S_{xy}{}^{h} = (I_p - V_{h-1} * \grave{V}_{h-1})\, \grave{S}_{xy}{}^{h-1}$.....(10)

وان V_{h-1} هو متعامد طبيعي ***orthonormal*** .

$S_{xy}{}^{h} = S_{xy}{}^{h-1} - P_{h-1}(\grave{P}_{h-1}, P_{h-1})^{-1}\grave{P}_{h-1}S_{xy}{}^{h-1}$.....(11)

وان $(V_1,.....V_{h-1})$ متعامد (orthogonal) الى مصفوفة التحميل X $(P_1,.....P_{h-1})$.

3- يتم الحصول على مقدرات المعلمات لإنموذج الانحدار الخطي

$\hat{B} = R_{ph}(\grave{R}_{h*p} * S_x * R_{p*1})\grave{R}_{h,p}S_{xy}$.....(12)

$\widehat{B_0} = \bar{Y} - \widehat{B_{i,p}}\bar{X}$.....(13)

حيث ان:

R: مصفوفة اوزان – X (X-weights) بأبعاد p*h حيث ان p عدد المتغيرات التوضيحية وان h عدد المركبات.

pT_h هي مصفوفة الدرجات x-score بأبعاد n*h

P_h هي مصفوفة التحميل x-loading بأبعاد p*h

S_{xy}: مصفوفة التباين والتباين المشترك الى XوY

S_x: مصفوفة التباين الى X.

5-الانحدار اللوجستي logistic regression [2][4][8]

يعد انموذج الانحدار اللوجستي من النماذج الإحصائية المهمة في تحليل البيانات اذ ان الهدف الأساسي من معظم الدراسات هو تحليل وتقويم العلاقات بين مجموعة من المتغيرات للحصول على صيغة نستطيع من خلالها ان نصف النموذج ويستعمل انموذج الانحدار اللوجستي لوصف العلاقة بين بين متغير الاستجابة من النوع المتقطع والمتغيرات التوضيحية ويكون على نوعين انموذج الانحدار اللوجستي ثنائي الاستجابة وانموذج الانحدار اللوجستي متعدد الاستجابة وسنركز على النوع الأول من الانحدار اللوجستي.

5-1 انموذج الانحدار اللوجستي ثنائي الاستجابة [2] [4] [8]

ان من خصائص الانحدار اللوجستي ثنائي الاستجابة ان المتغير التابع (y) متغير الاستجابة يتبع توزيع برنولي ويأخذ القيم (0) و (1) اي بمعنى ان (1) باحتمال قدره (p) احتمال نجاح، وباحتمال فشل (1-p) قدره (0) وبمعنى اخر احتمال حدوث الاستجابة (1) واحتمال عدم حدوث الاستجابة (0). ان نموذج الانحدار الخطي المتعدد MLR يكون كالاتي:

$$Y = XB + \in \quad \text{.....(14)}$$

حيث ان:

Y : موجه المتغير المعتمد بأبعاد (n*1)

X : مصفوفة المتغيرات التوضيحية ((n*(k+1))

B: موجه معلمات دالة الانحدار ((k+1)*1)

$\in$: موجه الأخطاء العشوائية والذي يشترط به تحقق الشروط الاتية:

1- $\in_i$ تتوزع توزيع طبيعي

2- $E(\in_i)=0$

3- $Cov(\in_i, \in_j) = 0$

4- $Var(\in_i) = \sigma^2$

n: حجم العينة.

k: عدد المعالم

ففي حالة وجود متغير توضيحي واحد فان متوسط قيم المشاهدة Y عند متجه معين للمتغير x فهو $E(y \mid x)$ وبذلك يمكن كتابة النموذج على النحو التالي:

$$E(y \mid x) = \beta_0 + \beta_1 x \quad \text{.....(15)}$$

ان الطرف الايمن لهذا النموذج يأخذ قيم (+∞،-∞)، لكن عندما يكون المتغير (y) ثنائيا فان الانموذج اعلاه لا يكون ملائما لأن:

$$E(y/x) = P_r(y = 1) = p \quad \text{.....(16)}$$

وفي هذه الحالة يكون الطرف الايمن محصور بين (0،1)، وهذا يعني ان الانموذج يكون غير قابل للتطبيق احصائيا. وللتخلص من هذه المشكلة سنقوم بإدخال تحويلة رياضيا على المتغير التابع y. وبما ان ($0 \leq p \leq 1$) وان ($\frac{p}{1-p}$) هو مقدار موجب محصور بين (0 ،∞) أي ($0 \leq \frac{p}{1-p} \leq \infty$) وبأخذ اللوغاريتم الطبيعي للاساس (e) للمقدار ($\frac{p}{1-p}$) فان مجال القيمة تصبح محصور بين (-∞ ،+∞) وتكون كالاتي ($-\infty \leq log e(\frac{p}{1-p}) \leq \infty$)، وبالنهاية يمكن كتابة انموذج الانحدار في حالة وجود متغير توضيحي واحد وكالاتي:

$$log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \text{.....(17)}$$

اما اذا كان لدينا اكثر من متغير توضيحي فتصبح صيغته كالاتي:

$$log_e\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \text{.....(18)}$$

اذ ان:

i= 1,2,3,.....,n j= 1,2,3,.....,k

$\beta_1, \beta_2, \dots\dots\dots, \beta_k$: موجه للمعالم المطلوب تقديرها.

x_{ij}: متغيرات توضيحية.

اما بالنسبة ($\frac{p}{1-p}$) نسبة افضلية النجاح (odds of success) او نسبة الافضلية للحدث المرغوب به وصيغته الرياضية هي كالاتي:

$$\frac{P(Y=1)}{1-P(Y=1)} = e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}} \text{.....(19)}$$

وصيغة احتمالات الاستجابة لأنموذج الانحدار اللوجستي تكتب كالاتي:

$$p = \frac{1}{1+(e^{\beta_0 + \beta_1 x_i})^{-1}} \text{.....(20)}$$

وان المقدار ($Log_e \frac{p}{1-p}$) يسمى لوغاريتم نسبة افضلية النجاح (logs odds of success).

2-5 الافتراضات الخاصة بالانحدار اللوجستي

ان الانحدار اللوجستي لا يتطلب افتراضات كثيرة فقط يتطلب عدم وجود ارتباط بين المتغيرات التوضيحية وان حجم المشاهدات كبيرة في كل مجموعة يفترض انها تكون اكبر من خمس مرات من عدد المعلمات المستعملة في الانموذج النهائي.

3-5 تقدير احتمالات الاستجابة Estimation of response probabilities (8)(4)(2)

ان صيغة احتمال الاستجابة لأنموذج الانحدار اللوجستي الثنائي يمكن الحصول عليها وذلك بإدخال ال (e) على المعادلة التالية:

$$log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \text{.....(21)}$$

وذلك صيغة احتمال الاستجابة تصبح كالاتي:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_i} \text{.....(22)}$$

ويمكن ايضا ان نكتب المعادلة اعلاه بالصيغة الاتية:

$$\frac{1}{\frac{1}{p}-1} = e^{\beta_0+\beta_1 x_i} \text{.....(23)}$$

وباستعمال الطرق الجبرية للاشتقاق تصبح صيغة احتمال الاستجابة لانموذج الانحدار اللوجستي الثنائي يكون كالتالي:

$$p = \frac{1}{1+(e^{\beta_0+\beta_1 x_i})^{-1}} \text{.....(24)}$$

وهذا يعني ان احتمال متغير الاستجابة (y) تأخذ القيمة (1) ويكون كالاتي:

$$P(y=1 \mid x) = \frac{1}{1+(e^{\beta_0+\beta_1 x_i})^{-1}} \text{.....(25)}$$

وعند القيمة (0) فان احتمال متغير الاستجابة (y) يكون على النحو التالي:

$$P(y=0 \mid x) = \frac{1}{1+e^{\beta_0+\beta_1 x_i}} \text{.....(26)}$$

وبما ان مجموع الاحتمالات يساوي (1) فأن:

$$P(y=1|x) + P(y=0|x) = 1 \text{.....(27)}$$

6-التحليل التمييزي: Discriminant analysis [3] [7] [9]

يعد التحليل التميزي من الأساليب الإحصائية المهمة في متعدد المتغيرات التي تهتم بتفريق (تميز) بين مجتمعين (مجموعتين) او أكثر من خلال إيجاد توافيق خطية للمتغيرات التوضيحية تعرف بدالة التميز التي يمكن عن طريقها الفصل او تميز بين مجموعتين او أكثر. وهناك عدة دوال تميزية منها الدالة المميزة الخطية والتي قدمت من قبل العالم فيشر عام (1936) والتي تستعمل في حالة تساوي مصفوفة التباينات المشتركة للمجاميع كافة وان كل مجموعة تعبر عن مجتمع طبيعي لمتعدد المتغيرات اما في حالة عدم تساوي مصفوفة التباينات المشتركة نلجأ الى الدالة المميزة التربيعية. ان من الافتراضات الخاصة بالتحليل المميز بأن المجاميع المدروسة ذات توزيع طبيعي متعدد متغيرات ، كذلك يتطلب ان تكون حجم العينة كبيرة حيث ان المجاميع المختلفة تتضمن على الأقل 20 مشاهدة لكل متغير توضيحي وأيضا يشترط عدم وجود ارتباط بين المتغيرات التوضيحية وعدم وجود قيم شاذة بينها.

6-1 الدالة المميزة الخطية لمجموعتين The linear discriminant function for two groups [3] [7] [9]

تسمى هذه الدالة بدالة فيشر (Fisher Function) تتوزع فيها المشاهدات توزيعاً طبيعياً. لنفترض ان لدينا مجتمعين ونريد المقارنة بينهما ولنفرض ان هذين المجتمعين لهما نفس مصفوفة التباين والتباين المشترك ($\Sigma_1 = \Sigma_2$) ولهما موجه متوسطات ($\underline{\mu}_1, \underline{\mu}_2$) بالتتابع، وتم اختيار عينتين عشوائيتان ($\underline{x}_{11}, \underline{x}_{12}, \underline{x}_{13}, \dots, \underline{x}_{1n_2}$) و ($\underline{x}_{21}, \underline{x}_{22}, \underline{x}_{23}, \dots, \underline{x}_{2n_2}$) لكل من المجتمعين.

وتكون صيغة الدالة المميزة الخطية لفيشر تكتب بالشكل الاتي:

$$Z = \underline{a}'\,\underline{x} = \left(\underline{\bar{x}}_1 - \underline{\bar{x}}_2\right)' S_p^{-1}\,\underline{x} \text{.....(28)}$$

حيث ان:

$$\underline{a} = S^{-1}\left(\underline{\bar{x}}_1 - \underline{\bar{x}}_2\right) \text{.....(29)}$$

وان مقياس (Mahalanobis) يعتمد بالأساس على القياسات ذات المسافة القليلة بين قيم المتغيرات للمشاهدات الجديدة وقيم المتوسطات للمتغيرات لكل مجموعة يمكن كتابتها بالشكل الاتي:

$$D_i^2 = \left(\underline{\bar{x}}_1 - \underline{\bar{x}}_i\right)' S_p^{-1} \left(\underline{\bar{x}_1} - \underline{\bar{x}}_i\right) \text{.....(30)}$$

إذ أن :

$\underline{\bar{x}}_i$: موجه متوسطات المتغيرات لكل قيمة من المجموعة (i)

S_p^{-1} : هو معكوس مصفوفة التباين والتباين المشترك المقدرة داخل العينتين.

7- تصنيف البيانات classification of data [11]

لنفرض ان:

لدينا عينة بحجم (n).

ان عدد المشاهدات من النوع (0) هي n_1

وان عدد المشاهدات من النوع (1) هي n_2

وان لدينا بيانات من نوع ثنائي (binary) وكانت لدينا تصنيف البيانات كما مبين في الجدول رقم (1) الاتي:

جدول رقم (1) يبين تصنيف البيانات

البيانات الثنائية	(0)	(1)
(0)	A_{11}	A_{12}
(1)	A_{21}	A_{22}

وتكون معايير التصنيف بالشكل الاتي:

1- نسبة التصنيف الصحيحة تحسب بالشكل الاتي:

$$= \left(\frac{A_{11}+A_{22}}{n}\right) * 100\% \text{.....(31)}$$

ومن ثم يمكن حساب معيار خطأ التصنيف الصحيح بالشكل الاتي:

$$\text{معيار خطأ التصنيف} = 100\% - \text{نسبة التصنيف الصحيح} \text{.....(32)}$$

2- نسبة التصنيف الجزئية على مستوى (0) و (1) وتحسب بالشكل الاتي:

$$p(0) = \left(\frac{A_{11}}{n_1}\right) * 100\% \text{.....(33)}$$

$$p(1) = \left(\frac{A_{22}}{n_2}\right) * 100\% \text{.....(34)}$$

1- نسبة التصنيف الكلية على مستوى (0) و (1) وتحسب بالشكل الاتي:

$$p(0) = \left(\frac{A_{11}+A_{21}}{n}\right) * 100\% \text{.....(35)}$$

$$p(1) = \left(\frac{A_{12}+A_{22}}{n}\right) * 100\% \text{.....(36)}$$

8- الجانب التجريبي

8-1 المحاكاة Simulation

تعرف المحاكاة بانها تصميم أنموذج افتراضي مشابه للأنموذج الحقيقي لغرض معرفة سلوك الانموذج للوصول الى الهدف المطلوب وبمعنى اخر هو إعادة صياغة الواقع الفعلي بواقع افتراضي باستعمال صيغ ونماذج معينة وفق أسلوب تكراري لعدة مرات من التكرارات لغرض الوصول الى نتائج تجريبية تهدف الى المقارنة بينها وبين نتائج الجانب التطبيقي.

ان الاعتماد على تجارب المحاكاة في المجال الاحصائي بصورة عامة وفي مجالات التقدير والتنبؤ والتصنيف بصورة خاصة كونه يقارن بين حجوم العينات وعدد المتغيرات في الدراسات ومن ثم التوصل الى أفضل الطرائق تحت الدراسة.

8-2 وصف لتجربة المحاكاة

ان موضوع الرسالة سيتم توليد بيانات بمتغيرات عدد (P=15) بحجم عينة (25 ،50، 100، 150، 250، 400) باستعمال برنامج Matlab (R2015b) وبالاعتماد على المعادلة الاتية:

$$x_1 = N(0,1) \text{.....(3-119)}$$

$$x_{p-1} = N(0,0.1) + x_1 \text{.....(3-120)}$$

$$Y = randi([0,1], n, 1) \text{.....(3-121)}$$

ان هذه البيانات تعاني من مشكلة تعدد خطي.

8-2-1 في حالة حجم العينة (100 – 150 – 250 – 400) وبعدد مركبات (2):

الجدول (4) بين نسب التصنيف
في المحاكاة للدالة المميزة بحجم عينة (100 – 150 – 250 – 400) ومركبات (2)

h=2 n=100	(0)	(1)	نسبة التصنيف الصحيحة
(0)	38	9	80.9%
(1)	11	42	79.2%
نسبة التصنيف الكلي	49%	51%	%80
h=2 n=150	(0)	(1)	نسبة التصنيف الصحيحة
(0)	59	18	76.6%
(1)	18	55	75.3%
نسبة التصنيف الكلي	51.4%	48.6%	%76
h=2 n=250	(0)	(1)	نسبة التصنيف الصحيحة
(0)	90	37	70.9%
(1)	45	78	63.4%
نسبة التصنيف الكلي	54%	46%	67.2%
h=2 n=400	(0)	(1)	نسبة التصنيف الصحيحة
(0)	130	72	64.6%
(1)	82	116	58.6%
نسبة التصنيف الكلي	53%	47%	%61.5

الجدول (5) يبين نسب التصنيف في المحاكاة
للانحدار اللوجستي الثنائي بحجم عينة (100 – 150 – 250 – 400) ومركبات (2)

h=2 n=100	(0)	(1)	نسبة التصنيف الصحيحة
(0)	21	26	49%
(1)	23	30	44.7%
نسبة التصنيف الكلي	44%	56%	%51
h=2 n=150	(0)	(1)	نسبة التصنيف الصحيحة
(0)	40	37	51.9%
(1)	38	35	47.9%
نسبة التصنيف الكلي	52%	48%	%50
h=2 n=250	(0)	(1)	نسبة التصنيف الصحيحة
(0)	64	63	50.4%
(1)	62	61	49.6%
نسبة التصنيف الكلي	50.4%	49.6%	%50
h=2 n=400	(0)	(1)	نسبة التصنيف الصحيحة
(0)	102	100	50.5%
(1)	100	98	49.5%
نسبة التصنيف الكلي	50.5%	49.5%	%50

نلاحظ من الجدولين (4) (5) أعلاه عندما يكون عدد المركبات (2) وبحجوم عينات (100 – 150 – 250 – 400) ان نسب تصنيف الدالة المميزة الخطية كانت اعلى من نسب التصنيف للانحدار اللوجستي الثنائي وبالتالي فأن خطأ التصنيف للدالة المميزة الخطية اقل من خطأ التصنيف للانحدار اللوجستي وهذا يعني ان الدالة المميزة الخطية افضل من الانحدار اللوجستي.

2-2-8 في حالة حجم العينة (100 – 150 – 250 – 400) وبعدد مركبات (4):

جدول (6) يبين نسب التصنيف في المحاكاة
للدالة المميزة بحجم عينة (100 – 150 – 250 – 400) ومركبات (4)

h=4 n=100	(0)	(1)	نسبة التصنيف الصحيحة
(0)	42	11	79.2%
(1)	10	37	84%
نسبة التصنيف الكلي	58%	42%	%79
h=4 n=150	(0)	(1)	نسبة التصنيف الصحيحة
(0)	52	23	69.3%
(1)	21	54	72%
نسبة التصنيف الكلي	48.6%	51.4%	%70.7
h=4 n=250	(0)	(1)	نسبة التصنيف الصحيحة
(0)	84	47	46.1%
(1)	50	69	57.9%
نسبة التصنيف الكلي	53.6%	46.4%	%61.2
h=4 n=400	(0)	(1)	نسبة التصنيف الصحيحة
(0)	122	80	60.4%
(1)	81	117	59.1%
نسبة التصنيف الكلي	50.8%	49.2%	%59.7

جدول (7) يبين نسب التصنيف في المحاكاة
للانحدار اللوجستي الثنائي بحجم عينة (100 – 150 – 250 – 400) ومركبات (4)

h=4 n=100	(0)	(1)	نسبة التصنيف الصحيحة
(0)	29	24	54.7%
(1)	26	21	44.7%
نسبة التصنيف الكلي	55%	45%	%50
h=4 n=150	(0)	(1)	نسبة التصنيف الصحيحة
(0)	38	37	50.7%
(1)	38	37	49.3%
نسبة التصنيف الكلي	50.6%	49.4%	%50
h=4 n=250	(0)	(1)	نسبة التصنيف الصحيحة
(0)	69	62	52..7%
(1)	63	56	47.6%
نسبة التصنيف الكلي	52.8%	47.2%	%50
h=4 n=400	(0)	(1)	نسبة التصنيف الصحيحة
(0)	102	100	50.5%
(1)	100	98	49.5%
نسبة التصنيف الكلي	50.5%	49.5%	50%

نلاحظ من الجدولين رقم (6) (7) أعلاه عندما يكون عدد المركبات (2) وبحجوم عينات (100 – 150 – 250 – 400) ان نسب تصنيف الدالة المميزة الخطية كانت اعلى من نسب التصنيف للانحدار اللوجستي الثنائي ومن ثم فأن خطأ التصنيف للدالة المميزة الخطية اقل من خطأ التصنيف للانحدار اللوجستي وهذا يعني ان الدالة المميزة الخطية افضل من الانحدار اللوجستي.

3-2-8 في حالة حجم العينة (100 – 150 – 250 – 400) وبعدد مركبات (7):

جدول (8) يبين نسب التصنيف في المحاكاة
للدالة المميزة بحجم عينة (100 – 150 – 250 – 400) ومركبات (7)

h=7 n=100	(0)	(1)	نسبة التصنيف الصحيحة
(0)	24	11	68.6%
(1)	12	53	81.5
نسبة التصنيف الكلي	36%	64%	77%
h=7 n=150	(0)	(1)	نسبة التصنيف الصحيحة
(0)	51	27	65.4%
(1)	24	48	66.7%
نسبة التصنيف الكلي	50%	50%	%66
h=7 n=250	(0)	(1)	نسبة التصنيف الصحيحة
(0)	79	40	66.4%
(1)	47	84	64.1%
نسبة التصنيف الكلي	50.4%	49.6%	%65.2
h=7 n=400	(0)	(1)	نسبة التصنيف الصحيحة
(0)	122	73	62.6%
(1)	81	124	60.5%
نسبة التصنيف الكلي	50.8%	49.2%	61.5%

جدول (9) يبين نسب التصنيف في المحاكاة
للانحدار اللوجستي الثنائي بحجم عينة (100 – 150 – 250 – 400) ومركبات (7)

h=6 n=100	(0)	(1)	نسبة التصنيف الصحيحة
(0)	10	25	28.6%
(1)	19	46	70.8%
نسبة التصنيف الكلي	29%	71%	56%
h=7 n=150	(0)	(1)	نسبة التصنيف الصحيحة
(0)	41	37	52.6%
(1)	38	34	47.2%
نسبة التصنيف الكلي	52.6%	47.4%	%50
h=7 n=250	(0)	(1)	نسبة التصنيف الصحيحة
(0)	56	63	48.3%
(1)	62	69	52.7%
نسبة التصنيف الكلي	47.2%	52.8%	%50
h=6 n=400	(0)	(1)	نسبة التصنيف الصحيحة
(0)	95	100	48.7%
(1)	100	105	51.2%
نسبة التصنيف الكلي	48.8%	51.2%	50%

نلاحظ من الجدولين رقم (8) (9) أعلاه عندما يكون عدد المركبات (2) وبحجوم عينات (100 – 150 – 250 – 400) ان نسب تصنيف الدالة المميزة الخطية كانت اعلى من نسب التصنيف للانحدار اللوجستي الثنائي وبالتالي فأن خطأ التصنيف للدالة المميزة الخطية اقل من خطأ التصنيف للانحدار اللوجستي وهذا يعني ان الدالة المميزة الخطية افضل من الانحدار اللوجستي.

9-الاستنتاجات

نستنتج من هذا البحث الاتي:

1- عند تكرار المحاولة 1000 مرة في الدالة المميزة الخطية كلما قل حجم العينة كانت نتائج خطأ التصنيف اقل، اما في الانحدار اللوجستي الثنائي فان زيادة او قلة حجم العينة لا يؤثر على خطأ التصنيف.

2- عند المقارنة بين الطريقتين باستخدام معيار خطأ التصنيف تم التوصل الى ان الدالة المميزة الخطية هي أفضل في تصنيف البيانات من الانحدار اللوجستي الثنائي ولجميع العينات والمتغيرات. وهذا يعني ان نموذج الدالة المميزة أفضل في التنبؤ من الانحدار اللوجستي الثنائي لأنها أعطت اقل خطأ تصنيف

10-التوصيات

بناءا على ما تم التوصل اليه من الاستنتاجات نوصي بالآتي:

1- اجراء دراسات إحصائية في حالة وجود قيم شاذة ومشكلة تعدد خطي واجراء دراسات مقارنة بين الانحدار اللوجستي والتحليل المميز بعد ان تتم المعالجة من وجود الشواذ بأحد الطرق الحصينة والارتباطات بين المتغيرات (بطريقة المربعات الصغرى الجزئية)

2- نوصي بتوسيع انحدار المربعات الصغرى الجزئية وذلك في حالة عدد المتغيرات التوضيحية اكبر من عدد المشاهدات ومن ثم استعمالها في المقارنة بين الانحدار اللوجستي والدالة المميزة الخطية سواء كانت خطية او تربيعية.

11-المصادر

1- البكري، رباب عبد الرضا (2015)، مقارنة بعض الطرائق الخطية لمعالجة مشكلة التعدد الخطي في النماذج مع تطبيق عملي، رسالة دكتوراه– كلية الادارة والاقتصاد – جامعة بغداد.

2- التميمي، رعد فاضل حسن، (2013)،"الانحدار والسلاسل الزمنية أساليب إحصائية تطبيقية متقدمة بأستخدام برنامج Minitab، كتاب، كلية الإدارة والاقتصاد، الجامعة المستنصرية.

3- الحمداني، بسمة رشيد،2014، "تميز الكادر الطبي حسب معرفتهم للتصنيف الدولي ((ICD-10 باستعمال الدالة المميزة"، رسالة ماجستير في جامعة بغداد، كلية الادارة والاقتصاد.

4- عباس، علي خضير، 2012، " استخدام أنموذج الانحدار اللوجستي في التنبؤ بالدوال ذات المتغيرات الاقتصادية التابعة النوعية"، مجلة كركوك للعلوم الادارية والاقتصادية، مجلد 2، العدد 2.

5- Abdi, hervi, 2010," Partial least squares regression and projection on latent structure regression (PLS Regression)", John Wiley & Sons.

6- Banh V. Nguyen, David M. Rocke, 2004, " On partial least squares dimension reduction for microarray-based classification: a simulation study, Division of Biostatistics, School of Medicine, University of California.

7- Boaz Nadler, Ronald R.Coifman, 2005, " Partial least squares, Beer's law and the net analyte signal: statistical modeling and analysis", Department of Mathematics, Yale University.

8- Erik Brorson, Asterios Geroukis, 2014, "A comparison between discriminant analysis and logistic regression using principal components", Department of Statistics, Uppsala University, Uppsala University.

9- Leo H. Chiang, Evan L. Russell, Richard D. Braatz,2000, " Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis", Department of Chemical Engineering, universty of Illinois.

10- Norliza Adnan, Ahmad Maizah Hura, Adnan Robiah, 2006, "A Comparative Study on Some Methods For Handling Multicollinearity Problems, Department of Mathematics, uivirsty of malysia.

11- Tormod Næs and Bjorn-Helge Mevik, "Understanding the collinearity problem in regression and discriminant analysis", Journal of Chemo metrics, P (413-426), 2001.

discriminate analysis and logistic regression by use partial least square_

Abstract

The method binery logistic regression and linear discrimint function of the most important statistical methods used in the classification and prediction when the data of the kind of binery (0,1) you can not use the normal regression therefore resort to binary logistic regression and linear discriminant function in the case of two group in the case of a Multicollinearity problem between the data (the data containing high correlation) It became not possible to use binary logistic regression and linear discriminant function, to solve this problem, we resort to Partial least square regression.

In this, search the comparison between binary logistic regression and linear discriminant function using error Category. Where the data has been generating a variable response (Y) binery data (0,1) containing Multicollinearity problem by the samples (50-100-150-250-400) and the variables (5-10-15). Multicollinearity problem has been processed using a method partial least square The research found that linear discriminant function It is the best in the classification of data from binary logistic regression classified as linear discriminant function the data correctly and more accurate than binary logistic regression.

Keyward: linear discriminant function- binary logistic regression- partial least square– multicollinearity problem – ratio of classification – simulation.