Iraqi Journal of Statistical Sciences, Vol. 17, No. 1, 2020, Pp(54-63)

المجلة العراقية للعلوم الإحصائية

http://stats.uomosul.edu.iq

بناء الدالة التمييزية – مراجعة مقال –

ندى نزار محمد ، نجلاء سعد ابراهيم و زيد طارق صالح

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

معلومات النشر

تاريخ المقالة:
تم استلامه في 28 كانون الثاني 2020
تم القبول في 7 اذار 2020
متاح على الإنترنت في 1 حزيران 2020

الكلمات الدالة:
عملية التمييز
طريقة حدود الثقة لـــ Roy-Bose
طريقة اختبارT-test
طريقة الاختيار الأمامي
طريقة الحذف العكسي
طريقة الانحدار المتدرج

المراسلة:
ندى نزار محمد
nada-nazar1984@uomsol.edu..iq

الخلاصة

يستخدم التحليل التمييزي لتصنيف البيانات إلى مجاميع مختلفة تبعا لبعض المعايير. وتعتمد عملية التصنيف على اختيار المتغيرات ذات الدلالة إحصائية المعنوية، ومن ثم استخدام هذه المتغيرات لبناء الدالة التمييزية التي مستخدمة لتصنيف البيانات. ولغرض استكشاف المعنوية الإحصائي للمتغيرات الداخلة في التحليل وقابلية كل متغير على المساهمة في عملية التمييز، والتي من خلال هذه المتغيرات نستطيع بناء دالة تمييزية، تم استخدام طريقة حدود الثقة لـ Roy–Bose وطريقة اختبار T–test، وهي من الطرائق المتبعة في اختيار المتغيرات في التحليل التمييزي. كذلك تم استخدام طرائق أخرى لاختيار المتغيرات ذات التأثير المعنوي في التحليل التمييزي، وهذه الطرائق هي طرائق الانحدار المتمثلة بـ (طريقة الاختيار الأمامي وطريقة الحذف العكسي وطريقة الانحدار المتدرج). بالإضافة إلى ذلك، تم تطبيق تحليل المركبات الرئيسية في التحليل العاملي، وهي أحد الطرق المستخدمة في تحليل متعدد المتغيرات، على البيانات، حيث تم بناء الدالة التمييزية بالاعتماد على المتغيرات التي تم اختيارها .

DOI:10.33899/IQJOSS.2020.165449 , ©Authors, 2020, College of Computer Science, and Mathematics University of Mosul.
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

1. المقدمة Introduction

يعد التحليل التمييزي أحد الإجراءات المهمة في تحليل متعدد المتغيرات Multivariate Analysis وذلك بالاستناد إلى مقاييس معينة وعلى خصائص المشاهدة التي لابد أن تتوافق مع خصائص المجتمع الذي ستنسب إليه بدرجة اكبر من درجة توافقها مع أي مجتمع آخر. ويعد التحليل التمييزي من الأساليب الإحصائية المهمة والذي يستخدم في كثير من مجالات الحياة؛ فعلى سبيل المثال، تستخدم دوال التمييز لغرض الوقوف على مدى إمكانية التنبؤ بحدوث الظواهر المختلفة اعتماداً على مقاييس محددة. كذلك يمكن استخدام هذه التقنية لمعرفة المتغيرات التي تساهم في التصنيف، وهي كما في تحليل الانحدار الذي لديه استخدامين الوصف (التمييز) والتنبؤ (Hamodat,2005). ان من الافتراضات الخاصة بالتحليل التمييزي يجب ان تكون المجاميع المدروسة تتوزع توزيع الطبيعي متعدد المتغيرات ،ففي حالة الدالة المميزة الخطية يشترط فيها ان تكون مصفوفة التباين والتباين المشترك متساوية للمجاميع كافة، ويتم استخدام الدالة المميزة التربيعية في حالة عدم التساوي.(Akkar,2008) . ان التحليل التمييزي يختلف عن تحليل الانحدار في أن المتغير المعتمد في التحليل التمييزي هو متغير اسمي (Nominal Variable) وهو من المتغيرات النوعية (Qualitative Variable) بينما المتغير المعتمد في تحليل الانحدار هو على الأكثر متغير مستمر (Continuous Variable) وهو من المتغيرات الكمية (Quantitative Variable). مع ذلك، فإنه هنالك أوجها عميقة من التشابه بين التحليلين من ناحية الهدف الأساسي للتحليل، حيث أن كلا التحليل يُستخدمان لوصف العلاقة بين المتغير المعتمد والمتغيرات المستقلة وذلك من خلال نمذجة البيانات (Al–Hamdani,2014). ويستخدم التحليل التمييزي في مجالات متعددة، فعلى سبيل المثال يوظف التحليل التمييزي في مجالات الطب والزراعة والتعليم وعلم الاجتماع والجغرافية وغيرها من المجالات التطبيقية. فعلى سبيل المثال في مجال التعليم، تستخدم دالة التمييز الخطية في معرفة مستوى كفاءة الأداء للطلبة وذلك بالاعتماد على درجات أو تقديرات الطلبة في مجموعة من الاختبارات؛ حيث يُمكِننا التحليل التمييزي من معرفة مدى تأثير الاختلاف في البيئة

والجنس ونوع التدريس (خصوصي، عام) على تفوق الطلاب وكفاءتهم. وكذلك فان القبول في المدارس يمكن معالجته من خلال إيجاد تركيبة خطية من المقاييس (المؤشرات) والتي من خلالها نستطيع تحديد المعايير الملائمة للقبول أو عدمه. أما في علم الجغرافية، فان الدالة المميزة تعد الأسلوب الأكثر استخداما في علوم الأرض من بين الأساليب متعددة المتغيرات الأخرى في دراسة مختلف الظواهر الجغرافية(Al-Zubaei & Al-Mashhadani,1998) .

2.التحليل التمييزي **Discriminant analysis** يعد التحليل التمييزي من الأساليب الإحصائية المهمة في متعدد المتغيرات التي تهتم بتميز بين مجموعتين أو أكثر من خلال إيجاد توافيق خطية للمتغيرات التوضيحية، حيث يعد التمييز والذي يسمى بدالة فيشر (Fisher) طريقة فعالة لنمذجة البيانات فيما لو تحققت فروض التحليل. حيث يُشترط أن يكون توزيع المتغيرات التوضيحية توزيعاً طبيعياً، وأن المجاميع المصنفة لها مصفوفات تباين وتباين المشترك متساوية، وأيضا يشترط عدم وجود ارتباط بين المتغيرات التوضيحية وعدم وجود قيم شاذة (Hamodat,2005).

2-1 الافتراضات الخاصة بالتحليل التمييزي

ان من الافتراضات الخاصة بالتحليل التمييزي ان تكون المجاميع ذات توزيع طبيعي متعدد متغيرات في حالة الدالة التمييزية الخطية يشترط فيها ان تكون مصفوفة التباين والتباين المشترك متساوية للمجاميع كافة ، كذلك يتطلب ان تكون حجم العينة كبيرة وان المجاميع المختلفة تتضمن على الأقل 20 مشاهدة لكل متغير توضيحي وأيضا يشترط عدم وجود ارتباط بين المتغيرات التوضيحية وعدم وجود قيم شاذة بينها(Demosthenes, 2006)

2-2 دالة التمييز الخطية **Linear Discriminant Function**

هي الدالة التي يمكن من خلالها التمييز بين المجموعات اي بمعني الفصل بين المشاهدات ووضع كل مشاهدة في المجموعة التي تعود لها، وتعد الدالة التمييزية نموذج رياضي بالإمكان صياغته من خلال مؤشرات عينة اختيرت بشكل عشوائي من مجموعتين مختلفتين وان هذه الدالة تمكننا من اختبار أي مفردة وتحدد المجموعة التي تعود إليها. ويستخدم التحليل التمييزي في عملية التوقع حيث يأتي الباحث بعدة متغيرات يتوقع أن تميز بين المجتمعين المراد دراستهما لنحصل على دالة تمييزية تستخدم في تصنيف المشاهدات بين المجتمعين تسمى هذه الدالة بدالة فيشر أو الدالة المميزة الخطية لمجموعتين. نفترض بأن لدينا مجتمعين، المجتمع الأول يُخصص له الرقم (0)، والمجتمع الثاني يُخصص له الرقم (1)، كذلك لدينا n_0، n_1 التي تم اختيارها من كل مجتمع على التوالي , نفترض أن لدينا قيم مشاهدة لـ m من المتغيرات العشوائية التي يمكن الاعتماد عليها بالتصنيف وهي X_m ,..., X_2 , X_1 لتكون الدالة التصنيفية هي: (anderson,1985)

$$Z = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_m X_m \quad (1)$$

حيث أن:

α: معلمات النموذج وتستخدم في عملية التصنيف.

m: عدد المتغيرات.

$\underline{X}$: متجه المتغيرات.

$$Z = \underline{X}' \underline{S}^{-1} [\underline{\bar{X}}_{(0)} - \underline{\bar{X}}_{(1)}] \quad (2)$$

علماً أن:

$$\alpha = \sum^{-1} [\underline{\bar{X}}_{(0)} - \underline{\bar{X}}(1)] \quad (3)$$

حيث أن: $\sum$ هي التباين المجمع pooled variance تحسب من مصفوفة التباين والتباين المشترك .

$$\sum = \frac{[(n_1-1)\underline{\sum}_{(0)} + (n_2-1)\underline{\sum}_{(1)}]}{n_1+n_2-2} \quad (4)$$

يمكن كتابة النقطة الفاصلة بين المجموعتين كما يلي :

$$(Cut\ Point) = \frac{1}{2}\left(\underline{\bar{X}}_0 - \underline{\bar{X}}_1\right)' \sum^{-1} \left(\underline{\bar{X}}_0 - \underline{\bar{X}}_1\right) \quad (5)$$

وهي النقطة الفاصلة بين المجموعتين فإذا كانت قيمة الدالة بعد تعويض قيم المفردة فيها اكبر من هذه النقطة إذن المفردة تعود للمجموعة الأولى أما إذا كانت قيمة الدالة اكبر من هذه النقطة إذن المفردة تعود للمجموعة الثانية.(Zhang, 2000)

3.الطرائق المستخدمة في اختيار المتغيرات للدالة التمييزية

أن من أهم الأهداف عند إدخال عدداً من المتغيرات في دراسة لظاهرة ما هو كيفية اختيار مجموعة المتغيرات التوضيحية المتضمنة في النموذج، أي بمعنى كيفية اختيار أفضل نموذج يتوصل إلى الهدف المنشود(Al-Rawi,1987) . ومن هذه الطرائق:

1.3 حدود الثقة لـ Roy -Bose

يتم استخدام حدود الثقة لروي– بوس لتعيين المتغيرات المهمة التي ستدخل في الدالة، ويمكن تلخيص هذه الطريقة بالخطوات التالية:

1–نجد قيمة F الجدولية بالرجوع إلى جداول خاصة به:

$$Tab'F = F(\alpha/2\ , m, n1 + n2 - 1) \qquad (6)$$

2– تعيين قيمة T الجدولية وكالآتي:

$$Tab\,T = \left[\frac{(n_1 + n_2 - 2)\,\mathrm{m}}{n_1 + n_2 - m - 1}\ tab.F\right]^{\frac{1}{2}} \qquad (7)$$

3– إيجاد Selection Vectorويرمز له بالرمز (a) وهو متجه غير صفري، حيث يكون عدد هذه المتجهات بعدد المتغيرات التوضيحية (m). حيث أن:

$a_1 = [1 \quad 0 \quad 0 \ldots. \quad]$
$a_2 = [0 \quad 1 \quad 0 \quad 0 \ldots.]$
.
.
$a_m = [0 \quad 0 \qquad \ldots.1]$

4– ثم بعد ذلك نجد حدود الثقة لروي–بوس وتكتب بالصيغة التالية:

$$\underline{a}'m(\underline{\bar{X}}_1 - \underline{\bar{X}}_2) - \sqrt{\underline{a}'m\underline{\sum}a_m\left[\tfrac{n_1+n_2}{n_1.n_2}\right]}\ T < \underline{a}'m < \underline{a}'m(\underline{\bar{X}}_1 - \underline{\bar{X}}_2) + \sqrt{\underline{a}'m\underline{\sum}a_m\left[\tfrac{n_1+n_2}{n_1.n_2}\right]} \qquad (8)$$

5. تكتب المتباينة $C.I(L) < a'm < C.I(u)$

والقرار هو انه إذا احتوت المتباينة على الصفر يعني عدم وجود أهمية لذلك المتغير أي أن المتغير لا يختلف وسطه الحسابي في كلا المجموعتين ، إما إذا لم تحتوي المتباينة على الصفر يعني وجود أهمية لذلك المتغير. (Morrison, 1976)

2.3 اختبار الفرق بين متوسطي المجموعتين (اختبار t)

لاختبار هل ان الفرق معنوي بين متوسطي مجموعتين نستخدم اختبار (t) للمقارنة بين الأوساط الحسابية، ، والفرضية الخاصة باختبار (t) هي:

$H_0 : \mu_0 = \mu_1$
$H_A : \mu_0 \neq \mu_1$

ويمكن حساب المختبر الاحصائي t حسب الصيغة التالية:

$$t = \frac{\bar{X}_{(0)} - \bar{X}_{(1)}}{\sum(\frac{1}{n_1} + \frac{1}{n_2})} \qquad (9)$$

حيث أن: $\sum$: التباين المجمع pooled variance للمجموعتين. $\bar{X}_{(i)}$:الوسط الحسابي للمجموعة i

وبمقارنة قيمة هذه الأحصاءة مع قيمة t الجدولية بدرجة حرية (n_1+n_2-2) ومستوى معنوية (α/2) يتم معرفة المتغير معنوياً أو غير معنوي. (الراوي،1979)

3.3 التحليل العاملي Factor Analysis

هو أسلوب إحصائي الغرض منه تبسيط الارتباطات بين مختلف المتغيرات الداخلة في التحليل من اجل الوصول إلى العوامل المشتركة التي تصف وتفسر العلاقة بين المتغيرات الداخلة في التحليل ومن طرق التحليل العاملي طريقة المركبات الرئيسية.(Dabdob,1990).

1.3.3 المركبات الرئيسية Principal Component

ان تحليل المركبات الرئيسية يستعمل لإيجاد مجموعة صغيرة من التراكيب الخطية للتباينات وتكون غير مترابطة مع بعضها وهذه التراكيب الخطية تختار أعلى التباينات وان تحليل المركبات الرئيسية في حالة وحدات القياس متشابهة للمتغيرات نستعمل القيم الأصلية ،وعندما تختلف تحول الى القيم المعيارية وبعدها يتم حساب مصفوفة الارتباط ثم إيجاد القيم الذاتية والمتجهات المميزة , والمركبات الرئيسية تعتبر تراكيب خطية لـ m من المتغيرات:

$$PC_{ij} = a_{1j}X_{1j} + a_{2j}X_{2j} + \ldots + a_{mj}X_{mj} \qquad (10)$$

حيث أن: PC_{ij}: القيمة i للمكون الرئيسي j. a_{ij}: قيم المتجهات المميزة (a_j) (Characteristic Vectors) وقد تسمى أيضاً Eigen Vector المرافقة للجذور المميزة λ_j (Characteristic Roots) وقد تسمى Eigen Values للمصفوفة المستخدمة. يمكن إيجاد العلاقة (10) من خلال مصفوفة التباين المشترك أو مصفوفة الارتباط. وباستخدام أسلوب المصفوفات:

$$\underline{PC} = AX \quad (11)$$

PC: تمثل مصفوفة المكونات الرئيسية $PC1, PC2, \dots, PC_m$

A: تمثل مصفوفة المتجهات المميزة $a1, a2, \dots, a_m$

X: تمثل مصفوفة المتغيرات التوضيحية $X1, X2, \dots, X_m$. (Aguilera, Escambia,2006)

4.3 أسلوب الاختيار الأمامي Forward Selection Procedure

تبدأ معادلة الانحدار بهذه الطريقة بدون أي متغير توضيحي، ثم يتم اختيار المتغيرات التوضيحية التي تدخل المعادلة واحداً تلو الأخرى ونتوقف عن الاختيار عندما تقل قيمة F المحسوبة الجزئية عن قيمة معينة من F الجدولية. ان المتغير التوضيحي الذي يرشح للدخول في الانحدار في أي خطوة يتم تثبيته نهائياً في الانحدار إذا ما ثبت تأثيره المعنوي في تلك الخطوة، فالمتغير التوضيحي الأول الذي يدخل المعادلة هو المتغير الذي له أعلى F محسوبة إما المتغير الثاني الذي يضاف إلى المعادلة أعلاه هو المتغير الذي له اعلى F جزئية بوجود المتغير الأول المنتخب بالخطوة الأولى وتزيد عن (F_{IN}) الجدولية المعينة لتلك الخطوة. وهكذا نستمر بإضافة المتغير الذي له أعلى F جزئية عن F_{IN} إلى ان نصل إلى أعلى F جزئية تقل عن F_{IN} فعندئذ نتوقف عن الإضافة، وينتهي الحل بأخذ المتغيرات التوضيحية في الخطوة السابقة لتلك الخطوة (طيب، 2005).

5.3 أسلوب الحذف الخلفي Backward Elimination Procedure

يتم في هذا الاختيار اختبار جميع المتغيرات التوضيحية في معادلة الانحدار ثم نبدأ بحذف المتغيرات التوضيحية ذات التأثير غير المعنوي واحداً بعد الأخر حتى نصل إلى الصيغة النهائية التي تحتوي على المتغيرات ذات التأثير المعنوي. أي ان المتغير الأول الذي يحذف من بين جميع المتغيرات التوضيحية في معادلة الانحدار هو المتغير الذي له اقل قيمة F جزئية والتي تكون اقل قيمة من قيمة F_{out} التي تنتخب قيمة لـ F جدوليه. إما اذا كانت اقل قيمة F جزئية اكبر من F الجدولية F_{out} فنتوقف عن الحذف وتكون بذلك جميع المتغيرات التي في المعادلة هي متغيرات توضيحية مهمة وذات تأثير معنوي على متغير الاستجابة Y. وبعد حذف المتغير الاول نحسب قيم F الجزئية لبقية المتغيرات التوضيحية ونحذف المتغير الذي له F جزئية التي تكون اقل من F_{out} المعينة وهكذا لكل خطوة، ونتوقف عن الحذف في المرحلة التي تكون قيم F الجزئية اكبر من قيمة F_{out} المعينة لتلك المرحلة (كاظم والدليمي ،1988).

6.3 أسلوب انحدار المتدرج Stepwise Regression Procedure

وهي طريقة تربط بين طريقتي الاختيار الأمامي والاختيار الخلفي، وهذه الطريقة تحتاج إلى قيمتين من قيم F الجدولية هما F_{IN} للاختيار الأمامي وF_{out} للحذف العكسي وخطواته:

الخطوة الأولى: نحسب قيمة F الجزئية لكل متغير توضيحي، ثم نختار أعلى F جزئية ذات معنوية إحصائية، وعند عدم وجود هكذا متغير نوقف العمليات الإحصائية ونتخذ قرار بعدم وجود أي متغير مؤثر على الظاهرة.

الخطوة الثانية: لاختيار المتغير التوضيحي الثاني نعيد الخطوة أعلاه وقبل أن نختار متغيراً توضيحياً ثالثاً نجري الطريقة العكسية لمعرفة أهمية بقاء المتغير الذي اختير في المرة الأولى وذلك لتحديد حذفه من المعادلة أو إبقاءه فيها، ويتم التوقف بإتباع خطوات هذه الطريقة عند عدم معنوية أي متغير في الاختيار الأمامي، وبهذا الأسلوب يمكن الاستمرار إلى أن نصل إلى النموذج النهائي .

4.تصنيف البيانات classification of data

لنفرض ان لدينا عينة بحجم (n) وان 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) الآتي: (الهبيل,2013)

جدول رقم (1) يبين تصنيف البيانات

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

وتكون معايير التصنيف بالشكل الآتي: دقة التصنيف تحسب بالشكل الآتي:

$$\text{دقةالتصنيف} = \left(\frac{A_{11}+A_{22}}{n}\right) * 100\% \quad (12)$$

وبالتالي يمكن حساب معيار خطأ التصنيف بالشكل الآتي:

$$\text{معيار خطأ التصنيف} = 100\% - \text{دقة التصنيف} \quad (13)$$

5.الجانب التطبيقي

تم اخذ البيانات من الموقع العالمي للبيانات (https://archive.ics.uci.edu/ml/index.php)، حيث أن عدد المتغيرات التوضيحية (12) و بحجم عينة ($n = 100$) ، إما المتغير المعتمد y فمصنف إلى مجموعتين (المجموعة الأولى يرمز لها بالرمز 0) و(المجموعة الثانية يرمز لها بالرمز 1) ، تم تطبيق التحليل التمييزي على هذه البيانات باستخدام برنامجي SPSS و R.

5-1 اختيار متغيرات الدالة التمييزية:

قبل اختبار معنوية المتغيرات تم إجراء التحليل الإحصائي للتعرف على عدد مشاهدات في كل مجموعة:

عدد المشاهدات	المجاميع
56	المجموعة الأولى 0
44	المجموعة الثانية 1
100	المجموع

1. طريقة Roy-Bose

لغرض تشخيص المتغيرات المعنوية نلجأ إلى إيجاد حدود الثقة لروي–بوس ولمعرفة معنوية أو عدم معنوية المتغيرات نطبق المعادلة (7) لإيجاد قيمة (T) ثم نجد حدود الثقة وذلك بتطبيق المعادلة (8) ، إذا احتوت حدود الثقة على الصفر معنى ذلك هو عدم وجود أهمية لذلك المتغير ،وهكذا يتم تطبيق هذه الطريقة على جميع المتغيرات، والجدول رقم (2) يوضح معنوية أو عدم معنوية المتغيرات حيث أظهرت المتغيرات $(X_4, X_5, X_6, X_7, X_8, X_9, X_{10})$ معنويتها بهذه الطريقة اذا كانت حدود الثقة لديها لا تساوي الصفر .

الجدول (2) اختيار متغيرات الدالة التمييزية باستخدام حدود الثقة لروي–بوس

المتغيرات	حدود الثقة لروي بوس	المعنوية
X_1	-0.07467<a_1<0.125111	غير معنوي
X_2	-6.41849<a_2<-1.202076	غير معنوي
X_3	-3.13536<a_3<-1.059996	غير معنوي
X_4	0.24619 <a_4<1.06869	معنوي
X_5	3.49555<a_5<2.044696	معنوي
X_6	.527520121.35870<a_6<	معنوي
X_7	2.975725<a_7<3.546078	معنوي
X_8	0.79253<a_8<1.643743	معنوي
X_9	0.55016 <a_9<1.360797	معنوي
X_{10}	2.982409<a_{10}<4.99863	غير معنوي
X_{11}	-0.129438<a_{11}<0.083642	غير معنوي
X_{12}	-0.11501<a_{12}<0.005873	غير معنوي

2. طريقة اختبار (T–Test)

تم استخدام اختبار t لاختبار معنوية المتغيرات ويتم ذلك عن طريق اختبار الفرضية: $H_0 : \mu_{i0} = \mu_{i1}$ وبالاعتماد على المعادلة (9) نستخرج قيمة t المحسوبة
$H_A : \mu_{i0} \neq \mu_{i1}$

ونقارنها مع قيمة t الجدولية تحت مستوى معنوية (0.05) ودرجات حرية (n_1+n_2-2) أي أن:

tab.t($\frac{\propto}{2}$,n_1+n_2-2)= 1.96

ومن الجدول أدناه نلاحظ ان كل من المتغيراتX_4($X_5, X_7, X_8, X_9, X_{10}$,) ظهرت فيه القيمة المحسوبة اكبر من القيمة الجدولية أي ترفض فرضية العدم وتقبل الفرضية البديلة، إما باقي المتغيرات فقد تم إهمالها.

الجدول رقم (3) يبين نتائج اختبار t

X_I	$Cal.t$
X_1	-0.221
X_2	-0.469
X_3	-1.351
X_4	-3.410*
X_5	-5.552*
X_6	-1.431
X_7	-3.289*

X_8	-10.874*
X_9	-6.312*
X_{10}	-5.597*
X_{11}	-0.885
X_{12}	-1.468

3. المركبات الرئيسية:

نلاحظ في الجدول رقم (4) أدناه أن تم تكوين 5 مركبات رئيسية، حيث كانت قيم الجذور المميزة لهذه المركبات اكبر من الواحد وكالاتي (3.094, 1.549, 1.317, 1.113, 1.010) على التوالي، إذ تم تمييز المتغيرات المؤثرة في كل مكون رئيسي من خلال مصفوفة المتجهات المميزة، ،حيث يتم اختيار هذه القيم عندما تكون اكبر او تساوي (0.5) حتى تكون ذات تأثير معنوي، وبذلك تم اختيار المتغيرات (x4, x5, x6, x7, x8, x9, x10, x12).

الجدول رقم (4): المركبات الرئيسية

المتغيرات	PC1	PC2	PC3	PC4	PC5
x1	-0.074	0.065	0420	-0.185	0.006
x2	0.315	0.488	0.401	0.274	0.394
x3	0.303	0720	0.078	0.133	-0.060
x4	0.318	0.142	-0.364	-0.264	0.665
x5	0.534	-0.638	-0.071	0.125	0.102
x6	0.348	-0.665	0.184	0.031	0.208
x7	0.684	-0.025	0.361	0.101	0.272
x8	0.791	0.000	-0.030	-0.015	-0.182
x9	0.807	0.126	-0.242	-0.058	-0.207
x10	0.709	0.205	-0.341	-0.161	-0.251
x11	0.369	-0.086	0.551	-0.282	-0.350
x12	0.086	-0.054	-0.059	8440	-0.140

2. طريقة الاختيار الأمامي **Forward Selection method**

تبدأ هذه الطريقة، كما أشرنا سابقا، بدون إي متغير توضيحي، ويتم استبعاد القيم التي ليس لها تأثير معنوي على النموذج. نلاحظ في الجدول رقم 5 أدناه أنه تم ترشيح المتغير (X_8) في الخطوة الأولى وفي الخطوة الثانية تم اختيار المتغير (X_5) بالإضافة إلى المتغير (X_8) في الخطوة الثانية. وفي الخطوة الثالثة والأخيرة تم إضافة المتغير (X_4). وبذلك تم اختيار المتغيرات (X_8, X_5, X_4) المهمة في بناء الدالة التمييزية.

الجدول رقم (5):المتغيرات المعنوية بطريقة الاختيار الأمامي

Step	Models	p-value
Step 1	(Constant)	0.045
	X_8	0.000
Step 2	(Constant)	0.148
	X_8	0.000
	X_5	0.001
Step 3	(Constant)	0.001
	X_8	0.000
	X_5	0.001
	X_4	0.004

3. أسلوب الحذف العكسي **Backward Elimination Method**

ان طريقة الحذف العكسي هي عكس طريقة الاختيار الأمامي، تبدأ بكل المتغيرات التوضيحية في النموذج ويتم استبعاد المتغير الذي لا يؤثر على النموذج. ويلاحظ من الجدول رقم 6 أدناه أنه في الخطوة الأولى تم إدخال جميع المتغيرات قيد الدراسة وعددها 12 متغيراً وفي الخطوة الثانية تم استبعاد المتغير X_7 وفي الخطوة الثالثة تم استبعاد المتغير X_3 وصولا إلى الخطوة الثامنة حيث تم اختيار المتغيرات $(X_4, X_5, X_8, X_{10}, X_{11})$ التي أثبتت معنويتها وفقا لهذه الطريقة.

جدول رقم (6): المتغيرات المعنوية بطريقة الحذف العكسي

steps	1	2	3	4	5	6	7	8
Constant	-.638-	-.635-	-.625-	-.638-	-.679-	-.782-	-.448	-.370
X1	.112	.111	.108	.104	.099	.100	.100	
X2	-.000-	-.003-	-.003-	-.003-				
X3	.003	.003						
X4	.191	.191	.190	.193	.182	.189	.177	.174
X5	.036	.036	.035	.035	.037	.032	.034	.033
X6	-.019-	-.019-	-.019-	-.019-	-.019-			
X7	.000							
X8	.575	.575	.581	.595	.581	.576	.584	.586
X9	.055	.054	.050					
X10	.013	.013	.014	.017	.017	.018	.017	.016
X11	-.108-	-.108-	-.107-	-.103-	-.110-	-.112-	-.125-	-.112-
X12	.160	.158	.162	.167	.149	.164		

4. طريقة الانحدار المتدرج:

لتحديد أفضل المتغيرات المؤثرة على المتغير المعتمد تبين ان هذه الطريقة مشابهة تماما لطريقة الاختيار الأمامي حيث تم اختيار المتغيرات (X_8, X_5, X_4) المهمة في بناء الدالة التمييزية.

5-2 بناء الدالة التمييزية :

بعد التعرف على المتغيرات المهمة والتي أظهرت معنويتها في جميع الطرائق سوف نقوم باستخدامها في بناء الدالة التمييزية:

1 . التحليل التمييزي بالاعتماد على اختبار Roy – Bose

يتم بناء الدالة التمييزية في روي- بوس بالاعتماد على المتغيرات $(x4, x5, x6, x7, x8, x9, x10)$ التي اظهرت معنويتها ويتم ذلك حسب المعادلة:

$$\underline{\alpha} = \underline{\sum}^{-1} \left(\overline{\underline{X}}_{(0)} - \overline{\underline{X}}_{(1)} \right)$$

سوف نقوم بإجراء التحليل التمييزي على جميع الطرائق المذكورة أعلاه:

$$\alpha = \begin{bmatrix} 0.1857 \\ -0.4219 \\ 0.4071 \\ -0.0938 \\ -4.1553 \\ -0.8770 \\ 0.2530 \end{bmatrix} \quad ; Z = 0.1857x_{i4} - 0.4219x_{i5} + 0.4071x_{i6} - 0.0938x_{i7} - 4.1553x_{i8} - 0.8770x_{i9} + 0.2530x_{i10}$$

2. التحليل التمييزي بالاعتماد على اختبار t

تم حساب الدالة التمييزية اعتمادا على المتغيرات(x4,x5,x7,x8,x9,x10) التي تم اختبارها باختبار t.

$$\alpha = \begin{bmatrix} -2.1625 \\ -0.3652 \\ 0.0819 \\ -6.2208 \\ -0.3564 \\ -0.1705 \end{bmatrix} \quad ; Z = -2.1625x_{i4} - 0.3652x_{i5} + 0.0819x_{i7} - 6.2208x_{i8} - 0.3564x_{i9} - 0.1705x_{i10}$$

ومن خلال قيمة (α) يتم بناء الدالة التمييزية بالاعتماد على متغيرات المعنوية لاختبار t

3.التحليل التمييزي بالاعتماد على المركبات الرئيسية

تم حساب الدالة التمييزية اعتمادا على المتغيرات(x4,x5,x6,x7,x8,x9,x10,x11)

$$\alpha = \begin{bmatrix} -2.5922 \\ -0.1641 \\ -0.7798 \\ 0.1847 \\ -5.9117 \\ -0.0368 \\ -0.2789 \end{bmatrix} \quad Z = -2.5922x_{i4} - 0.1641x_{i5} - 0.7798x_{i6} + 0.1847x_{i7} - 5.9117x_{i8} - 0.0368x_{i9} - 0.8789x_{i10} - 2.7903x_{i11}$$

4. التحليل التمييزي بالاعتماد على طريقة الاختيار الأمامي

وفي طريقة الاختيار الأمامي كانت الدالة التمييزية لها:

$$\alpha = \begin{bmatrix} 1.1963 \\ 0.0267 \\ 0.4229 \end{bmatrix} \quad ; \quad Z = 1.19639X_{i4} + 0.0267X_{i5} + 0.4229X_{i8}$$

5. **التحليل التمييزي بالاعتماد على أسلوب الحذف العكسي**

وفي طريقة الاختيار العكسي تم تكوين الدالة التمييزية كالاتي:

$$\alpha = \begin{bmatrix} -1.995 \\ -0.3745 \\ -6.6696 \\ -0.1897 \\ 1.2714 \end{bmatrix} \; ; \; Z = -1.995X_{i4} - 0.3745X_{i5} - 6.6696X_{i8} - 0.1897X_{i10} + 1.2714X_{i11}$$

6. التحليل التمييزي بالاعتماد على الانحدار المتدرج

وفي طريقة الاختيار التدريجي تبين ان هذه الطريقة مشابهة تماما لطريقة الاختيار الأمامي كانت الدالة التمييزية لها:

$$9X_{i4} + 0.0267X_{i5} + 0.4229X_{i8} Z = 1.1963$$

من الجدول أدناه نلاحظ في الدالة التمييزية لجميع المتغيرات ان النسب المئوية لدقة التصنيف تعني كم من البيانات صنفت ضمن المجموعة الأولى او المجموعة الثانية او صنفت بشكل خاطئ لمشاهدات المتغير المعتمد ومن خلال الجدول (7) في طريقة الروي بوس نجد ان 47 مشاهدة صنفت بشكل صحيح من مجموع 56 مشاهدة صنفت ضمن المجموعة الاولى وان 41 مشاهدة من مجموع 44 صنفت على صنفت ضمن المجموعة الثانية وان دقة التصنيف بلغت 88.0% وهذا يعني ان احتمال خطأ التصنيف 12.0% ، وفي اختبار t نجد دقة التصنيف بلغت 85.0% وهذا يعني ان احتمال خطأ التصنيف 15.0% , وفي المركبات الرئيسية بلغت دقة التصنيف الصحيحة 87.0% وهذا يعني ان احتمال خطأ التصنيف 13.0%, وفي طريقة الاختيار الامامي وطريقة الانحدار المتدرج نجد دقة التصنيف بلغت 86.0% وهذا يعني ان احتمال خطأ التصنيف 14.0% , وفي طريقة الحذف العكسي دقة التصنيف بلغت 85.0% وهذا يعني ان احتمال خطأ التصنيف 15.0% , نلاحظ ان نسبة الخطأ لجميع الطرائق متقاربة وقليلة نسبيا وهذا دليل على كفاءة الطرائق المستخدمة .

جدول (7): يبين نسب التصنيف

الدالة التمييزية باستخدام Roy-Bose					الدالة التمييزية باستخدام اختبار t			الدالة التمييزية باستخدام المركبات الرئيسية		
y		Predicted Group Membership		المجموع	Predicted Group Membership		المجموع	Predicted Group Membership		المجموع
		0	1		0	1		0	1	
Count	0	47	9	56	48	8	56	84	8	56
	1	3	41	44	7	37	44	5	39	44
خطأ التصنيف		12.0%			15.0%			13.0%		
الدالة التمييزية باستخدام الاختيار الامامي					الدالة التمييزية باستخدام الحذف العكسي			الدالة التمييزية باستخدام الانحدار المتدرج		
y		Predicted Group Membership		المجموع	Predicted Group Membership		المجموع	Predicted Group Membership		المجموع
		0	1		0	1		0	1	
Count	0	48	8	56	45	8	56	48	8	56
	1	6	38	44	4	40	44	6	38	44
خطأ التصنيف		14.0%			15.0%			14.0%		

الاستنتاجات:

1. طريقة الاختيار الأمامي مشابهة تماما لطريقة الاختيار المتدرج في تحليل الانحدار والتحليل التمييزي.
2. لا يوجد اختلاف كبير بين طريقة حدود الثقة لروي–بوس وبين المكونات الرئيسية حيث ان الطريقتين اثبتت معنوية المتغيرات($x4, x5, x6, x7, x8, x9, x10$) والاختلاف الوحيد هو اضافة المتغير x12 بطريقة المركبات الرئيسية.
3. لوحظ ان المتغيرات (x4,x5,x8) أثبتت معنويتها في جميع الطرائق المستخدمة وهذا دليل على ان هذه المتغيرات مهمة جداً وذات تأثير كبير. وان المتغيرات (x1,x2,x3) لم تظهر معنويتها في كل الطرائق وهذا يعني ان هذه المتغيرات غير مهمة .
4. نلاحظ ان جميع النتائج التصنيف متقاربة وقليلة نسبيا يعني ذلك كفاءة الطرائق المستخدمة.

Refference

1- Al-Hamdani, Basma Rasheed, 2014, "The Medical Staff Distinguished According to Their Knowledge of the International Classification (ICD-10) by Using the Distinctive Function," a master's thesis in statistics from the University of Baghdad, College of Administration and Economics.
2- Al-Rawi, Khasha'a Mahmoud, (1979), "The Introduction to Regression Analysis", printed by Dar Al-Kutub for Printing and Publishing, University of Mosul.
3- Al-Rawi, Khasha'a Mahmoud, (1987), "The Introduction to Regression Analysis", printed by Dar Al-Kutub for Printing and Publishing, University of Mosul.
4- Al-Rikabi, Abdel-Qader Kazem, (1988), "The Characteristic Analysis of the Results of Classroom Students' Examinations".
5- First in Al-Mustansiriya University", Master Thesis, Al-Mustansiriya University.
6- Al-Zobaie, Obaid Mahmoud and Al-Mashhadani, Kamal Alwan (1988), "A proposed statistical model for classifying students into the scientific and literary branches," a research published in the Journal of Economic Sciences - Issue Two.
7- Al-Shukrji, Thanon Younes and Al-Nuaimi, Aswan Muhammad (2007), "Construction of the Discrimination Function Depending on the Variables of Regression Analysis," a research published in the Tikrit Journal of Administrative and Economic Sciences - Seventh Issue.
8- Al-Ezzi, Muhammad Shakir Mahmoud (2017), "Comparison of the classification process with the method of the linear characteristic function and logistic regression in the presence of the problem of polylinearity with the application of" master's thesis, University of Baghdad.
9- Al-Habil, Abdullah and Al-Jazzar, Majed, 2013, "A comparison between the linear characteristic function and multiple logistic regression", An-Najah University, Education Journal, Volume 28, Number (6), pp. 1525-1548.
10- Al-Naimi, Aswan Muhammad Tayeb, (2005), "Choice of Variables in Letter Decline", unpublished master's thesis, College of Computer Science and Mathematics, University of Mosul.
11- Hamoudat, Alaa Abdel Sattar (2005), "The discriminatory function and methods of determining its variables," master's thesis, University of Mosul.
12- Dabdoub, Marwan Abdel-Aziz, (1991), "Using Rows and Columns Matrixes in Analyzing Principal Components," Tanmiat Al-Rafidain Journal, University of Mosul, No. 49, Volume VIII.
13- Akkar, Ahmed Abd Ali, 2008, "Classification of Certain Types of Dates Using Distinctive Analysis," Journal of Administration and Economics, Issue Sixty-eighth, pp. 96-105.
14- Kazem, Amari Hadi and Al-Dulaimi, Muhammad Munajid, (1988), "Introduction to Linear Regression Analysis", Dar Al-Kutub for Printing and Publishing, University of Mosul.
15- Anderson, T.W. (1985), An Introduction to "Multivariate Statistical Analysis", John Wiley, New York, London.
16- Morrison, D.F. (1976)," Multivariate Statistical Methods", McGraw-Hill Kogakusha, LTD, Pennsylvania.
17- Zhang, M.Q., (2000), "Discriminant Analysis and its Application in DNA Sequence Motif Recognition",
18- Demosthenes B. Panagiotakos, 2006 ," A comparison between Logistic Regression and Linear Discriminant Analysis for the Prediction of Categorical Health Outcomes", International Journal of Statistical Sciences, Number 5, pp (73-84).
19- Aguilera, A., Escabias, M., . Valderrama, M.(2006)' Using principal components for estimating logistic regression with high- dimensional multicollinear data Computational Statistics & Data Analysis 50 (2006) 1905– 1924 www.elsevier.com/locate/csda

Building Discriminate Function-Review-

Nada Nizar Muhammad Naglaa Saad Ibrahim Zaid Tariq Saleh
Department of Statistics and Informatics, College of Computer science and Mathematics, University of Mosul, Mosul, Iraq

Abstract : Discriminant Analysis has been widely used to classify data into subgroups based on certain criteria. The classification process depends on choosing any variable that shows a statistical significance, then use the selected variables to build the discriminant function. In order to investigate the statistical significance of the variables in our data, we used Roy-Bose procedure for finding confidence intervals and t-test, which is one of the popular variable-selection methods in discriminant analysis. In addition, some other variable-selection techniques has been employed, namely, Forward-Selection, Backward-Selection, and Stepwise-Selection methods, which are usually used to select variables in linear regression analysis. Furthermore, a principal component analysis has been carried out for the purpose of choosing the variables with high statistical significance. The selected variables have been used to build the discriminant function.
Keyword: discrimination process,Roy-Bose boundary confidence method ,T-test method ,forward check method ,Reverse elimination method ,Stepwise regression method