المركز الجامعي عبد الحفيظ بوالصوف – ميلة

المحور الخامس

التحليل التمييزي The discriminate ate analysis

1/ مقدمة:

التحليل التمييزي هو أحد طرق البحث العلمي الذي يستخدم في تحليل وتفسير الظواهر الاجتماعية والصحية والاقتصادية وغيرها، ويهتم بكيفية التمييز بين مجموعتين أو أكثر من الأفراد أو الأشياء وتصنيف المفردات الجديدة في هذه المجموعات. ويعتبر التحليل التمييزي استكشافيا، ويسعى إلى التمييز بين المشاهدات باستخدامه العديد من المتغيرات، وذلك من خلال إيجاد تركيبات خطية لمجموعة من المتغيرات يطلق عليها إسم: "متغيرات التمايز".

2/ أهمية وأهداف التحليل التمييزي

تنبع أهمية التحليل التمييزي من مقدرته على التمييز بين مجموعتين أو أكثر من خلال مجموعة من المتغيرات، وذلك بإنشاء (بناء) دوال تمييزية تعمل على تعظيم الاختلاف (الفروق) بين المجموعات بأقل خطأ للتصنيف.

يهدف التحليل التمييزي إلى تحقيق مجموعة من الغايات أهمها:

- بناء دوال تمييزية للفصل أو التمييز بين فئات المتغير التابع النوعي؛
- تعمل هذه الدوال على تعظيم الفروق بين المجموعات (فئات المتغير التابع)؛
- ترتيب المتغيرات التي تسهم بقدر كبير في التمييز بين المجموعات (الفئات)؛
- تصنيف المشاهدات الجديدة وتوزيعها على المجموعات (الفئات)؛
- الوصول إلى أقل نسبة خطأ للتوصيف- تقديم دقة التصنيف كنسبة مئوية.

3/ شروط التحليل التمييزي:

1- عدم تساوي متوسطات المجموعات (فئات المتغير التابع).
2- تساوي مصفوفة التباين -التغاير بين المجموعتين.
3- أن تكون المجموعات منفصلة وقابلة للتحديد.
4- أن تتوزع المتغيرات التابعة والكمية توزيعا طبيعيا.
5- يتم اختيار العينة عشوائيا.
6- استقلالية المشاهدات، أي عدم وجود تعدد خطي بين متغيرات الدراسة.
7- عدم وجود قيمة متطرفة لكون التحليل التمييزي أكثر حساسية وتأثرا بالقيم الشاذة.

4/ الدالة التمييزية (Discriminant Function)

يقوم التحليل التمييزي على فكرة بناء دالة تمييز (Discriminant Function) تعمل على تعظيم الفروق بين متوسطات المجموعات وتدنية أخطاء التصنيف في الوقت ذاته. وهو أسلوب احصائي بتناسب مع

حالة **"المتغير التابع النوعي، والمتغيرات المستقلة الكمية"**، وتهدف إلى تسهيل تصنيف أي مشاهدة جديدة إلى إحدى المجموعات بأقل خطأ للتصنيف، كما تعمل على استبعاد المتغيرات التي ليس لها تأثير معنوي في التمييز والفصل بين المجموعات.

يمكن صياغة التوليفة الخطية لهذه الدالة التمييزية على النحو التالي:

$$L = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k$$

حيث: Z = الدالة التمييزية Discriminate Score

α_i = الدرجة المميزة Discriminate Weight

x_i =المتغيرات المستقلة Independent Variables

K = عدد المتغيرات المستقلة في النموذج

كيفية بناء الدالة التمييزية:

في حالة تعدد المجموعات تتعدد الدوال التمييزية، لذا سنقتصر عملنا هذا على الدالة التمييزية بين مجموعتين فقط.

أولا– حساب متوسطات المتغيرات في كل مجموعة وإيجاد الفرق بين المتوسطات

متوسطات متغيرات مج1	متوسطات متغيرات مج2	الفرق بين متوسطات متغيرات المجموعتين
$\bar{x}_{i(1)} = \begin{pmatrix} \bar{x}_{1(1)} \\ \bar{x}_{2(1)} \\ \vdots \\ \bar{x}_{k:(1)} \end{pmatrix}$	$\bar{x}_{i(2)} = \begin{pmatrix} \bar{x}_{1(2)} \\ \bar{x}_{2(2)} \\ \vdots \\ \bar{x}_{k(2)} \end{pmatrix}$	$d_i = \bar{x}_{i(1)} - \bar{x}_{i(2)} = \begin{pmatrix} \bar{x}_{1(1)} - \bar{x}_{1(2)} \\ \bar{x}_{2(1)} - \bar{x}_{2(2)} \\ \vdots \\ \bar{x}_{k(1)} - \bar{x}_{k(2)} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{pmatrix}$

ثانيا- إيجاد التباينات المشتركة والتغايرات المشتركة بين المجموعتين

التباينات	التغايرات
$S_{ii} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$	$S_{ij} = \sum x_i x_j - \frac{\sum x_i \sum x_j}{n}$
التباينات المشتركة	**التغايرات المشتركة**
$v_{ii} = \frac{S_{ii(1)} + S_{ii(2)}}{n_1 + n_2 - 2}$	$v_{ij} = \frac{S_{ij(1)} + S_{ij(2)}}{n_1 + n_2 - 2}$

ثالثا- مصفوفة التباين المشترك-التغاير المشترك بين المجموعتين: $V = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kk} \end{pmatrix}$

وهي مصفوفة مربعة متناظرة، يمثل قطرها الرئيسي التباين المشترك وتمثل باقي العناصر التغاير المشترك.

رابعا- صيغة الدالة التمييزية

تأخذ الدالة التمييزية بمعاملات معيارية الشكل التالي: $L = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k$

حيث: $\alpha = V^{-1}.d \Leftrightarrow \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kk} \end{pmatrix}^{-1} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{pmatrix}$

α_i = معاملات الدالة التمييزية المعيارية

d_i = مصفوفة المسافة بين متوسطات المتغيرات في المجموعتين.

V^{-1} = مصفوفة التباين المشترك- التغاير المشترك.

5/ الأهمية النسبية للعوامل المؤثرة (للمتغيرات المستقلة)

بعد عملية بناء الدالة التمييزية، ينبغي تحديد الأهمية النسبية للمتغيرات المستقلة المؤثرة في عملية التمييز والفصل بين المجموعات وترتيبها، بعد استبعاد إشارات المعاملات المعيارية لدالة التمييز، وصاحب أعلى قيمة هو الأكثر أهمية.

أما عن نسبة المساهمة في عملية التمييز، فتحدد من خلال معامل الارتباط القانوني " Canonical Correlation".

6/ اختبارات الدالة التمييزية

لاختبار قدرة الدالة على التمييز والفصل بين المجموعات نستخدم الاختبارات الآتية:

أولا- اختبار F ($F\ test$)

ويعتمد هذا الاختبار على قياس الاختلافات بين المجموعات وداخل المجموعات بين المفردات من خلال جدول تحليل التباين التالي:

Source	SS	Df	MS	F
Between x's بين المجموعات	SSB	k-1	MSB	$\frac{MSB}{MSE}$
Within x's داخل المجموعات (الخطأ)	SSE	n-k	MSE	
Total الكلي	SST	n-1	/	/

حيث:

مجموع مربعات الأخطاء: $SSE = D^2 = \alpha_1 d_1 + \alpha_2 d_2 + \cdots + \alpha_k d_k$

مجموع بين المجموعات: $SSB = \frac{n_1.n_2}{(n_1+n_2)(n_1+n_2-2)} \times (D^2)^2$

مجموع المربعات الكلي: $SST = SSB + SSE$

ويتم الاختبار كالآتي:

1- صياغة الفرضيات:
الفرضية الصفرية H_0 ليس للدالة القدرة على التمييز عند مستوى دلالة معين.
الفرضية البديلة H_1 للدالة القدرة على التمييز عند مستوى الدلالة المعين.
2- القيمة المحسوبة: $F_c = \frac{MSB}{MSE}$
3- القيمة الجدولية: $F_t(k-1, n-k)$
4- القرار: إذا كانت $(F_c > F_t)$ نرفض الفرضية الصفرية ونقبل الفرضية البديلة لها التي تنص على أن للدالة قدرة تمييزية بين المجموعات، والعكس صحيح.

ثانيا- اختبار ويلكس لمدا (Λ) Wilks Lambda

1- صياغة الفرضيات

تأخذ الفرضيات الإحصائية الشكل التالي:

الفرضية الصفرية $H_0 : \mu_1 = \mu_2$ ليس للدالة القدرة على التمييز عند مستوى دلالة معين.
الفرضية البديلة $H_1 : \mu_1 \neq \mu_2$ للدالة القدرة على التمييز عند مستوى الدلالة المعين.

2- القيمة المحسوبة: $\boldsymbol{\Lambda = \prod_{i=1}^{k} \left(\frac{1}{1+\lambda_i}\right) = \frac{1}{1+\lambda_1} \cdot \frac{1}{1+\lambda_2} \ldots\ldots \frac{1}{1+\lambda_k}}$

λ_i الجذر الكامن لكل المتغيرات، k عدد المتغيرات

3- **القرار**: تنحصر قيمة $0 < \boldsymbol{\Lambda} < 1$
إذا كان: $\boldsymbol{\Lambda} = 1$ معناها تساوي متوسطات المجموعتين، وبالتالي عدم مقدرة الدالة على التمييز والفصل.
إذا كان: $\boldsymbol{\Lambda} = 0$ معناها عدم تساوي متوسطات المجموعتين، وبالتالي مقدرة الدالة على التمييز والفصل.
إذا اقتربت قيمة $\boldsymbol{\Lambda}$ **من الواحد دليل على عدم قدرة الدالة على التمييز، وإذا اقتربت من الصفر دليل على قدرة الدالة على التمييز.**

تستخدم إحصائية "ويلكس لمدا" لاختبار معنوية المتغيرات الداخلة في النموذج حيث يتم الإبقاء على المتغيرات التي لها أدنى قيمة الإحصائية "Wilk's lambda" وأعلى قيمة لـ: F.

ثالثا– اختبار هوتلنج Hotelling-Lawely test (T^2)

1– إحصاءة هوتلنج تأخذ الشكل التالي: $\boldsymbol{T^2 = \sum_{i=1}^{s} \lambda_i}$
حيث: λ_i هي الجذور الكامنة للمتغيرات
S: عدد المتغيرات

2- وتعادل لإحصائية هوتلنج قيمة F من جدول تحليل التباين، ويمكن تحويله إلى قيمة لها توزيع F تقريبي صيغته كالتالي: $F = \frac{n_1+n_2-k-1}{(n_1+n_2-2)k} * T^2$

3- **القيمة الجدولية:** $F_\alpha(k-1, n_1+n_2-k-1)$

4- **القرار**: إذا كانت F المحسوبة أكبر من F المجدولة، نرفض الفرضية الصفرية ونقبل الفرضية البديلة التي تنص على أن للدالة قدرة عالية على التمييز بين المجموعتين.

7/ نقطة الفصل (القطع) Cut of Point

بعد بناء الدالة التمييزية واختبار قدرتها على التمييز والفصل بين المجموعتين، يبدأ الاستخدام الثاني لها وهو كيفية تصنيف المشاهدة الجديدة إلى أي المجموعتين تنتمي، ويتم ذلك من خلال الخطوات الآتية:

أولا- تحديد نقطة الفصل وهي تمثل متوسط المتوسطين: $\bar{L} = \frac{\bar{L}_{(1)}+\bar{L}_{(2)}}{2}$ **حيث:**

$\bar{L}$ هي نقطة الفصل، $\bar{L}_{(1)}$ **:** تمثل متوسط قيم المجموعة الأولى، $\bar{L}_{(2)}$ **:** تمثل متوسط قيم المجموعة الثانية.

ثانيا- قاعدة التصنيف Classification Role

من خلال هذه القاعدة، يمكن تصنيف أو التنبؤ بانتماء مفرجة جديدة لإحدى المجموعتين بأقل خطأ تصنيف على النحو التالي:

1- إذا كان $\bar{L}_{(1)} > \bar{L}_{(2)}$

- وإذا كانت القيمة التمييزية للمفردة الجديدة أكبر من نقطة الفصل، تصنف ضمن المجموعة الأولى.
- وإذا كانت القيمة التمييزية للمفردة الجديدة أصغر من نقطة الفصل، تصنف ضمن المجموعة الثانية.
- وإذا ساوت نقطة الفصل، تصنف عشوائيا ضمن أي مجموعة من المجموعتين.

2- إذا كان $\bar{L}_{(1)} < \bar{L}_{(2)}$

- وإذا كانت القيمة التمييزية للمفردة الجديدة أكبر من نقطة الفصل، تصنف ضمن المجموعة الثانية.
- وإذا كانت القيمة التمييزية للمفردة الجديدة أصغر من نقطة الفصل، تصنف ضمن المجموعة الأولى.
- وإذا ساوت نقطة الفصل، تصنف عشوائيا ضمن أي مجموعة من المجموعتين.

8/ أخطاء التصنيف

يقصد بأخطاء التصنيف، وضع المفردة في مجموعة غير مناسبة لها، ويعتبر خطأ التصنيف عامل مهم عند الحكم على كفاءة الدالة التمييزية. وهناك نوعان من أخطاء التصنيف هما:

أولا- خطأ التصنيف الظاهري: يتم حسابه من خلال جدول التصنيف التالي:

المجموعة	تابع المجموعة الأولى (1)	تابع المجموعة الثانية (2)	المجموع
الأولى (1)	n_{11}	n_{12}	n_1
الثانية (2)	n_{21}	n_{22}	n_2

$\boldsymbol{n_{11}}$: عدد المفردات من المجموعة الأولى التي تم تصنيفها في نفس المجموعة، وبالتالي فقد صنفت بطريقة صحيحة.

$\boldsymbol{n_{12}}$: عدد المفردات من المجموعة الأولى التي تم تصنيفها في المجموعة الثانية، وبالتالي فقد صنفت بطريقة خاطئة.

$\boldsymbol{n_{21}}$: عدد المفردات من المجموعة الثانية التي تم تصنيفها في المجموعة الأولى، وبالتالي فقد صنفت بطريقة خاطئة.

$\boldsymbol{n_{22}}$: عدد المفردات من المجموعة الثانية التي تم تصنيفها في نفس المجموعة، وبالتالي فقد صنفت بطريقة صحيحة.

ويحسب الخطأ الظاهري كما يلي:

$\boldsymbol{P_{12} = \frac{n_{12}}{n_1}}$: وهي نسبة المفردات التي تنتمي إلى المجموعة الأولى وصنفت خطأ في المجموعة الثانية.

$\boldsymbol{P_{21} = \frac{n_{12}}{n_1}}$: وهي نسبة المفردات التي تنتمي إلى المجموعة الثانية وصنفت خطأ في المجموعة الأولى.

ويمكن حساب معدل الخطأ الظاهري باستخدام المعادلة التالية: $\boldsymbol{\frac{n_{12}+n_{21}}{n_1+n_2}}$.

ثانيا- الخطأ الحقيقي: يمثل نسبة التصنيف الخاطئ في المجتمع: $\boldsymbol{P_{12} = P_{21} = F\left[\frac{-\sqrt{D^2}}{2}\right]}$

حيث: F هي دالة التوزيع الطبيعي المعياري، D هي إحصائية "Mahanobis". تحسب القيمة بين قوسين ويحسب الاحتمال المقابل لها من جدول التوزيع الطبيعي المعياري.

القرار: كلما اقترب الاحتمال من الصفر، دل على انخفاض خطأ التوصيف، وبالتالي قدرة الدالة على التمييز والتصنيف. أما إذا كان الاحتمال قريبا من الواحد، دل على ارتفاع خطأ التوصيف وانخفاض قدرة الدالة على التمييز والتصنيف.

9/ الدالة التمييزية بمعاملات غير معيارية

تأخذ هذه الدالة الشكل التالي: $\boldsymbol{y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k}$

y : الدالة التمييزية غير المعيارية $\quad$ b_0 : ثابت التمييز

b_n : معاملات دالة التمييز غير المعيارية $\quad$ x_n : المتغيرات غير المعيارية

وللحكم على جودة النموذج التمييزي من خلال: "معامل الارتباط القانوني" *Canonical Correlation*، حيث أن القيم المرتفعة لمعامل الارتباط القانوني تكون مؤشرا على جودة التوفيق العالي للنموذج التمييزي.

وبتربيع قيمة معامل الارتباط القانوني، نحصل على قيمة معامل التحديد "R^2" الذي يحدد نسبة مساهمة المتغيرات المستقلة في التمييز والتصنيف.

انتهى بعون الله