المركز الجامعي عبد الحفيظ بوالصوف-ميلة

دراسة حالة باستخدام طريقة التحليل التمييزي

Case study by discriminatory analysis method

تمهيد:

عندما يتعلق الأمر بعدد كبير من المتغيرات لابد من اللجوء إلى التحليل متعدد المتغيرات، والذي من أساليبه والشائع استخدامه أسلوب: " التحليل التمييزي"، ويتمثل دوره في التمييز وفصل مفردات العينة المدروسة إلى مجموعتين رئيستين.

I- دراسة حالة:

إليك عينة مكونة من 350 شخص منهم 168 مصاب بمرض السكري و182 غير مصاب بهدف الوصول إلى دالة تمييزية يتم من خلالها تصنيف الأشخاص أو المشاهدات الجديدة في إحدى المجموعتين بناء على فرضيات معينة.

متغيرات البحث

تتمثل في متغير تابع نوعي ثنائي القيمة (y) غير مصاب (0) ومصاب (1) ومجموعة من العوامل المؤثرة (المتغيرات المستقلة) وهي:

1- الوراثة (x_1) تأخذ $(0$ لا يوجد، 1 يوجد$)$
2- الوزن (x_2)
3- ضغط الدم (x_3) $(0$ طبيعي، 1 مرتفع$)$
4- العمر (x_4)
5- الجنس (x_5) تأخذ (0 أنثى، 1 ذكر)
6- التدخين (x_6) تأخذ (0 لا يدخن، 1 يدخن)
7- ممارسة الرياضة (x_7) تأخذ (0 لا يمارس، 1 يمارس)
8- مرض التفرس (x_8) تأخذ (0 لا يوجد، 1 يوجد)
9- الكوليسترول (x_9) تأخذ (0 لا يوجد، 1 يوجد)
10- الحالة الاجتماعية (x_{10}) تأخذ (0 أعزب، 1 متزوج)
11- أمراض القلب والكلى (x_{11}) تأخذ (لا يوجد، 1 يوجد

ولاستخدام التحليل التمييزي هناك مجموعة من الفرضيات لابد من توافرها

1/ اختبار التوزيع الطبيعي للبيانات

نظرا لكون حجم العينة يزيد عن 30 مفردة طبقا لنظرية "النهاية المركزية" فإن البيانات تتبع التوزيع الطبيعي، ولا داعي لإجراء هذا الاختبار.

2/ اختبار تساوي متوسطات المجموعتين

من خلال مخرجات البرنامج الجدول (1) SPSS الواردة في الجدول التالي:

الجدول (1) حساب متوسطات المجموعتين والدلالة الإحصائية

المتغير	المجموعة		Sig
	الأولى (مصابين)	الثانية (غير مصابين)	
	المتوسط	المتوسط	
(x_1)	0.7381	0.3352	.000
(x_2)	100.3274	73.7691	.000
(x_3)	0.7202	0.3187	.000
(x_4)	48.3333	38.2363	.000
(x_5)	0.6667	0.5385	.014
(x_6)	0.3512	0.6154	.000
(x_7)	0.3512	0.6644	.000
(x_8)	0.5476	0.5549	**.891**
(x_9)	0.7143	0.4121	.000
(x_{10})	0.7381	0.6099	.011
(x_{11})	0.5179	0.2802	.000

من خلال الجدول (1) نلاحظ أن القيمة الاحتمالية لكل المتغيرات هي أصغر من 0.05، وبالتالي نرفض الفرضية الصفرية H_0 ونقبل الفرضية البديلة لها التي تنص على أن الفرق بين متوسطي كل متغير في المجموعتين ذو دلالة إحصائية، ما عدا المتغير الثامن x_8.

كما نلاحظ أن المتوسطات الأعلى في المجموعة الأولى (المصابين) في كل من x_2، x_4، x_1، x_9، وx_3 وهي: الوزن، العمر، الوراثة، الكوليسترول، وضغط الدم، وذلك أمر بديهي فإن المصابين بمرض السكري نسبة الضغط العالي والوزن الزائد والعمر المتقدم. أما المتوسطات الأعلى في المجموعة الثانية (غير المصابين) تتمثل في x_2، x_4، x_6.

ومن خلال إحصائية ويلكس لامدا Wilks's Lambda distribution نحصل على الجدول التالي:

الجدول (2) اختبار Lambda'Wilks

Tests of Function (s)	Wilks'Lambda	Chi-square	df	Sig
1	.365	346.661	8	.000

تصاغ الفرضيات الإحصائية على النحو التالي:

الفرضية الصفرية: $\boldsymbol{H_0 : \mu_1 \neq \mu_2}$

الفرضية البديلة: $\boldsymbol{H_1 : \mu_1 = \mu_2}$

حيث أن القيمة الاحتمالية $Sig(.000)$لاختبار كاي تربيع أصغر من 0.05، إذن نرفض الفرضية الصفرية ونقبل الفرضية البديلة لها التي تنص على أن هناك اختلاف جوهري بين متوسطي المجموعتين عند مستوى الدلالة الإحصائية 0.05، كما أن إحصاءة ويلس لامدا تساوي 0.365، وهي تقترب من الصفر دليل على وجود اختلاف بين متوسطي المجموعتين، بمعنى أن الدالة التمييزية لها القدرة على التمييز وتصنيف المشاهدات.

3/ اختبار فرضية تجانس التباين بين المجموعتين

تصاغ الفرضيات الإحصائية على النحو التالي:

الفرضية الصفرية: $H_0 : \Sigma_1 \neq \Sigma_2$

الفرضية البديلة: $H_1 : \Sigma_1 = \Sigma_2$

وباستخدام اختبار Box's M كانت النتائج كما يلي:

الجدول (3) Log Determinants

y	Rank	Log Determinants
N	8	1.230
Y	8	.976
Pooled within-Groups	8	1.244

الجدول (4) Test Results

Box² s M	47.341
F Approx.	1.283
df1	36
df2	402161.071
Sig	0.119

من خلال الجدول (4) نلاحظ أن القيمة الاحتمالية (0.119) Sig أكبر من 0.05، وبالتالي نقبل الفرضية الصفرية التي تنص على تساوي التباين التغاير للمجموعتين، بمعنى تحقق تجانس التباين بين المجموعتين.

4/ اختبار معنوية العوامل المؤثرة (المتغيرات المستقلة)

تم اختبار جميع العوامل المؤثرة في النموذج التمييزي لمعرفة أهمية كل متغير ومدى إسهامه في عملية التمييز والتصنيف فكانت كالآتي:

الجدول (5) Tests of Equality Group of Means

	Wilks'Lambda	F	df1	df2	Sig
(x_1)	.857	58.046	1	348	.000
(x_2)	.477	381.840	1	348	.000
(x_3)	.839	66.816	1	348	.000
(x_4)	.873	50.479	1	348	.000
(x_5)	.983	6.052	1	348	.014
(x_6)	.930	26.082	1	348	.000
(x_7)	.936	23.829	1	348	.000
(x_8)	1.000	0.019	1	348	.891
(x_9)	.908	35.422	1	348	.000
(x_{10})	.981	6.594	1	348	.011
(x_{11})	.941	21.827	1	348	.000

من خلال الجدول (5) نلاحظ أن القيمة الاحتمالية لكل المتغيرات المفسرة Sig أصغر من مستوى الدلالة 0.05 ما عدا المتغير الثامن، وهو ما يدل على أن المتغيرات لها تأثير معنوي كبير في عملية التمييز بين المجموعتين. ومن ثم تم التوصل إلى نموذج التحليل التمييزي المناسب لبيانات مرضى السكري.

// – بناء الدالة التمييزية

أولا– نحدد المتغيرات الداخلة في تكوين الدالة التمييزية، إذ أن اختزال عدد المتغيرات في نموذج التمييز يفيد في قياس المتغيرات ذات العلاقة المعنوية وذات الصلة الأكبر بالموضوع محل الدراسة.

ثانيا– هناك عدة معايير إحصائية للإدخال والحذف وهي: "الإبقاء على التمييز صاحب القيمة الأكبر لـ: F، والأقل قيمة لإحصاءة ويلكس لامدا كما هو في الجدول التالي:

الجدول (6) Variables Entered Removed

Step	Entered	Wilks'Lambda							
						Exact F			
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig
1	(x_2)	.477	1	1	348.000	381.840	1	348.000	.000
2	(x_4)	.441	2	1	348.000	219.946	2	347.000	.000
3	(x_3)	.417	3	1	348.000	160.993	3	346.000	.000
4	(x_1)	.400	4	1	348.000	129.221	4	345.000	.000
5	(x_7)	.390	5	1	348.000	107.676	5	344.000	.000
6	(x_5)	.380	6	1	348.000	93.114	6	343.000	.000
7	(x_6)	.370	7	1	348.000	83.038	7	342.000	.000
8	(x_9)	.365	8	1	348.000	74.142	8	341.000	.000

المصدر: SPSS V23

من خلال الجدول (6) نلاحظ أنه تم استبعاد ثلاث متغيرات وتم الإبقاء على ثمان متغيرات التي لها قدرة أعلى التمييز والفصل بين مجموعتي المصابين وغير المصابين والتي لها أعلى قيمة لـ: F وأقل قيمة لـ: ويلكس لامدا. وتم الاختيار بناء على الاختبار التدريجي على 8 خطوات، وبالتالي المتغيرات الداخلة في النموذج هي: $x_2, x_4, x_3, x_1, x_7, x_5, x_6, x_9$.

ثالثا– إيجاد الدالة التمييزية

1– من الجدول رقم (1) تم إيجاد متوسطات المتغيرات في كلتا المجموعتين.

2– إيجاد الفرق بين متوسطات كل متغير من كلتا المجموعتين: $d_i = \bar{x}_{i(1)} - \bar{x}_{i(2)} = \begin{pmatrix} \bar{x}_{1(1)} - \bar{x}_{1(2)} \\ \bar{x}_{2(1)} - \bar{x}_{2(2)} \\ \vdots \\ \bar{x}_{k(1)} - \bar{x}_{k(2)} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{pmatrix}$

3– **مصفوفة التباين التغاير بين المجموعتين**

الجدول (6) مصفوفة التباين المشترك التغاير المشترك

V=

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_9
x_1	**0.284**	2.969	0.50	1.446	0.049	−0.41	−0.41	**0.57**
x_2		**337.47**	3.136	62.014	0.293	−1.640	−1.528	**2.410**
x_3			**0.251**	0.960	−0.007	−0.047	−0.024	**0.035**
x_4				**201.25**	−0.348	−1.571	−0.814	**0.916**
x_5					**0.241**	0.053	−0.013	**−0.014**
x_6						**0.251**	0.016	**−0.041**
x_7							**0.250**	**−0.035**
x_9								**0.247**

وبذلك، الدالة التمييزية بمعاملات معيارية هي: $L = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k$

حيث:

$$\alpha = V^{-1}.d = \begin{pmatrix} 0.160 \\ 0.797 \\ 0.267 \\ 0.295 \\ 0.268 \\ -0.203 \\ -0.187 \\ 0.154 \end{pmatrix}$$

زمنه: $\boldsymbol{L = 0.160x_1 + 0.797x_2 + 0.267x_3 + 0.295x_4 + 0.268x_5 - 0.203x_6 - 0.187x_7 + 0.154x_9}$

رابعا– الأهمية النسبية للعوامل المؤثرة ونسبة المساهمة في التمييز والتنبؤ في النموذج التمييزي

لتحديد أكثر العوامل تأثيرا على مستوى الإصابة ومساهمة العامل في التمييز والتصنيف نستخدم الجدول الآتي:

الجدول (7) الأهمية النسبية للعوامل المؤثرة ونسبة المساهمة في التمييز

المتغير	المعامل α_i	الأهمية النسبية (معامل الارتباط القانوني التمييزي)
x_2	0.797	0.794
x_4	0.295	0.289
x_5	0.268	0.10
x_3	0.267	0.332
x_6	-0.203	-0.208
x_7	-0.187	-0.198
x_1	0.16	0.310
x_9	0.154	0.242

لمعرفة أهم العوامل المؤثرة ننظر لعمود المعاملات المعيارية α_i، إذ أن القيمة المطلقة الكبيرة يقابلها العامل الأكثر أهمية في التأثير على الإصابة وتكون هذه الأهمية موجبة أو سالبة. وأن أكثر المتغيرات أهمية: "x_2 الوزن، x_4 العمر، x_5 الجنس، x_3 ضغط الدم، x_6 التدخين، x_7 ممارسة الرياضة، x_1 الوراثة، x_9 الكوليسترول.

أما الأهمية النسبية فإن (x_2) الوزن يساهم بنسبة أكبر في عملية تمييز المجموعتين 79.4%، يليه ضغط الدم 33.2%، ثم الوراثة 31%، ثم العمر 28.9%، والكوليسترول 24.2%، ثم التدخين 20.8%، بعدها ممارسة الرياضة 19.8%، وأخيرا الجنس 10%.

خامسا- اختبار قدرة الدالة على التمييز

1- اختبار فيشر باستخدام جدول تحليل التباين

- من خلال الفرضيات الآتية:

الفرضية الصفرية: الدالة ليس لها القدرة على التمييز H_0

الفرضية البديلة: الدالة لها القدرة على التمييز H_1

- إيجاد قيمة مجموع مربعات الأخطاء (داخل المجموعات)

$$SSE = \boldsymbol{\alpha_1 + \alpha_2 d_2 + \alpha_3 d_3 + \alpha_4 d_4 + \alpha_5 d_5 + \alpha_6 d_6 + \alpha_7 d_7 + \alpha_9 d_9}$$

$$SSE = [0.160 \quad 0.797 \quad 0.267 \quad 0.295 \quad 0.268 \quad -0.203 \quad -0.187 \quad 0.154].\begin{bmatrix} 0.4029 \\ 26.55 \\ 0.4015 \\ 10.09 \\ 0.1282 \\ -0.2642 \\ -0.2532 \\ 0.3022 \end{bmatrix} = 24.50$$

- مجموع مربعات بين المتغيرات:

$$SSB = \frac{n_1.n_2}{(n_1+n_2)(n_1+n_2-2)} \times (D^2)^2 = \frac{168 \times 182}{(168+182)(168+182-2)} \times 24.5^2 = 150.68$$

- مجموع المربعات الكلي: $SST = SSE + SSB = 24.5 + 150.68 = 175.18$

الجدول (8) جدول تحليل التباين

Source	SS	df	MS	F
Between x's بين المجموعات	150.68	7	21.525	300.5
Within x's داخل المجموعات (الخطأ)	24.5	342	0.072	
Total الكلي	175.18	349	/	/

القيمة المحسوبة: $F = 300.5$

القيمة الجدولية: $F_{0.05}(7, 342) = 1.40$

القرار: $F_{0.05}(7, 342) = 1.40 < F_c = 300.5$، إذن نرفض الفرضية الصفرية ونقبل الفرضية البديلة التي تنص على أن للدالة قدرة عالية على التمييز والفصل بين المجموعتين.

2- اختبار ويلكس لامدا:

تصاغ الفرضيات الإحصائية على النحو التالي:

الفرضية الصفرية: الدالة ليس لها القدرة على التمييز $H_0 ; \mu_1 \neq \mu_2$

الفرضية البديلة: الدالة لها القدرة على التمييز $H_1 ; \mu_1 = \mu_2$

من خلال الجدول (2) نلاحظ أن:

أولا– قيمة اختبار ويلكس لامدا تساوي: 0.365 وهي أقرب للصفر، وهو دليل على القدرة العالية للدالة على التمييز.

ثانيا– القيمة الاحتمالية (0.365) Sig وهي أصغر من 0.05، وبالتالي نرفض الفرضية الصفرية ونقبل الفرضية البديلة لها التي تنص على أن للدالة القدرة على التمييز والفصل بين المجموعتين.

الجدول (9) الجذر الكامن (القيمة الذاتية)

Function	Eigenvalue	%of variance	Cumulative %	Canonical Correlation
1	**1.739[a]**	**100.0**	**100.0**	**0.797**

تشير قيمة الجذر الكامن $\lambda = 1.739$ إلى أن نسبة التباين المفسر بين مجموعتي المصابين وغير المصابين والتي تعود للفروق بينها في النموذج التمييزي الوحيد وقيمة معامل الارتباط القانوني 0.797 معامل الارتباط بين مجموعة العوامل المؤثرة ونموذج التمييز الوحيد. وبتربيع هذه القيمة نحصل على 63.5%، وهذا يعني نسبة مساهمة العوامل المؤثرة في التباين والاختلاف في التمييز بين المجموعتين.

الاستخدام الثاني للنموذج التمييزي التصنيف: فكانت النتائج كما يلي:

الجدول (10) استخدام النموذج التمييزي في التصنيف

/	غير مصاب (0)	مصاب (1)	مجموع
غير مصاب (0)	**169**	**13**	**182**
مصاب (1)	**20**	**148**	**168**
90.60% النسبة الإجمالية			

النموذج التمييزي يتكون من ثمان متغيرات هي: $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_9$. قام بالتصنيف الصحيح 169 مفردة من غير المصابين بمرض السكري وتصنيف غير صحيح 13، أي نسبة تصنيف صحيح 92.9% في الحالات بشكل صحيح. أما بالنسبة للمصابين قام بالتصنيف الصحيح 148 مفردة وغير الصحيح 30 مفردة بنسبة إجمالية صحيحة 88.1% وأن الدقة الإجمالية للتصنيف هي: 90.60% بخطأ 9.4%.

نسبة الخطأ الظاهري:

المجموعة الأولى: $P_{12} = \frac{n_{12}}{n_1} = \frac{13}{182} = 0.0714$، المجموعة الثانية: $P_{21} = \frac{n_{21}}{n_2} = \frac{20}{168} = 0.119$

///- تقدير النموذج التمييزي بمعاملات غير معيارية

يتم تقدير النموذج التمييزي بمعاملات غير معيارية كما يلي:

المتغير	b
x_1	0.324
x_2	0.063
x_3	0.581
x_4	0.022
x_5	0.551
x_6	-0.420
x_7	-0.386
x_9	0.325
$(Constant)$	-6.974

$$\hat{y}_i = -6.974 + 0.324x_1 + 0.063x_2 + 0.581x_3 + 0.022x_4 + 0.551x_5 - 0.420x_6 - 0.386x_7 + 0.325x_9$$

وهذا النموذج فعال وقادر على التصنيف الصحيح للمرضى بنسبة **90.60%** وأقل خطأ تصنيف 9.4%.

تصنيف المشاهدات الجديدة

- حساب متوسطي المجموعتين: $\bar{y}_1 = 1.369$ ، $\bar{y}_2 = 1.263$
- **نقطة الفصل**: $\bar{y} = \frac{\bar{y}_1 - \bar{y}_2}{2} = \frac{1.369 - .1.263}{2} = 0.052$

إذن: $\bar{y}_1 > \bar{y}_2$

إذا كانت المفردة الجديدة أكبر من النقطة الفاصلة تصنف المفردة في المجموعة الأولى، وإذا كانت المفردة الجديدة أصغر من النقطة الفاصلة تصنف المفردة في المجموعة الثانية.

مثال (1) - إذا كانت هناك سيدة $(x_5 = 0)$ لديها مرض السكري وراثة $(x_1 = 1)$ والوزن $(x_2 = 100)$ والعمر $(x_4 = 70)$ ولديها ضغط الدم $(x_3 = 1)$ والتدخين $(x_6 = 1)$ ولا تمارس الرياضة $(x_7 = 0)$ وليس لديها كوليسترول $(x_9 = 0)$ فإن:

$$\hat{y}_i = -6.974 + 0.324(1) + 0.063(100) + 0.581(1) + 0.022(70) + 0.551(0) - 0.420(0) - 0.386(0) + 0.325(0)$$

$$\hat{y}_i = 1.771$$

إذن القيمة التمييزية للمشاهدة الجديدة أكبر من نقطة الفصل، وبالتالي تصنف ضمن المجموعة الأولى (المصابين) لمرض السكري.

<u>مثال (2)</u> – إذا كانت هناك رجل $(x_5 = 1)$ ليس لديه مرض السكري وراثي $(x_1 = 0)$ وزنه $(x_2 = 65)$ والعمر $(x_4 = 60)$ وضغط الدم عادي $(x_3 = 0)$ ويدخن $(x_6 = 1)$ ولا تمارس الرياضة $(x_7 = 0)$ ولديه كوليسترول عالي $(x_9 = 1)$ فإن:

$$\widehat{y}_i = -6.974 + 0.324(0) + 0.063(65) + 0.581(0) + 0.022(60) + 0.551(1) - 0.420(1) - 0.386(0) + 0.325(1)$$

$$\widehat{y}_i = -1.103$$

إذن القيمة التمييزية للمشاهدة الجديدة أصغر من نقطة الفصل، وبالتالي تصنف ضمن المجموعة الثانية (غير المصابين) لمرض السكري.