الفصل السادس

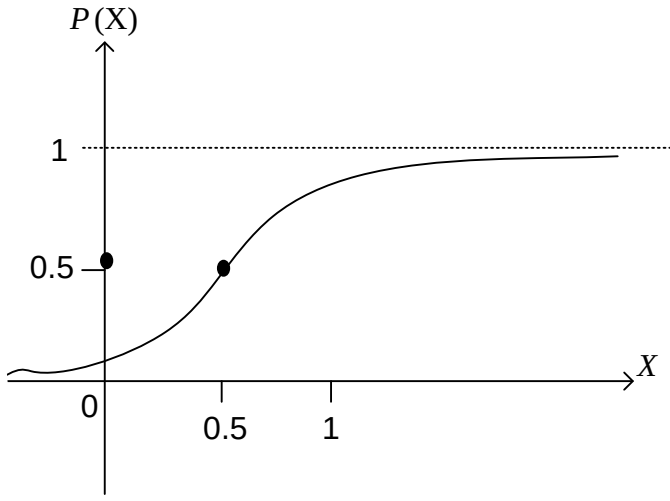
التحليل اللوجستي

6-1 تمهيد:

يهدف التحليل اللوجستي إلى تصنيف عناصر المجتمع المدروس إلى مجموعتين أو أكثر، وذلك باستخدام التوزيع الاحتمالي اللوجستي المعرف بالعلاقة التالية:

$$P(X) = \frac{1}{1 + \bar{e}^{(\beta_0 + \beta X)}} \qquad -\infty < X < \infty \qquad (1-6)$$

حيث X هو شعاع المتحولات المؤثرة في عمليات التصنيف و $P(x)$ هو الاحتمال المقابل له ويأخذ قيمه في المجال $[0,1]$، وهو يرسم في المستوى المنحني التالي :



الشكل (6-1) منحني التوزيع اللوجستي

وهناك نوعان للتحليل اللوجستي هما: الثنائي والمتعدد :

1- **التحليل اللوجستي الثنائي:** وفيه يشترط أن يكون التابع Y نوعياً ويأخذ حالتين متنافيتين (نجاح أو فشل، ربح أو خسارة، مدخن أو غير مدخن، حامل للمرض أو غير حامل له، محقق لشرط ما أو غير محقق له، ...الخ)، وأن يأخذ مقابل الحالة المرغوبة الأولى القيمة العددية (1) واحد، وأن يأخذ مقابل الحالة الثانية القيمة العددية (0) صفر. أما المتحولات X المؤثرة في التصنيف فيمكن أن تكون كمية أو نوعية أو مختلطة، وتأخذ قيمها ضمن مجالات أو فئات محددة، ولا يشترط عليها أن تحقق أية شروط مسبقة .

2- **التحليل اللوجستي المتعدد:** وفيه يشترط أن يكون التابع Y نوعياً ويأخذ عدة حالات متنافية (مستوى التعليم، حالة العمل، الحالة الاجتماعية، ...الخ) وأن يأخذ مقابل إحدى الفئات القيمة (1) ومقابل الفئات المتبقية القيمة (0) .

ولتوضيح الأساس الرياضي للنموذج اللوجستي الثنائي نعود إلى العلاقتين (3-32) و(3-33) من الفصل الثالث، اللتين تعطيانا المنطقتين R_1 و R_2 المقابلتين للمجموعتين G_1 و G_2 والاحتمالين السابقين P_1 و P_2 وللمتحولين الخاضعين للتوزيع الطبيعي X_1 و X_2 .

ولنفترض الآن أن تكاليف التصنيف الخاطئ متساوية $\left[\text{أي أن} C(1/2) = C(2/1)\right]$ فعندها نجد أن العلاقة (3-32) تأخذ الشكل التالي:

$$R_1: (\bar{X}_1 - \bar{X}_2)' * S_P^{-1} * X - \frac{1}{2}(\bar{X}_1 - \bar{X}_2)' S_P^{-1}(\bar{X}_1 - \bar{X}_2) \geq \ln\left(\frac{P_2}{P_1}\right) \qquad (2-6)$$

وبما أن: $\ln\left(\frac{P_2}{P_1}\right) = -\ln\left(\frac{P_1}{P_2}\right)$ فإنه يمكننا كتابة (6-2) كما يلي:

$$+\ln\left(\frac{P_1}{P_2}\right) \geq +\frac{1}{2}(\bar{X}_1 - \bar{X}_2)'\, S_P^{-1}(\bar{X}_1 - \bar{X}_2) - (\bar{X}_1 - \bar{X}_2) S_P^{-1} * X \qquad (3-6)$$

وبما أن $P_2 = 1 - P_1$ فإنه يمكننا كتابة (6-3) كما يلي:

$$\ln\left(\frac{P_1}{P_2}\right) = \ln\left(\frac{P_1}{1-P_1}\right) \geq \propto + \beta X \qquad (4-6)$$

حيث أن $\propto$ و B تقدران من بيانات العينة وحصراً من العلاقتين :

$$\tilde{\propto} = a = +\frac{1}{2}(\bar{X}_1 - \bar{X}_2)'\, S_P^{-1}(\bar{X}_1 - \bar{X}_2) \qquad (5-6)$$

$$\tilde{\beta} = b = -(\bar{X}_1 - \bar{X}_2) S_P^{-1} \qquad (6-6)$$

نعود إلى العلاقة (6-4) فنجد أنه يمكننا صياغتها كما يلي:

$$\frac{P_1}{1-P_1} \geq e^{a+bX} \qquad (7-6)$$

نقسم بسط ومقام الطرف الأيسر على P_1 الموجب فنجد أن:

$$\frac{1}{\frac{1}{P_1} - 1} \geq e^{a+bX}$$

ثم نأخذ مقلوب الطرفين فنجد أن:

$$\frac{1}{P_1} - 1 \leq \frac{1}{e^{a+bX}} = \bar{e}^{(a+bX)}$$

$$\frac{1}{P_1} \leq +1 + \bar{e}^{(a+bX)}$$

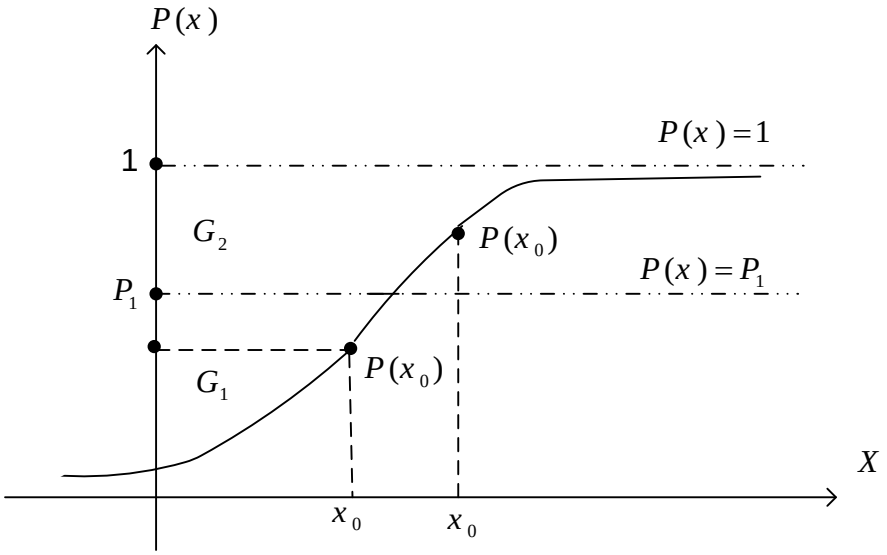
ثم نأخذ مقلوب الطرفين مرة أخرى فنجد أن:

$$\boxed{P_1 \geq \frac{1}{1+\bar{e}^{(a+bX)}} = P(x)} \qquad (8-6)$$

وهي صيغة النموذج اللوجستي الذي يأخذ قيمه المستمرة في المجال $[0,1]$، وهكذا تصبح القاعدة (3-32) على الشكل التالي :

إذا كان لدينا x_0 عنصراً جديداً من المجتمع فإننا نصنفه في G_1 إذا كانت قيمة التابع اللوجستي : $P(x_0) = \frac{1}{1+\bar{e}^{(a+bx_0)}}$ أصغر من قيمة الاحتمال السابق P_1 (وهو معلوم لأنه عبارة عن نسبة المجموعة G_1 في المجتمع) .

أما إذا كانت قيمة $P(x_0)$ أكبر من الاحتمال P_1، فإننا نصنف x_0 في المجموعة G_2، ويمكننا تمثيل ذلك بيانياً كما يلي:



الشكل (6–2): قاعدة التصنيف للنموذج اللوجستي

ومنه نلاحظ أنه إذا كانت قيمة التابع $P(x_0)$ أصغر من P_1 فإن x_0 ينتمي إلى G_1، وإذا كانت أكبر من P_1 فإن x_0 ينتمي إلى G_2 .

مثال (6–1): لنفترض أن دراسة شملت (20) طالباً، لمعرفة علاقة متوسط عدد ساعات الدراسة يومياً X مع نتيجة اجتيازهم للامتحان Y، الذي يأخذ القيمة (1) في حالة النجاح والقيمة (0) في حالة الرسوب . وكانت نتائج الاستبيان كما في الجدول التالي :

جدول (6–1): نتائج الاستبيان لعلاقة عدد الساعات بنتيجة الامتحان [Wikipediai.rg] :

i رقم الطالب	1	2	3	4	5	6	7	8	9	10
X: عدد الساعات	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Y: نتيجة الامتحان	0	0	0	0	0	0	1	0	1	0

يتبع

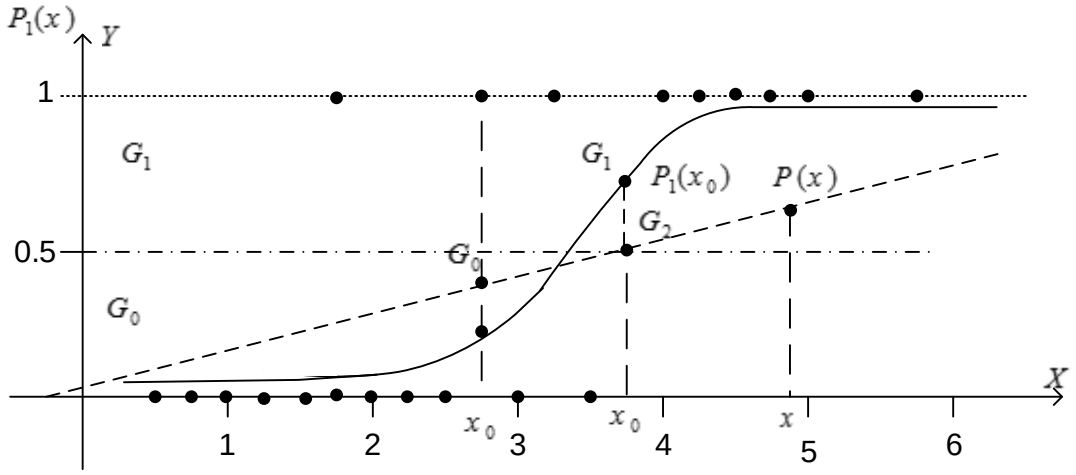
i رقم الطالب	11	12	13	14	15	16	17	18	19	20
X: عدد الساعات	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Y: نتيجة الامتحان	1	0	1	0	1	1	1	1	1	1

ونريد الآن معرفة مدى تأثير عدد ساعات الدراسة على احتمال النجاح .

ومن الجدول السابق نلاحظ أن عدد الناجحين $m = 10$، أي أن المعدل العام للنجاح $P = \frac{10}{20} = 0.50$، وبالمقابل نجد أن المعدل العام للرسوب $q = 0.50$.

وإذا أردنا رسم شكل الانتشار لهذه البيانات، فإننا نلاحظ أن المتحول المستقل X هو متحول مستمر ويأخذ قيمه في المجال $[0,6]$.

أما المتحول التابع Y فهو متحول منقطع وثنائي القيمة، فهو يأخذ القيمة $(Y=1)$ في حالة النجاح، ويأخذ القيمة (0) في حالة الرسوب. وإذا قمنا برسم النقاط (x_1,y_1) على المستوى نجد أن قيم Y الصفرية تتوضع على المحور $0X$ وتميل نحو الجانب الأيسر، أما قيم Y المساوية للواحد فتتوضع على المستقيم$(Y=1)$ وتميل نحو الجانب الأيمن . ويوجد بعض النقاط المتقابلة في المنطقة الوسطى كما هو مبين على الشكل التالي :



الشكل (6−3): شكل الانتشار للبيانات

والسؤال الآن كيف سنتعامل مع هذا الشكل العجيب؟

وللإجابة على هذا السؤال ندرس بعض خصائص الظواهر الثنائية من خلال بيانات المثال (5−1) السابق .

6−2 خواص الظواهر الثنائية :

إن الظواهر الثنائية هي عبارة عن توابع ثنائية تأخذ حالتين A و$\bar{A}$ فقط [موافق أو غير موافق، نعم أو لا، نجاح أو فشل، ربح أو خسارة، قبول أو رفض، ...الخ] . ولقد أصطلح على إعطاء التابع الثنائي Yقيمة الواحد (1) عندما تتحقق الحالة المرغوبة A، وقيمة الصفر (0) عندما تحقق الحالة غير المرغوبة $\bar{A}$ (عدم تحقق A). ولنفترض أنه عند إجراء $(n=100)$ تجربة على أية ظاهرة ثنائية كانت نتائج تلك التجارب التي تخضع لتوزيع (برنويللي) كما في بالجدول التالي :

جدول (6−2): مخطط جدولي لتوزيع (برنويللي) لـ 100 تجربة على Y

الحالة	$A=G_1$	$\bar{A}=G_0$	المجموع
قيمة التابع Y	1	0	----
احتمال التحقيق	p	q	$p+q=1$
عدد التكرارات المطلقة	n_1	n_0	$n_1+n_0=n$
توزع عدد التجارب	60	40	$100=n$

وعندما نكرر هذه التجربة n مرة سنحصل على عينة من قيم Yبحجم n ، وإن عناصرها تتوزع حسب الجدول (6-2) على مجموعتين هما:

- المجموعة G_1 : وهي مجموعة العناصر التي تقابل القيم$(Y = 1)$ [مجموعة الناجحين]، وتضم n_1 عنصراً، ويفترض أن يكون احتمال تحققها في كل تجربة يكون ثابتاً ويساوي p، وإن p يقدر من $\tilde{p} = \frac{n_1}{n} = \frac{60}{100}$.
- المجموعة G_0 : وهي مجموعة العناصر التي تقابل القيم $(Y = 0)$ [مجموعة الراسبين] وتضم n_0 عنصراً، ويفترض أن يكون احتمال تحققها في كل تجربة يكون ثابتاً ويساوي $q = 1 - p$. ويقدر من: $\tilde{q} = \frac{n_0}{n} = \frac{40}{100}$.

وبناءً على نتائج هذه التجارب يمكننا تعريف عدة مؤشرات تستخدم في التحليل اللوجستي أهمها الأرجحية $(odds)$.

- **مفهوم الأرجحية $(odds)$ وتعريفها:** يعود ظهور مفهوم الأرجحية إلى عمليات الرهان في الظواهر الثنائية (نجاح أو فشل) .

حيث يقال: إن إمكانية فوز اللاعب A تساوي n_1 مقابل n_0 للاعب $\bar{A}$. وإذا قام اللاعبان بإجراء $n = 100$ تجربة وفاز اللاعب A بـ $n_1 = 60$ تجربة وخسر $n_0 = 40$ تجربة منها، فإن تعريف الأرجحية لحادث فوز اللاعب A على اللاعب $\bar{A}$ يعطى بالعلاقة التالية:

$$odds(A) = \frac{n_1}{n_0} = \frac{\text{عدد مرات تحقق فوز } A}{\text{عدد مرات عدد فوز } \bar{A}} = \frac{60}{40} = \frac{3}{2} = \frac{1.5}{1} \qquad (9-6)$$

وعندها نقول أن إمكانية فوز A على $\bar{A}$ تساوي 60 مقابل 40 . وهنا يفضل اختصار الكسر $\frac{60}{40}$ إلى آخر عددين صحيحين مثل $\left(\frac{3}{2}\right)$، ونقول إن إمكانية فوز A على $\bar{A}$ تساوي 3 مقابل 2 . وإنه من الأفضل تحويل الكسر الأخير إلى نسبة عدد إلى الواحد مثل $\left(\frac{1.5}{1}\right)$، ونقول أن فوز A على $\bar{A}$ تساوي 1.5 مقابل 1 . ونكتب ذلك على الشكل 1.5 : 1 . (ويكتب بالعكس بالنسبة لـ $\bar{A}$ *1,5:1*) .

- وبطريقة مشابهة نعرف الأرجحية لفوز اللاعب $\bar{A}$ بالعلاقة :

$$odds(\bar{A}) = \frac{n_0}{n_1} = \frac{\text{عدد مرات تحقق فوز } \bar{A}}{\text{عدد مرات عدد فوز } A} = \frac{40}{60} = \frac{2}{3} = \frac{1}{1.5} \qquad (10-6)$$

- ومن التعريفين السابقين نستنتج أن:

$$odds(\bar{A}) = \frac{1}{odds(A)} \qquad (11-6)$$

$$odds(A) * odds(\bar{A}) = 1 \qquad (12-6)$$

- تعريف احتمال تحقق فوز اللاعب (A): ويعرف بالعلاقة التالية :

$$P(A) = \frac{n_1}{n_1 + n_0} = \frac{n_1}{n} = \tilde{p} = \frac{60}{100} = 0.60 = \frac{1.5}{1 + 1.5} \qquad (13-6)$$

- احتمال تحقق فوز اللاعب $(\bar{A})$: ويعرف بالعلاقة التالية :

$$P(\bar{A}) = \frac{n_1}{n_1 + n_0} = \frac{n_0}{n} = q = \frac{40}{100} = 0.40 = \frac{1}{1 + 1.5} \qquad (14 - 6)$$

ومن العلاقتين (6-13) و(6-14) نستخلص أن:

$$P(A) + P(\bar{A}) = p + q = 1 \qquad (15 - 6)$$

كما يمكننا استخلاص العلاقة التي ترتبط بين الأرجحية واحتمال تحقق حالتها، حيث نجد أنه يمكننا كتابة العلاقة (6-9) كما يلي:

$$odds(A) = \frac{n_1}{n_0} = \frac{\frac{n_1}{n}}{\frac{n_0}{n}} = \frac{P(A)}{P(\bar{A})} = \frac{p}{q} = \frac{p}{1 - p} \qquad (16 - 6)$$

وكذلك نجد أن:

$$odds(\bar{A}) = \frac{n_0}{n_1} = \frac{\frac{n_0}{n}}{\frac{n_1}{n}} = \frac{P(\bar{A})}{P(A)} = \frac{q}{p} = \frac{1 - p}{p} \qquad (17 - 6)$$

وسنستخدم العلاقة (6-16) في عمليات استخراج التابع اللوجستي [انظر الفقرة (6-7) في آخر هذا الفصل حول مفهوم الأرجحية وعلاقتها باحتمالات الظواهر الثنائية].

6-3 استخراج النموذج المنطقي :

لقد رأينا أن التابع Y (نتيجة الطالب في المثال (6-1)) هو تابع ثنائي ويأخذ إحدى القيمتين (1) للنجاح و(0) للرسوب، ومن شكل الانتشار (6-2) نلاحظ أن هذا التابع Y لا يصلح من وجهة نظر نظرية الانحدار، لأن يكون نتيجة لأي تركيب خطي (أو غير خطي) للمتحول المستقل X. لذلك يجب البحث عن بديل للتابع Y مرتبط به ويعبر عنه ويوصلنا معه .

ومن جهة أخرى نجد أن الاحتمال الشرطي لأن يأخذ التابع Y القيمة $(Y = 1)$، عند قيمة معطية x يساوي احتمال أن ينتمي العنصر المعلوم x إلى المجموعة G_1، ونكتب ذلك على الشكل التالي:

$$P(G_1/x) = P(Y = 1/x) = P_1(x) = Y \text{ الاحتمال اللاحق لـ} \qquad (18 - 6)$$

وهو احتمال النجاح عند أية قيمة معطية x، وهو تابع مستمر ويأخذ قيمه في المجال $[0\ ,\ 1]$، وهو يصلح لأن يكون بديلاً عن Y، لأنه أصبح من الممكن رياضياً دراسة علاقة $P_1(x)$ مع المتحول المستقل x .

وإذا استطعنا أن نجد العلاقة بين هذا الاحتمال $P_1(x)$ والمتحول X، فإننا نكون قد تجاوزنا المشكلة، التي واجهتنا أثناء تمثيل Y عبر X .

وهكذا نجد أنه يجب علينا الآن أن نقوم بإيجاد قيم $P_1(x)$ المقابلة لجميع قيم X، حتى نستطيع أن نقابلها مع قيم X، ثم استخلاص علاقة الانحدار بينهما دون وضع شروط مسبقة على المتحول X .

لذلك نقوم بحساب قيم الاحتمالات $P_1(x)$ اللاحقة من علاقات (بايز) (2-33) التي تأخذ الشكل التالي:

$$P_1(x) = P(G_1/x) = \frac{P * \not{f}(x/G_1)}{P * \not{f}(x/G_1) + q * \not{f}(x/G_0)} \qquad (19-56)$$

حيث أن: $\not{f}(x/G_1)$ و $\not{f}(x/G_0)$ هما التوزيعان التجريبيان لـ X ضمن المجموعتين G_1 و G_0 على الترتيب، وهما يحسبان (يعد تكرار التجربة n مرة) من التكرارات النسبية المقابلة لقيم X المختلفة كمايلي:

$$\not{f}(x/G_1) = \frac{n_1(x)}{n} \qquad (20-6)$$

$$\not{f}(x/G_0) = \frac{n_0(x)}{n} \qquad (21-6)$$

حيث أن: $n_1 + n_0 = n$

وأن: $n_1(x)$ هو عدد تكرار مرات النجاح مقابل القيمة (x) .

وأن: $n_0(x)$ هو عدد تكرار مرات الرسوب مقابل القيمة (x) .

وبعدها يمكننا أن نفترض أن العلاقة بين $P_1(x)$ وX هي علاقة انحدار خطية من الشكل التالي :

$$\tilde{P}_1(x) = \propto + \beta x \qquad (22-6)$$

ثم نقوم بحساب تقدير لـ $\propto$ و β بطريقة المربعات الصغرى أو بطريقة الإمكانية العظمى، فنحصل على مستقيم محدد يفصل بين المجموعتين G_1 و G_0. كما هو مبين على الشكل (6-3) السابق.

ومنه نحسب القيم النظرية للاحتمالات اللاحقة $\tilde{P}_1(x)$ الواقعة على ذلك المستقيم مقابل كل قيمة لـ X .

ثم نقوم بحساب الاحتمالات اللاحقة المتممة له: $\tilde{P}_0(x)$ من العلاقة :

$$\tilde{P}_0(x) = 1 - \tilde{P}_1(x) \qquad (23-5)$$

وأخيراً نقوم بمقارنة $\tilde{P}_1(x)$ مع $\tilde{P}_0(x)$ ونصنف أي عنصر جديد x وفق القاعدة التالية:

$(24-5)$ إذا $P_1(x) \geq P_0(x)$ نصنف x في المجموعة G_1 (في مجموعة الناجحين)

وإذا كان $P_1(x) < P_0(x)$ نصنف x في المجموعة G_0 (في مجموعة الراسبين)

والخط المستقيم على الشكل (6-3) يوضح ذلك .

ولكن الشكل (6-3) يظهر لنا أن جودة التمثيل لذلك المستقيم ضعيفة جداً (لأن قيمة R^2 صغيرة) . لذلك كان لابد من البحث عن حل آخر أو نموذج آخر لتمثيل العلاقة بين $P_1(x)$ وX، ومن أجل البحث عن تلك العلاقة، سنحاول الاستفادة من شكل العلاقة (6-16) ونستبدل $P_1(x)$ بتابع مستمر جديد ومناسب، وهو متحول الأرجحية *(odds)*، والذي يعرف بدلالة الاحتمال $P_1(x)$ من خلال العلاقة (6-16) والتي تأخذ الشكل التالي :

$$odds(x) = \frac{P_1(x)}{1 - P_1(x)} = \frac{\text{احتمال تحقق } Y}{\text{احتمال عدم تحقق } Y} = \frac{n_1}{n_0} \qquad (25-6)$$

حيث أن: $P_1(x)$ هو احتمال أن يأخذ التابع Y القيمة (1) عند القيمة x، أو احتمال أن ينتمي العنصر x إلى المجموعة G_1، ونكتب ذلك كما يلي :

$$P_1(x) = P(Y = 1/x) = P(G_1/x) \qquad (26-6)$$

ولإيجاد علاقة الانحدار بين هذه الأرجحية $(odds)$ والمتحول المستقل X، نفترض أنهما يرتبطان بعلاقة خطية لوغاريتمية كالعلاقة (6−4) السابقة، والتي نكتبها كما يلي:

$$ln(odds) = ln\left(\frac{P_1(x)}{1-P_1(x)}\right) = \propto + \beta x \qquad (27-6)$$

ويسمى التابع اللوغاريتمي الأيسر باسم $logit(P_1(x))$ ويكتب على الشكل التالي :

$$logit[P_1(x)] = ln\left[\frac{P_1(x)}{1-P_1(x)}\right] = ln(odds) \qquad (28-6)$$

أي أن التابع $logit[P_1(x)]$ هو عبارة عن تحويل الاحتمال $P_1(x)$ حسب (6−26) إلى $(odds)$ ثم إلى الشكل اللوغاريتمي $ln\left[\frac{P_1(x)}{1-P_1(x)}\right]$، وهو عبارة عن تابع مستمر ويأخذ قيمه في المجال $]-\infty, +\infty[$ ، لأنه لدينا $0 \le P_1(x) \le 1$ ، فيكون $0 \le \frac{P_1(x)}{1-P_1(x)} < +\infty$، وبالتالي فإن $-\infty < ln\left[\frac{P_1(x)}{1-P_1(x)}\right] < +\infty$ ، ومن (6−27) و(6−28) يمكننا أن نفترض أن العلاقة بين التابع $logit[P_1(x)]$ والمتحول X هي خطية وتأخذ الشكل التالي :

$$logit[P_1(x)] = \propto + \beta X = ln(odds) \qquad (29-5)$$

وبعد حساب القيم العددية لـ $logit[P_1(x)]$ من العلاقة (6−28)، يمكننا إيجاد تقديرات لـ $\propto$ و β بتطبيق طريقة المربعات الصغرى أو طريقة الامكانية العظمى .

والآن نعود إلى العلاقة (6−27) فنجد أنه يمكننا كتابتها على الشكل التالي :

$$\frac{P_1(x)}{1-P_1(x)} = e = \propto + \beta X \qquad (30-6)$$

ومنها يمكننا أن نستخرج $P_1(x)$ كما يلي:

نقسم البسط والمقام في الطرف الأيسر على $P_1(x)$ فنجد أن:

$$\frac{1}{\frac{1}{P_1(x)} - 1} = e^{\propto + \beta X}$$

ثم نأخذ مقلوب الطرفين فنجد أن:

$$\frac{1}{P_1(x)} - 1 = \frac{1}{e^{\propto + \beta X}} = \bar{e}^{(\propto + \beta X)}$$

$$\frac{1}{P_1(x)} = 1 + \bar{e}^{(\propto + \beta X)}$$

ثم نأخذ مقلوب الطرفين مرة أخرى فنجد أن:

$$P_1(x) = \frac{1}{1 + \bar{e}^{(\propto + \beta X)}} = \frac{e^{\propto + \beta X}}{1 + e^{\propto + \beta X}} \qquad (31-6)$$

وبناء على (6-18) نجد أن:

$$P(G_1/x) = P_1(x) = \frac{e^{\propto + \beta X}}{1 + e^{\propto + \beta X}} = \frac{1}{1 + \bar{e}^{(\propto + \beta X)}} \qquad (32-6)$$

وهو عبارة عن منحني التابع اللوجستي المرسوم على الشكل (3−6) . وهنا نلاحظ أن هذا المنحني يختلف جذرياً عن المستقيم المرسوم على نفس الشكل، لأنه يقترب بطرفه الأيسر من نقاط المجموعة G_0، ويقترب بطرفه الأيمن من نقاط المجموعة G_1، وهو يعطينا بدقة أفضل، احتمال أن ينتمي x إلى G_1 مقابل كل قيمة x من قيم X .

ولحساب الاحتمال المتمم له نقوم بحساب $P_0(x)$ من العلاقة :

$$P_0(x) = 1 - P_1(x) = 1 - \frac{1}{1+\bar{e}^{(\propto+\beta X)}} = \frac{\bar{e}^{(\propto+\beta X)}}{1+\bar{e}^{(\propto+\beta X)}} \qquad (33-6)$$

وبتقسيم البسط والمقام على البسط نحصل على أن:

$$P_0(x) = P(G_0/x) = \frac{1}{1+e^{\propto+\beta X}} \qquad (34-6)$$

قاعدة: ولاتخاذ قرار حول انتماء أي عنصر x لإحدى المجموعتين نطبق القاعدة التالية:

$$\begin{array}{l} \text{إذا كان } P_1(x) \geq P_0(x) \text{ نصنف } x \text{ في المجموعة } G_1 \\ \text{إذا كان } P_1(x) < P_0(x) \text{ نصنف } x \text{ في المجموعة } G_0 \end{array} \qquad (35-6)$$

والمنحني المنطقي الملتوي على الشكل (6−3) يوضح ذلك [مع ملاحظة أن G_1 حلت محل G_2 وأن G_0 حلت محل G_1 من الشكل (6−2)] .

ويمكننا تطوير أو تعديل القاعدة (6−35) السابقة لتصنيف العناصر x، وذلك بأخذ نسبة الاحتمالين التالين $\frac{P_1(x)}{P_0(x)}$ فنجد أن :

$$\frac{P_1(x)}{P_0(x)} = \frac{\frac{1}{1+\bar{e}^{(\propto+\beta X)}}}{\frac{\bar{e}^{(\propto+\beta X)}}{1+\bar{e}^{(\propto+\beta X)}}} = e^{\propto+\beta X} \qquad (36-6)$$

وبذلك تصبح قاعدة التصنيف لأي عنصر x كما يلي:

قاعدة: نصنف أي عنصر x_0 إلى المجموعة G_1 إذا كانت النسبة :

$$\frac{P_1(x_0)}{P_0(x_0)} \geq 1 \quad \Leftrightarrow \quad e^{\propto+\beta x_0} \geq 1 \qquad (37-6)$$

ونصنف x إلى المجموعة G_0 إذا كانت النسبة :

$$\frac{P_1(x_0)}{P_0(x_0)} < 1 \quad \Leftrightarrow \quad e^{\propto+\beta x_0} < 1 \qquad (38-6)$$

ويمكن تحويل هذه القاعدة إلى الشكل الخطي بأخذ اللوغاريتم الطبيعي للطرفين فنحصل على القاعدة التالية .

قاعدة: نصنف أي عنصر x_0 إلى المجموعة G_1 إذا كانت قيمة التركيب الخطي موجبة أو غير سالبة، أي إذا كان :

$$\propto +\beta x_0 \geq 0 \qquad (39-6)$$

نصنف أي عنصر x إلى المجموعة G_0 إذا كانت قيمته سالبة، أي إذا كان :

$$\propto +\beta x_0 < 0 \qquad (40-6)$$

وبذلك نحصل على توابع تمييزية فاصلة بين المجموعات بأساليب متعددة . ولكن الاختلاف بين هذه القاعدة والقواعد الخطية (في الفصول السابقة)، هو أنها تستند على النسبة بين التوزيعات الاحتمالية اللاحقة للمجموعات بينما كانت القواعد السابقة تعتمد على النسبة بين الاحتمالات السابقة P_2 و P_1 .

كما إن شكل العلاقة (6-32) لـ$P_1(x)$ يساعدنا في الحصول على تقدير لـ $\propto$ و β بطريقة الامكانية العظمى .

6-4 : تقدير معالم النموذج اللوجستي (بطريقة الامكانية العظمى MLE) بمتحول واحد:

(*Maximum Likelihood Estimation*) بتصرف وإضافة *[Webb P.159]*. X

إن تقدير معالم النموذج اللوجستي $\propto$ و β بطريقة الإمكانية العظمى MLE. يعتمد على بيانات إحصائية معينة، ويتطلب حساب المواصفات الميدانية لتلك للبيانات، ويرتبط بتصميم المعاينة التي تطبق على المجموعتين G_1 و G_0 .

وهناك عدة تصاميم لهذه المعاينة هي :

1- المعاينة المختلطة : وتكون من توزيعات مختلطة بين المجموعتين G_1 و G_0 (توزيع مشترك) .

2- المعاينة الشرطية لـ X: تجري بحيث يكون x ثابتاً، ثم نسحب عينة أو أكثر من العناصر (التي يمكن أن تنتمي إلى G_1 أو إلى G_0) .

3- المعاينة المنفصلة من كل مجموعة على حدة : حيث تكون التوزيعات الشرطية $P(x/G_1)$ أو $P(x/G_0)$ هي التوزيعات المعتمدة .

علماً بأن طريقة الإمكانية العظمى تعطينا تقديرات للمعلم β، تكون مستقلة عن شكل تصميم المعاينة . وإن بعض تصاميم المعاينة تعطينا تقديرات أفضل من تصاميم أخرى للمعلم β_0 (تصميم المعاينة المنفصلة) .

والآن لنفترض إننا نعمل ضمن المعاينة المختلطة، التي تفترض أن العينة العشوائية مسحوبة من المجتمع المختلط للمجموعتين بحجم n، ومؤلفة من: n_1 من G_1 و n_0 من المجموعة G_0، والتي سنرمز لها لضرورات رياضية بالرمز G_2 ولعدد عناصرها بـ n_2، وعندها نجد أن تابع الإمكانية العظمى L في هاتين المجموعتين يأخذ الشكل التالي:

$$L = L_1 * L_2 = \prod_{i=1}^{n_1} P(x_{1i}/G_1) * \prod_{i=1}^{n_2} P(x_{2i}/G_2) \qquad (41-6)$$

حيث أن $i: 1\ 2\ 3 \dots n_s$ وحيث أن: x_{si} هو المشاهدة المسحوبة من المجموعة s وأن: $s: 1\ 2$

وبما أن التوزيع الشرطي لـ $P(x\ /G_s)$ يساوي :

$$P(x\ /G_s) = \frac{P(x) * P(G_s/x)}{P(G_s)} \qquad : s = 1\ 2 \qquad (42-6)$$

نقوم الآن بتعويض $P(x/G_s)$ من (6-42) في العلاقة (6-41) فنحصل على أن:

$$L = \prod_{i=1}^{n_1} P(G_1/x_{1i}) \frac{P(x_{1i})}{P(G_1)} * \prod_{i=1}^{n_2} P(G_2/x_{2i}) \frac{P(x_{2i})}{P(G_2)} \qquad (43-6)$$

وبإخراج التوزيعين $P(G_1)$ و $P(G_2)$ خارج الجداء لأنه ليس لهما علاقة بدليل الجداء i ، فنجد أن :

$$L = \frac{1}{P(G_1) * P(G_2)} * \prod_{i=1}^{n_1} P(x_{1i}) * P(G_1/x_{1i}) * \prod_{i=1}^{n_2} P(x_{2i}) * P(G_2/x_{2i}) \quad (\mathbf{44-6})$$

وبما أن $P(x_{1i})$ و $P(x_{2i})$ ليس لهما علاقة بالمعلمتين $\propto$ و β للنموذج، لذلك نكتب جداءاتهما على الشكل التالي :

$$\prod_{i=1}^{n_1} P(x_{1i}) * \prod_{i=1}^{n_2} P(x_{2i}) = \prod_{i=1}^{n} P(x_i) : \quad (\mathbf{45-6})$$

وهنا نلاحظ أن الجداء الأخير قد أصبح مأخوذاً على كامل حجم العينة $\boldsymbol{n}$

وبذلك نجد بأن تابع الامكانية العظمى يأخذ الشكل التالي:

$$L = \frac{\prod_{i=1}^{n} P(x_i)}{P(G_1) * P(G_2)} * \prod_{i=1}^{n_1} P(G_1/x_{1i}) * \prod_{i=1}^{n_2} P(G_2/x_{2i}) \qquad (46-6)$$

وبما أن الحد $\frac{\prod P(x_i)}{P(G_1)*P(G_2)}$ ليس له علاقة بالمعلمتين $\propto$ و β للنموذج اللوجستي . لذلك يمكننا افتراض (كما فعل اندرسون 1967). أن L مستقل عن الاحتمالات السابقة $P(x)$. وبذلك يمكننا اختصار التابع L إلى تابع مكافئ له L' يساوي :

$$L' = \prod_{i=1}^{n_1} P(G_1/x_{1i}) * \prod_{i=1}^{n_2} P(G_2/x_{2i}) \qquad (47-6)$$

والآن نأخذ اللوغاريتم الطبيعي للطرفين في (6-47) فنجد أن :

$$ln\, L' = \sum_{i=1}^{n_1} ln\, P(G_1/x_{1i}) + \sum_{i=1}^{n_2} ln\, P(G_2/x_{2i})$$

وبتعويض ما تحت اللوغاريتمات بما تساويها من العلاقتين (6-32) و(6-34)، نحصل على أن $ln\, L'$ يساوي :

$$ln\, L' = \sum_{i=1}^{n_1} (\propto + \beta X_{1i}) - \sum_{i=1}^{n_1} ln\left(1 + e^{\propto + \beta x_{1i}}\right) - \sum_{i=1}^{n_2} ln\left(1 + e^{\propto + \beta x_{2i}}\right) \quad (\mathbf{48-6})$$

وبعد دمج المجموعين الأخيرين وأخذ المجموع على كامل العينة n نجد أن :

$$ln\, L' = \sum_{i=1}^{n_1} (\propto + \beta X_{1i}) - \sum_{i=1}^{n} ln\left(1 + e^{\propto + \beta x_i}\right) \qquad (49-6)$$

والآن نقوم بأخذ المشتقات الجزئية لـ($\ell n\, L'$) بالنسبة لـ $\propto$ و β ونضعها مساوية للصفر، فنحصل بعد الإصلاح والاستبدال على أن هذين المشتقين يساويان:

$$\frac{\partial\, \ell n\, L'}{\partial \propto} = n_1 - \sum_{i=1}^{n} P(G_1/x_i) = 0 \qquad (50-6)$$

$$\frac{\partial\, \ell n\, L'}{\partial \beta} = \sum_{i=1}^{n_1} x_{1i} - \sum_{i=1}^{n} (x_i) * P(G_1/x_i) = 0 \qquad (51-6)$$

مع الانتباه إلى أن المجموعين الأخيرين ( ) $\sum_{i=1}^{n}$ مأخوذين على جميع قيم x في العينة n، ثم نقوم بحل هاتين المعادلتين فنحصل على تقدير لـ $\propto$ و β، ومنهما نحصل على النموذج اللوجستي المطلوب .

ملاحظة: إذا كان عدد المتحولات المؤثرة X يساوي p متحولاً فإننا سنرمز لها بشعاع واحد كما يلي :

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

وعندها يمكننا كتابة العلاقة (6-27) كما يلي :

$$ln\left[\frac{P_1(x)}{1-P_1(x)}\right] = \propto + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p = \propto + \beta' X \qquad (52-6)$$

حيث أن: $\beta'(\beta_1\ , \beta_2 \ldots\ \beta_p\ )$

وعندها أيضاً فإن صيغة النموذج اللوجستي تأخذ الشكل التالي :

$$P_1(x) = \frac{1}{1+\bar{e}^{(\propto+\beta' X)}} = \frac{e^{(\propto+\beta' X)}}{1+e^{(\propto+\beta' X)}} \qquad (53-6)$$

ثم نجد أن $P_0(x)$ تحسب من العلاقة:

$$P_0(x) = 1 - P_1(x) = \frac{1}{1+e^{\propto+\beta_1' X}} \qquad (53\ a-6)$$

وعندها فإن معادلات الإمكانية العظمى لحساب $\propto$ و β' تأخذ الشكل التالي :

$$\frac{\partial\, ln\, L'}{\partial \propto} = n_1 - \sum_{i=1}^{n} P(G_1/x_i) = 0 \qquad (54-6)$$

$$\frac{\partial\, ln\, L'}{\partial \beta_j} = \sum_{i=1}^{n_1} (x_{1i})_j - \sum_{i=1}^{n} (x_i)_j * P(G_1/x_i)_j = 0 \qquad (55-6)$$

حيث أن J تأخذ القيم $j: 1\ 2\ 3 \ldots P$

ومن هذه المعادلات يمكننا حساب تقديرات لـ $\propto$ و β_p ... β_2 , β_1، علماً بأن المعادلات (6-54) و(6-55) هي معادلات غير خطية بالنسبة لـ $\propto$ و β، وإن حلها يحتاج إلى أساليب وبرامج متقدمة .

مثال (6–2): بناءً على بيانات المثال (6–1) السابق تم تقدير معالم النموذج (6–31) بطريقة الامكانية العظمى، فحصلنا على الجدول التالي: [المصدر: *Wikipedia*] :

جدول (6–3): نتائج علاقة الانحدار لـ *logit P(X)* على *X*

البيان	قيم الأمثال	الانحراف المعياري	قيمة Z	قيمة P حسب اختبار *(Wald)*
الثابت	$\propto = -4.0777$	1.7610	-2.316	0.0206
الساعات X	$\beta = 1.5046$	0.6287	2.393	0.0167

نلاحظ أن هذه المخرجات تشير إلى أن عدد ساعات الدراسة X ترتبط معنوياً مع احتمال النجاح في الامتحان (وذلك لأن قيمة P حسب اختبار $Wald$ تساوي $P = 0.0167$، وهي أصغر من قيمة مستوى الدلالة 0.05).

وإن معادلة العلاقة بين $logit[P_1(x)]$ وX تساوي :

$$logit[P_1(x)] = ln\left[\frac{P_1(x)}{1-P_1(x)}\right] = 1.5046X - 4.0777 = 1.5046[X - 2.71]$$

وهكذا نجد أن معادلة الأرجحية مع x هي:

$$odds(Y) = \frac{P_1(x)}{1-P_1(x)} = e^{1.5046(X-2.71)}$$

ومنها نجد أنه إذا كان عدد الساعات $X = 2.71$ فإن:$odds(Y) = e^0 = 1$، وفي هذه الحالة نستنتج أن إمكانية النجاح $(odds)$ تساوي إمكانية الرسوب، وإن احتمال كل منهما يساوي 0.50، وفي هذه الحالة (عندما $X=2.71$) يمكننا أن نقول أن : أرجحية النجاح مقابل الرسوب هي كما يلي: واحد مقابل واحد ونكتب ذلك كما يلي (1 : 1).

ولحساب الاحتمال $P_1(x)$ من النموذج نلاحظ أن:

$$P_1(x) = \frac{1}{1+\bar{e}^{(\tilde{\propto}+\tilde{\beta}X)}} = \frac{1}{1+\bar{e}^{(-4.0777+1.5046X)}}$$

فعندما يكون $X = 2$ نجد أن احتمال النجاح يساوي :

$$P_1(2) = \frac{1}{1+\bar{e}^{(-4.0777+1.5046*2)}} = 0.26 \quad \Longrightarrow \quad P_0(2) = 0.74$$

وهذا يعني أن احتمال نجاح من يدرس ساعتين $(X=2)$ أصغر من احتمال رسوبه (0.74)، لذلك نصنف ذلك الطالب في المجموعة G_0 ونعتبره من مجموعة الراسبين .

أما عندما يكون $X=2.71$ فإننا نجد أن احتمال النجاح يساوي:

$$P_1(x) = \frac{1}{1+e^0} = \frac{1}{2} = 0.50$$

وهذا ما لاحظناه سابقاً وهذا يعني أن القيمة X=2,71 هي النقطة الفاصلة بين النجاح والرسوب .

أما عندما تكون $X=4$ فإن احتمال النجاح $P_1(x)$ يساوي :

$$P_1(4) = \frac{1}{1 + \bar{e}^{(-4.0777+1.5046*4)}} = 0.87 \quad \Longrightarrow \quad P_0(4) = 0.13$$

وهذا يعني أن احتمال نجاح من يدرس (X=4) ساعات أكبر من احتمال رسوبه (0.13)، لذلك نصنف ذلك الطالب في المجموعة G_1 [مجموعة الناجحين] .

أما عندما يكون X=3 فإن احتمال النجاح يساوي :

$$P_1(3) = \frac{1}{1 + \bar{e}^{(-4.0777+1.5046X)}} = 0.61 \quad \Longrightarrow \quad P_0(3) = 0.39$$

أي أن احتمال نجاح من يدرس (X=3) ساعات أكبر من احتمال رسوبه (0.39)، ولذلك نصنف ذلك الطالب في المجموعة G_1 [مجموعة الناجحين]، رغم إنه كان من بين الراسبين (انظر الجدول ($6-1$)).

وأخيراً يمكننا أن ننظم بعض النتائج الممكنة لهذا النموذج في جدول مناسب كالجدول التالي :

جدول ($6-4$): قيم التحليل اللوجستي

عدد ساعات الدراسة X	مؤشرات النجاح في الامتحان		
	قيمة $ln\ (odds)$	قيمة الـ $odds$	احتمال النجاح $P_1(x)$
1	-2.57	$0.078 \approx 1:13.1$	0.07
2	-1.07	$0.034 \approx 1:12.91$	0.26
2.71	0.00	$1 \approx 1:1$	0.50
3	0.44	$1.55 \approx \cdots$	0.61
4	1.94	$6.96 \approx \cdots$	0.87
5	3.45	$31.4 \approx \cdots$	0.97
6	4.95	141.16085	0.99

وبالعودة إلى الجدول ($6-3$) نجد أن احتمال الدلالة تساوي $P = 0.0167$، علماً بأن هذه القيمة محسوبة استناداً إلى علامة اختبار $Wald - Z$. ولكن هناك طريقة أفضل من طريقة $Wald$ – الطريقة المعتمدة في حساب قيمة P للتوابع اللوجستية – وهي طريقة اختبار نسبة الامكانية العظمى (LRT)، والتي تعطينا من هذه البيانات أن قيمة P للنموذج اللوجستي المدروس $P = 0.00006$.

6-5 : التحليل اللوجستي المتعدد: [webb P. 161 بتصرف]

إن التحليل اللوجستي المتعدد هو تعميم للتحليل اللوجستي الثنائي، وهو يعالج الحالات التي يكون فيها المجتمع مؤلفاً من عدة مجموعات (أو فئات) منفصلة نرمز لها بـ :

$$G_1 \ G_2 \ \dots \ G_j \dots \ G_g \qquad (58-6)$$

وعندها فإننا نشكل تابع الـ $logit$ لكل زوج من هذه المجموعات، وإذا أخذنا المجموعة G_j مقابل أية مجموعة أخرى ولتكن G_g فإن تابع الـ $logit$ يأخذ الشكل التالي:

$$logit\left[P_j(x)\right] = ln\left[\frac{P_j(x)}{P_g(x)}\right] = \propto + * \beta_j' X \qquad (59-6)$$

حيث أن: $j = 1\ 2\ 3 \dots g-1$ وأن: $X'(X_1 \ X_2 \ \dots \ X_P)$ وأن: $\beta'(\beta_1 \ \beta_2 \ \dots \ \beta_P)$

وهي تمثل $(g-1)$ تابعاً تمييزياً تفصل بين تلك المجموعات .

وهذا يعني أن لوغاريتم نسبة الإمكانية، $\left[\frac{P_j(x)}{P_g(x)}\right]$ لأي زوج ممكن من المجموعات يرتبط مع المتحولات X بعلاقة خطية. وهي تشكل المستوى الفاصل بينهما .

وبطريقة مشابهة لما عرضناه في التحليل اللوجستي الثنائي يمكننا صياغة الاحتمالات اللاحقة $P_j(x)$ و $P_g(x)$ بدلالة X بواسطة العلاقتين التاليتين :

$$P_j(x) = P(G_j/x) = \frac{e^{\propto_j + \beta_j' X}}{1 + \sum_{j=1}^{g-1} e^{\propto_j + \beta_j' X}} \qquad (60-6)$$

حيث أن: $j = 1\ 2\ 3 \dots g-1$ وأن $(g-1)$ هو عدد التوابع في $(59-6)$ أما التابع المتمم $P_g(x)$ فيساوي :

$$P_g(x) = P(G_g/x) = \frac{1}{1 + \sum_{j=1}^{g-1} e^{\propto_j + \beta_j' X}} \qquad (61-6)$$

وهكذا نجد أن قاعدة التصنيف التمييزي تصبح تابعة للعلاقة الخطية $(\propto_j + \beta_j' X)$. وتأخذ الصيغة التالية:

قاعدة: نصنف x_0 إلى المجموعة G_k إذا كانت $0 < \propto_k + \beta_k' x_0$ وكانت أكبر من القيم الأخرى: أي إذا كانت:

$$0 < \propto_k + \beta_k' x_0 = Max[\propto_j + \beta'] \qquad (62-6)$$

حيث أن: $j = 1\ 2\ 3 \dots g-1$ وإذا كان العكس نصنفه إلى المجموعة G_g .

وكذلك يمكننا أن نستخلص تابع الإمكانية العظمى من العلاقة:

$$L = \prod_{j=1}^{g} \prod_{i=1}^{n_i} P(x_{ji}/G_j) \qquad (\mathbf{63-6})$$

وبإجراء نفس العمليات والمعالجات على L نحصل على التابع المكافئ له L' ونأخذ لوغاريتمه فنجد أن:

$$ln\left(L'\right) = \sum_{j=1}^{g} \sum_{i=1}^{n_j} ln\, P(G_j/x_{ji}) \qquad (64-6)$$

وباشتقاقه نحصل على المعادلات التالية:

$$\frac{\partial ln(L')}{\partial \propto} = n_j - \sum_{X\text{كل}}^{n} P(G_j/X) = 0 \qquad (65-6)$$

$$\frac{\partial ln(L')}{\partial (\beta_j)_s} = \sum_{i=1}^{n_j} (x_{ji})_s - \sum_{X\text{كل}}^{n} x_s * P(G_j/X)_s = 0 \qquad (\mathbf{66-6})$$

حيث أن: *j: 1 2 3 ..... g-1*

وبحل هذه المعادلات نحصل على تقديرات للمعلمين $\propto$ و β ، ولكن بما أن هذه المعادلات ليست خطية، فإنه يتم البحث عن حلول تقاربية لها (بطريقة نيوتن أو بطريقة المعاودة)، وعندها نحتاج إلى حل ابتدائي ننطلق منه لإيجاد الحلول المتتالية والمتقاربة، ويمكن أن نأخذ الحل الابتدائي المقابل للقيم الصفرية، ونضع في البداية $\propto_j = 0$ و $\beta_1 = \beta_2 = \cdots \beta_g = 0$، ثم نتابع البحث عن قيم $\propto$ و β المثالية .

6-6 : تقييم جودة النموذج اللوجستي: [Wikipedia بتصرف] :

إن عملية تقييم جودة النموذج اللوجستي تختلف عن عملية تقييم الجودة في الانحدار الخطي، ومع أن الانحدار اللوجستي يقدر معالم النموذج β_j بطريقة الإمكانية العظمى، فهو لا يعتمد على معامل التحديد R^2 لتقييم جودة التمثيل . ولكنه يعتمد في تقييم جودة التمثيل على مفاهيم جديدة هي:

- **النموذج المشبع (Saturated Model)**: وهو النموذج الذي يمثل (نظرياً) البيانات المدروسة تمثيلاً تاماً، ونرمز لتابع الإمكانية العظمى لهذا النموذج بالرمز L_S . مأخوذة من (Likelihood of the saturated model) ولكن عملية الحصول على هذا النموذج في الحالة العامة قد تكون غير ممكنة . ويبقى L_S مجهولاً .
- **النموذج الصفري (Null Model)**: وهو النموذج الذي يتوافق مع فرضية العدم $(H_0: \beta_j = 0)$. أي أنه النموذج الذي لا يتضمن أي من المتحولات X، ويأخذ قيمة ثابتة هي قيمة الثابت β_0 . ونرمز لتابع الإمكانية لهذا النموذج بالرمز L_0 .
- **النموذج المقدر (Fitted Model)**: وهو النموذج الذي ينتج عن حساب وتقدير المعالم β_i بطريقة الإمكانية العظمى، وهو يتضمن بعض المتحولات X (واحد على الأقل). ونرمز لتابع الإمكانية العظمى لهذا النموذج بالرمز L_M . علماً بأن توابع الإمكانية العظمى L_0 و L_M تحسب اعتماداً على العلاقة (6-41) أو على العلاقة (6-47) وهي تأخذ شكل جداءات توزيع (بيرنويللي) التالي:

$$L = \prod_{\substack{i=1 \\ j=1,2}}^{n} P(G_j/x_{ji}) = \prod_{i=1}^{n} \left(P_1(x_i)\right)^{y_i} [1 - P(x_i)]^{1-y_i} \qquad (67-6)$$

حيث y_i تأخذ (1) أو (0) .

وبناء على ذلك تم تعريف المؤشرات التالية:

1- **حيدان النموذج المقدر (Deviance Model)**: ويعرف بالعلاقة التالية:

$$D_{\text{ff}} = -2\ln\frac{\left(\text{قيمة تابع الإمكانية العظمى للنموذج المقدر}\right)}{\left(\text{قيمة تابع الإمكانية العظمى للنموذج المشبع}\right)} = -2\ln\left[\frac{L_M}{L_S}\right] \qquad (68-6)$$

وهو يعبر عن الاختلاف النسبي بين النموذج (بمتحول واحد أو أكثر) وبين النموذج المشبع، لقد تم وضع الإشارة السالبة قبل اللوغاريتم لأن $L_M < L_S$.

- **حيدان النموذج الصفري** ويعرف كما يلي:

$$D_0 = -2\ln\frac{\left(\text{قيمة تابع الإمكانية العظمى للنموذج الصفرى}\right)}{\left(\text{قيمة تابع الإمكانية العظمى للنموذج المشبع}\right)} = -2\ln\left[\frac{L_0}{L_S}\right] \qquad (69-6)$$

- الفرق Δ ويحسب للتخلص من L_S المجهولة، وعند حساب الفرق بين D_0 و $D_{\not f}$ نجد أن:

$$\Delta = D_0 - D_{\not f} = -2\,ln\left[\frac{L_0}{L_S}\right] + 2\,ln\left[\frac{L_M}{L_S}\right] = -2\left[ln\left(\frac{L_0}{L_S}\right) - ln\left(\frac{L_M}{L_S}\right)\right]$$

$$\Delta = D_0 - D_{\not f} = -2\,ln\left[\frac{\frac{L_0}{L_S}}{\frac{L_M}{L_S}}\right] = -2\,ln\left[\frac{L_0}{L_M}\right] = -2[ln\,L_0 - ln\,L_M\,] \qquad (70-6)$$

ويستخدم هذا الفرق Δ في تقييم جودة التمثيل، لأنه من هذه العلاقات يمكننا أن نستنتج أن: $L_0 < L_M < L_S$ وأن: $D_0 > D_{\not f}$ وأنه كلما ازدادت قيمة L_M واقتربت من L_S كان التمثيل جيداً، ولكن زيادة L_M تعني زيادة $D_{\not f}$، وهذا يؤدي إلى تناقص الفرق $\Delta = D_0 - D_{\not f}$، وهذا يعني أنه كلما تناقص الفرق Δ كان التمثيل جيداً . لذلك يجب علينا عند تقدير المعالم $\propto$ و β البحث عن النموذج الذي يجعل Δ أصغر ما يمكن، ولهذا السبب تخصص البرامج الحاسوبية عموداً خاصاً للفرق $\Delta = -2\,ln\left[\frac{L_0}{L_M}\right]$ وتراقب بتغيراته وتتخذه كمعيار للتوصل إلى الحل المثالي للمعادلات (6-54) و(6-55) بطريقة المعاودة . وذلك عندما يأخذ ذلك الفرق Δ قيمة ثابتة خلال المعاودات الأخيرة .

2- **دراسة قيم معاملات التحديد اللوجستية:** وهناك عدة معاملات تحديد لتقييم جودة النموذج اللوجستي وهي :

- معامل نسبة الإمكانية العظمى (Likelihood Ratio:

$$R_L^2 = \frac{D_0 - D_{\not f}}{D_0} = 1 - \frac{ln\left[\frac{L_M}{L_S}\right]}{ln\left[\frac{L_0}{L_S}\right]} \qquad (71-6)$$

- معامل cox & snele:

$$R_{cs}^2 = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}} = 1 - e^{\frac{2[ln\,L_0 - ln\,L_M]}{n}} \qquad (72-6)$$

- معامل Mc Fadden:

$$R_{MCF}^2 = 1 - \frac{ln(L_M)}{ln(L_0)} \qquad (73-6)$$

وإن هذه المعاملات ترتبط مع بعضها بالعلاقات التالية :

$$R_{cs}^2 = 1 - \left(\frac{1}{L_0}\right)^{\frac{2}{n}R_{MCF}^2} \qquad (74-6)$$

$$R_{MCF}^2 = \frac{n}{2}\frac{(1 - R_{cs}^2)}{ln(L_0)} \qquad (75-6)$$

3-**حساب جدول تقاطع حالات التصنيف اللوجستي مع التصنيف الفعلي** (السابق) والذي يأخذ الشكل التالي (في حالة مجموعتين) :

جدول (6-5)

اللوجستي / الفعلي	G_1	$G_2 = G_0$	المجموع
G_1	n_{11}	n_{12}	n_1'
$G_2 = G_0$	n_{21}	n_{22}	$n_2' = n_0$
المجموع	n_1	n_2	n

ومنه يتم حساب معدل التصنيف الصحيح بحساب نسبة مجموع عناصر القطر الرئيسي على المجموع الكلي n فنجد أن:

$$R = \frac{n_{11} + n_{22}}{n} 100\% \qquad (76-6)$$

ومنه يمكن حساب معدل التصنيف الخاطئ :

$$MR = 1 - R = \frac{n_{12} + n_{21}}{n} 100\% \qquad (77-6)$$

4- **حساب المعامل** kappa من العلاقة:

$$kappa = \frac{n \sum^2 n_{ii} - \sum^2 n_i * n_i'}{n^2 - \sum^2 n_i * n_i'} 100\% \qquad (78-6)$$

وهناك مقاييس أخرى لتحليل جودة مثل هذه الجداول، وكالحساسية (SE) والخصوصية (SP) ونسبة الأرجحية (OR) وغيرها . وهي مذكورة ومعرفة في العلاقات (6-85) و(6-87) و(6-89) من الفقرة اللاحقة (6-7) .

5-**اختبار (هوسمير- ليمشو** Hosmer- Lemshow**): لجودة المطابقة:** وهو يختبر صحة فرضية العدم التالية H_0: يتساوى عدد الحالات المشاهدة مع عدد الحالات المتوقعة (أي أن النموذج يمثل البيانات بشكل صحيح). وعندها نتخذ القرار بقبول فرضية العدم H_0، إذا كان مستوى المعنوية أو احتمال الدلالة p لاختبار (كاي مربع) أكبر من مستوى الدلالة المحدد بـ $\propto$.

6- **دراسة معنوية المتحولات الداخلة فيه:** وذلك من خلال مقارنة قيم احتمال الدلالة $(P = sig)$ مع مستوى الدلالة $\propto = 0.05$. فإذا كانت قيمة P_j المقابلة للمتحول X_j أصغر من 0.05 ، تكون علاقة ذلك المتحول مع النموذج معنوية والعكس بالعكس . ولكن دراسة هذه المعنوية تعتمد على اختبار جديد هو اختبار $(Wald)$ وهو اختبار شبيه بالاختبار الطبيعي ويعرف بالعلاقة: $Wald = \left(\frac{\beta_i}{SE(\beta_i)}\right)^2$.

7- **دراسة قوة تأثير كل متحول** X_j **على النموذج:** وذلك من خلال حساب نسبة الأرجحية له OR (Odds Ratio) والتي يمكن تعريفها وحسابها من العلاقة التالية :

$$\text{OR} = \frac{odds(x_j + 1)}{odds(x_j)} = \frac{e^{\widetilde{\propto} + \widetilde{\beta}(x_j+1)}}{e^{\widetilde{\propto} + \widetilde{\beta}(x_j)}} = e^{\beta} \qquad (78\,a-6)$$

لذلك يقوم البرنامج الحاسوبي بحساب القوة e^{β}، لجميع المتحولات الداخلة في النموذج . لأنها تدل على قوة تأثير المتحول X_j عندما يزداد بمقدار واحد (1) على النموذج ككل . وهو يعبر عن مرونة المتحول X_j بالنسبة للتابع (odds) .

مثال (6-3): في دراسة نشرها [شاهين 2014] حول تطبيق التحليل اللوجستي على مرضى سرطان الدم (اللوكيميا) في محافظة البصرة، لاحظ أولاً أن الجهات الطبية تصنف هؤلاء المرضى إلى نوعين أو مجموعتين هما: مرض سرطان الدم النخاعي الحاد (AME) ومرض سرطان الدم اللمفاوي (ALL) .
لذلك قام الباحث بسحب عينيتين عشوائيتين بحجمين متساويين $n_1 = n_2 = 80$ من ملفات هاتين المجموعتين، ثم قام بدراستها وتحليلها وتحديد المتحولات التي تفسر عملية الإصابة بهذين المرضين. فكانت ثمانية متحولات مختلطة (كمية ونوعية)، ثم قام بتحويلها إلى متحولات ثنائية وصنف قيم كل منها ضمن فئتين محددتين كما يلي:

X_1 – جنس المريض وصنفه إلى (ذكر = 1 ، انثى = 2) .

X_2 – عمر المريض وصنفه إلى [$(X_2 \leq 50) \ , \ 1 = (X_2 > 50) \ = 2$].

X_3 – وزن المريض وصنفه إلى [$(X_3 \leq 40) \ , \ 1 = (X_3 > 40) \ = 2$].

X_4 – نسبة الكريات الحمراء (P,C,V) وصنفه إلى [$(X_4 \leq 0.35) \ , \ 1 = (X_4 > 0.35) \ = 2$].

X_5 – هيموغلوبين الدم (H,B) وصنفه إلى [$(X_5 \leq 11.5) \ , \ 1 = (X_5 > 11.5) \ = 2$].

X_6 – معدل الكريات البيضاء (W,B,C) وصنفه إلى [$(X_6 \leq 7.5) \ , \ 1 = (X_6 > 7.5) \ = 2$].

X_7– **سرعة ترسب الكريات الحمراء** (E,S,Q) **وصنفه إلى** [$(X_7 \leq 22.5) , 1 = (X_7 > 22.5) \ = 2$]

X_8 – عدد الصفائح الدموية (P,C) وصنفه إلى [$(X_8 \leq 150) \ , \ 1 = (X_8 > 150) \ = 2$].

كما قام بتسمية تابع الاستجابة الثنائي Y، الذي يعبر عن نوع الإصابة بمرض سرطان الدم وصنف حالتيه (النخاعي واللمفاوي) كما يلي:

إذا كان المريض مصاب بسرطان الدم النخاعي فإن قيمة $Y = 0$.

إذا كان المريض مصاب بسرطان الدم اللمفاوي فإن قيمة $Y = 1$.

ثم قام بجمع البيانات اللازمة له من ملفات عناصر العينة المسحوبة من المجموعتين، وعمل على تحليلها وتحويلها إلى متحولات ثنائية، ووضعها في جداول منظمة، ثم قام بتطبيق برنامج التحليل اللوجستي على المتحولات الثنائية التالية :

$$Y: X_1 , X_2 , X_3 , X_4 , X_5 , X_6 , X_7 , X_8$$

واستخدم لذلك طريقة (enter)، واختار طريقة (نيوتن) لإجراء المعاودة لإيجاد الحلول التقريبية للمعادلات غير الخطية الواردة في العلاقات (6-54) و(6-55)...الخ . وذلك من أجل الحصول على معالم النموذج اللوجستي التالي :

$$\ln\left(\frac{P_1(x)}{1-P_1(x)}\right) = \propto + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_8 X_8$$

فكانت النتائج بعد إجراء (6) دورات للمعاودة كما يلي:

جدول (6-6): قيم الأمثال للتابع اللوجستي خلال دورات المعاودة:

رقم الدورة	$\Delta = -2\ell n\left[\frac{L_0}{L_M}\right]$	الثابت C	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	162.878	0.758	-2.268	-0.207	-0.871	-1.080	0.557	-0.359	-0.195	0.949
2	156984	0.901	-3.387	-0.236	-1.184	-1.370	0.669	-0.474	-0.194	1.246
3	156.243	0.913	-3.998	-0.228	-1.258	-1.130	0.679	-0.497	-0.185	1.212
4	156.200	0.912	-4.188	-0.227	-1.264	-1.416	0.679	-0.499	-0.183	1.316
5	156.200	0.912	-4.206	-0.227	-1.264	-1.416	0.679	-0.499	-0.183	1.316
6	156.200	0.912	-4.206	-0.227	-1.264	-1.416	0.679	-0.499	-0.183	1.316

وهناك نلاحظ أن قيم أمثال التابع اللوجستي قد استقرت تماماً بعد إجراء خمس أو ست دورات لمعاودة الحسابات .

كما إن قيمة مؤشر الجودة للنموذج المتمثل بالعمود الثاني الذي يتضمن قيم الفرق $(D_0 - D_\ell)$، والتي استقرت عند القيمة 156,200 في الدورة الرابعة ومابعدها. لذلك توقف عن الحسابات عند الدورة السادسة . ثم انتقل الباحث إلى تقدير معالم النموذج اللوجستي المتعدد فحصل على النتائج التالية:

جدول (6-7): نتائج الانحدار اللوجستي (أمثال النموذج اللوجستي وخواصها):

المقدرات المتحولات	β	S. E	Z Wald	df	Sig	Exp (B)	Lower of Ic 0.95	Upper of Ic 0.95
Constant	0.912	0.625	2.133	1	0.144	2.490	-	-
X_1 – جنس المريض	-4.206	1.071	15.426	1	0.000	0.015	0.002	0.122
X_2 – عمر المريض	-0.227	0.430	0.278	1	0.598	0.797	0.343	1.853
X_3 – وزن المريض	-1.264	0.435	8.461	1	0.004	0.283	0.121	0.662
X_4 – نسبة الكريات الحمراء	-1.416	0.878	2.603	1	0.107	0.243	0.043	1.356
X_5 – هيموغلوبين الدم	0.679	0.852	0.635	1	0.425	1.973	0.371	10.494
X_6 – معدل الكريات البيضاء	-0.499	0.459	1.179	1	0.278	0.607	0.247	1.494
X_7– سرعة ترسب	-0.183	0.407	0.202	1	0.653	0.833	0.375	1.850
X_8 – عدد الصفائح	1.316	0.404	10620	1	0.001	3.730	1.690	8.233

وعند دراسة هذا الجدول نجد أن عناصر β في العمود الأول هي نفسها العناصر التي استقرت عليها الحلول في السطر الأخير من الجدول (6-6) السابق. وهذا يجعلنا نكتب النموذج على شكل العلاقة (6-52) كما يلي:

$$ln(odds) = ln\left(\frac{P_1(X)}{1-P_1(X)}\right) = 0.912 - 4.206X_1 - 0.227X_2 - 1.264X_3 - 1.416X_4 + \\ +0.679X_5 - 0.499X_6 - 1.183X_7 - 1.316X_8$$

ومنها يمكننا حساب الاحتمال $P_1(X)$ وهو احتمال الانتماء إلى G_1 وكتابته حسب العلاقة (6-53) كمايلي:

$$P_1(X) = \frac{1}{1 + e^{-(0.912 - 4.206X_1 - 0.227X_2 - \cdots\ldots\ldots + 1.316X_8)}}$$

ومنها أيضاً نقوم بحساب احتمال الانتماء إلى G_0 من العلاقة:

$$P_0(X) = 1 - P_1(X) = \frac{1}{1 + e^{+(0.912 - 4.206X_1 - 0.227X_2 - \cdots\ldots\ldots + 1.316X_8)}}$$

ولاختبار جودة التوفيق استخدم معيار نسبة الامكانية العظمى ورمز لها بـ $\chi^2 = -2\,ln\left[\frac{L_0}{L_1}\right]$ ، وهي تتبع تقاربياً للتوزيع χ^2 وبدرجة حرية $(p_1 - p_0)$. حيث: p_0 و p_1 هما أبعاد L_0 و L_1 على الترتيب .

وحيث أن L_0 هي قيمة دالة الامكانية العظمى عند الفرضية H_0 في G_0

وإن L_1 هي قيمة دالة الامكانية العظمى عند الفرضية H_1 في G_1 . فنجد أن :

جدول (6-8): اختبار x^2 للتوفيق:

	$\chi^2 = -2\,ln\left[\frac{L_0}{L_1}\right]$	df	sig
Model	65.607	8	0.008

وهذا يدل على معنوية النموذج بشكل عام (لأن قيمة sig أقل بكثير من 0.05)، ونلاحظ أن الجدول (6-7) السابق يعطينا عموداً خاصاً باسم Wald، وهو عبارة عن مؤشر Wald لتعريف معنوية ومصداقية تقدير كل من الأمثال β_i وهو يحسب من العلاقة التالية:

$$Wald_i = \left(\frac{\tilde{\beta}_i}{S.E(\tilde{\beta}_i)}\right)^2$$

ومن خلال عمود sig نلاحظ أن هناك ثلاثة متحولات فقط ، ذات تأثير معنوي وهي X_1 و X_3 و X_8، لأن قيم sig المقابلة لها أقل من 0.05، أما بقية المتغيرات فليس لها تأثيرات ذات أهمية أو معنوية .

كما نلاحظ أن العمود الذي يضم $EXP(\beta_1) = e^{\beta_i}$ فهو يعبر عن مقدار زيادة تابع الاستجابة Y أو التابع $logit\ (P)$ عندما يزداد المتحول المرافق لـ β_i بمقدار واحد، فمثلاً نجد أن :

$$eXP(\beta_1) = e^{-4.206} = 0.015$$

وهو يعني أن تابع الاستجابة Y سيزداد بمقدار 0.015 إذا تعبر المتحول الأول X_1 بمقدار (1): أي إذا تعبر نوع الجنس من ذكر إلى أنثى .

وبناء على صيغة النموذج الأخيرة والمحددة نقوم بحساب الاحتمالات $P_1(X)$ لكل مفردة i من مفردات العينة ثم نقوم بحساب الاحتمالات المكملة $P_0(X)$ لكل مفردة i من العلاقة التالية:

$P_0(X) = 1 - P_1(X)$، ثم نقوم بمقارنة هذين الاحتمالين لكل مفردة i . ونصنف المفردة i كما يلي:

القاعدة: فإذا كانت $P_{1i}(X) \geq P_{0i}(X)$ فإننا ننسب تلك المفردة i إلى G_1 .

أما إذا كانت $P_{1i}(X) < P_{0i}(X)$ فإننا ننسب تلك المفردة i إلى G_0 .

وذلك كما فعلنا على الشكل (6-3) السابق.

وبعد إجراء كل هذه العمليات وتصنيف كل مفردات العينة في إحدى المجموعتين G_1 أو G_0، نقوم من جديد تبويب النتائج الجديدة للتصنيف مع نتائج التصنيف الأصلي المستخدم في الإدارة . وعند إجراء ذلك التبويب حاسوبياً حصل الباحث على الجدول التالي:

جدول (6-9):

		التصنيف المستنبط من النموذج		النسبة المئوية	مجموع الأعداد
		G_0	G_1	%	n_i'
التصنيف الأصلي الإداري	G_0	57	23	71.3	80
	G_1	21	59	73.8	80
	-	49.75	51.25	72.5	-
مجموع الأعداد	n_i	78	82	-	160

ومن هذا الجدول نستنتج أن احتمال التصنيف الصحيح في المجموعة G_0 فقط كان يساوي 71.3% وفي المجموعة G_1 كان يساوي 73.8% ولكن الاحتمال الاجمالي للتصنيف الصحيح كان يساوي 72.5% وهذا يعني أن المعدل الاجمالي للتصنيف الخاطئ يساوي 27.5% . ولتقدير جودة التصنيف نحسب المؤشر kappa فنجد أن:

$$kappa = \frac{n * (\sum n_{ii}) - \sum n_i n_i'}{n^2 - \sum n_i n_i'} = \frac{160(57 + 59) - [(80 * 78) + (80 * 82)]}{(160)^2 - [(80 * 78) + (80 * 82)]} = 0.45$$

وهي قيمة ضعيفة نسبياً . وتدل على جودة ضعيفة لعملية التصنيف المجراة في ذلك البحث .

6-7 : إضافات رياضية عن الأرجحية (odds):

لتوضيح مفهوم الأرجحية وعلاقتها بالاحتمالات للظواهر الثنائية، نفترض إننا نريد معرفة معدلات الإصابة بإحدى الأمراض (كالسكري مثلاً) بين الأشخاص المعرضين له (وراثياً وصحياً وسلوكياً)، فأخذنا عينة عشوائية مؤلفة من (1000) شخص من مجتمع المعرضين لذلك المرض، وأجرينا عليهم الفحوصات المخبرية والسريرية، ثم قمنا بتبويب نتائج هذه الاختبارات حسب حالة المريض الفعلية (D^+= مصاب فعلاً و D^-= غير مصاب)، وحسب نتيجة الاختبار الإيجابية (T^+= مصاب مخبرياً) والسلبية (T^-= غير مصاب مخبرياً)، وضعناها في الجدول التالي:

جدول (6-10): نتائج تبويب المرضى حسب حالة المريض الفعلية ونتيجة الاختبار (فرضية)

حالة المريض / نتيجة الاختبار	حالة المريض الفعلية		المجموع
	D^+= مصاب بالمرض	D^-= غير مصاب	
T^+ = الإصابة إيجابية	$a = 200$	$b = 85$	$n_1' = 285$
T^- = عدم إصابة	$c = 15$	d =700	$n_2' = 715$
المجموع	$n_1 = 215$	$n_2 = 785$	$n = 1000$

واعتماداً على الجدول السابق (6–10) نلاحظ أن حجم العينة $n = 1000$ شخص وهي تتألف من مجموعتين: مجموعة المصابين فعلاً (D^+) وحجمها $n_1 = 215$ شخصاً، ومجموعة غير المصابين (D^-) وحجمها $n_2 = 785$ شخصاً .

ولكن نتائج الاختبارات أظهرت لنا أن عدد المصابين مخبرياً T^+ يساوي $(n_1' = 285)$ شخصاً، وعدد غير المصابين مخبرياً $n_2' = 715$ شخصاً .

وبناءً على ذلك يمكننا أن نعرف الاحتمالات التالية (وهي تحسب من هوامش الجدول):

$$P(D^+) = \frac{n_1}{n} = \frac{215}{1000} = 0.215 = \tilde{P} \quad \text{احتمال أن يكون الشخص مصاباً :}$$

$$P(D^-) = \frac{n_2}{n} = \frac{785}{1000} = 0.785 = \tilde{q} \quad \text{احتمال أن يكون الشخص غير مصاب :}$$

$$P(T^+) = \frac{n_1'}{n} = \frac{285}{1000} = 0.285 \quad \text{احتمال أن تكون نتيجة اختبار إيجابية :}$$

$$P(T^-) = \frac{n_2'}{n} = \frac{715}{1000} = 0.715 \quad \text{احتمال أن تكون نتيجة اختبار سلبية :}$$

$$(79-6)$$

وهنا نلاحظ أن: $P(D^+) + P(D^-) = 1$ وأن: $P(T^+) + P(T^-) = 1$

وبناء على ذلك يمكننا أن نعرف الأرجحيات (odds) المختلفة. لذلك نضع التعريف العام لأرجحية تحقق حادث (A) خل n تجربة عليه كما يلي:

$$\frac{n_1(A)}{n_2(\bar{A})} = \text{قيمة الأرجحية للحادث}(A) = \frac{\text{عدد مرات تحقق الحادث}(A)\text{ خلال التجربة}}{\text{عدد مرات عدم تحقق الحادث}(A)\text{ خلال التجربة}}$$

ونرمز لها بالرمز odds (A) ونكتبها رياضياً كما يلي:

$$odds(A) = \frac{n_1(A)}{n_2(\bar{A})} : \quad \left[\text{ونكتب ذلك كمايلي :} n_1(A) : n_2(\bar{A})\right] \qquad (80-6)$$

حيث أن: $n_1(A) + n_2(\bar{A}) = n$

وبناء على ذلك يمكننا أن نحسب قيم الأرجحيات المختلفة لحالات المرضى ولنتائج الاختبارات من بيانات الجدول (6–10) كمايلي:

$$odds(D^+) = \frac{n_1}{n_2} = \frac{215}{785} = \frac{43}{157} \quad \text{يوجد } \mathbf{43} \text{ مريضاً مقابل كل } \mathbf{157} \text{ غير مريض :}$$

$$odds(D^-) = \frac{n_2}{n_1} = \frac{785}{215} = \frac{157}{43} \quad \text{يوجد } \mathbf{157} \text{ غير مريض مقابل كل } \mathbf{43} \text{ مريضاً :}$$

$$odds(T^+) = \frac{n_1'}{n_2'} = \frac{285}{715} = \frac{57}{143} \quad \text{هناك } \mathbf{57} \text{ نتيجة إيجابية مقابل كل } \mathbf{143} \text{ نتيجة سلبية :}$$

$$odds(T^-) = \frac{n_2'}{n_1'} = \frac{715}{285} = \frac{143}{57} \quad \text{هناك } \mathbf{143} \text{ نتيجة سلبية مقابل كل } \mathbf{57} \text{ نتيجة إيجابية :}$$

$$(81-6)$$

كما يمكننا ببساطة استخراج العلاقات التي تربط هذه الأرجحيات بالاحتمالات (6–79) السابقة. وذلك بتقسيم بسط ومقام كل أرجحية على حجم العينة (n) فنجد أن:

$$odds(D^+) = \frac{\frac{n_1}{n}}{\frac{n_2}{n}} = \frac{P(D^+)}{P(D^-)} = \frac{p}{q} = \frac{p}{1-p} = \frac{0.215}{0.785} = 0.274$$

أي أنه يوجد مقابل كل 274 مصاب بالمرض 1000 شخص غير مصاب به .

$$odds(D^-) = \frac{\frac{n_2}{n}}{\frac{n_1}{n}} = \frac{P(D^-)}{P(D^+)} = \frac{q}{p} = \frac{1-p}{p} = \frac{1}{odds(D^+)} = 3.65 \qquad (82-6)$$

أي أنه يوجد مقابل كل 365 غير مصاب يوجد 100 مصاب .

$$odds(T^+) = \frac{\frac{n_1'}{n}}{\frac{n_2'}{n}} = \frac{P(T^+)}{P(T^-)} = \frac{p'}{q'} = \frac{p'}{1-p'} = \frac{0.285}{0.715} = 0.363$$

أي أنه مقابل كل 363 نتيجة إيجابية يوجد 1000 نتيجة سلبية .

$$odds(T^-) = \frac{n_1'/n}{n_2'/n} = \frac{P(T^-)}{P(T^+)} = \frac{q}{p'} = \frac{1-p'}{p'} = \frac{1}{odds(T^+)} = 2.75$$

أي أنه مقابل كل 275 نتيجة سلبية يوجد 100 نتيجة إيجابية .

ومما سبق يمكننا استنتاج أن أرجحية أي حادث (A) ترتبط باحتمال تحققه وعدم تحققه وفق العلاقة التالية:

$$odds(A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1-P(A)} \qquad (83-6)$$

وأن نستنتج أن: $odds(\bar{A}) = \frac{1}{odds\,(A)}$ وأن: $odds(\bar{A}) = \frac{P(\bar{A})}{P(A)}$

$$odds(A) * odds(\bar{A}) = 1 \qquad (84-6)$$

كما يمكننا أن نعرّف عدداً من المؤشرات الاحصائية اعتماداً على التكرارات الداخلية والتكرارات الهامشية والمبينة في الجدول (6–10) السابق. وهذه المؤشرات هي الاحتمالات الشرطية التالية:

1–الحساسية (sensibility): وهي نسبة عدد المرضى (D^+) الذين اعتبرتهم الاختبارات إنهم مصابين بالمرض (أي كانت نتيجة الاختبار T^+ متطابقة مع الحالة الفعلية للمريض D^+) ونرمز لها بالرمز $P(T^+/D^+)$ ونحسبها من الاحتمال الشرطي التالي:

$$P(T^+/D^+) = \frac{a}{a+c} = \frac{a}{n_1} = \frac{200}{215} = 0.93023 \qquad (85-6)$$

وهو احتمال أن تكون نتيجة اختبار المريض متطابقة مع حالته المرضية (أي أن تكون نتيجة الاختبارات الإيجابية صحيحة) .

2–الخصوصية أو النوعية (specificity): وهي نسبة عدد غير المرضى (D^-) الذين اعتبرتهم الاختبارات إنهم غير مصابين (أي كانت نتيجة الاختبارات السلبية T^- متطابقة مع الحالة الفعلية للمريض D^-)، ونرمز لها بـ $P(T^-/D^-)$ ونحسبه من الاحتمال الشرطي التالي:

$$P(T^-/D^-) = \frac{d}{b+d} = \frac{d}{n_2} = \frac{700}{785} = 0.89172 \qquad (86-6)$$

وهو احتمال أن تكون نتيجة اختبار غير المريض متطابقة مع حالته الصحية (أي أن تكون نتيجة الاختبارات السلبية الصحيحة) .

3-نسبة الاختبارات الإيجابية الكاذبة (غير الصحيحة): وهي نسبة عدد غير المرضى (D^-) الذين اعتبرتهم الاختبارات مصابين بالمرض (T^+) ونرمز لها بالرمز $P(T^+/D^-)$ ونحسبها من الاحتمال الشرطي التالي:

$$P(T^+/D^-) = \frac{b}{b+d} = \frac{85}{285} = 0.10828 \qquad (87-6)$$

وهو احتمال أن تكون نتيجة اختبار غير المريض إيجابية (النتيجة غير صحيحة).

4- نسبة الاختبارات السلبية الكاذبة (غير الصحيحة): وهي نسبة عدد غير المرضى (D^+) الذين اعتبرتهم الاختبارات أنهم غير مصابين (T^-) ونرمز لها بالرمز $P(T^-/D^+)$ وتحسب من الاحتمال الشرطي التالي:

$$P(T^-/D^+) = \frac{c}{a+c} = \frac{15}{215} = 0.06977 \qquad (88-6)$$

وهو احتمال أن تكون نتيجة اختبار المريض سلبية (النتيجة غير صحيحة).

5- مصداقية الاختبارات: وهي نسبة كل الاختبارات الصحيحة (الإيجابية والسلبية) إلى المجموع الكلي للاختبارات (n) ويُرمز لها بـ AT وتحسب من العلاقة:

$$AT = \frac{a+d}{n} = \frac{a+d}{a+b+c+d} = \frac{200+700}{1000} = 0.90 \qquad (89-6)$$

6-معدل التصنيف الخاطئ: وهو نسبة كل الاختبارات غير الصحيحة إلى المجموع الكلي للاختبارات (n) ويرمز له بـ MR وتحسب من العلاقة:

$$MR = \frac{b+c}{n} = \frac{b+c}{a+b+c+d} = \frac{85+15}{1000} = 0.10 \qquad (90-6)$$

وهنا نلاحظ أن: $MR = 1 - AT$

كما يمكننا تعريف عدداً من الأرجحيات الشرطية. بناء على التكرارات الداخلية فقط المبينة في الجدول (6-10) السابق وهذه الأرجحيات هي:

$$odds(T^+/D^+) = \frac{a}{c} = \frac{200}{15} = \frac{40}{3} \qquad (91-6)$$

أي أنه يوجد 40 نتيجة إيجابية مقابل كل 3 نتائج سلبية وذلك عند مجموعة المرضى D^+

$$odds(T^+/D^-) = \frac{b}{d} = \frac{85}{700} = \frac{17}{140}$$

أي أنه يوجد 17 نتيجة إيجابية مقابل كل 140 نتيجة سلبية عند مجموعة غير المرضى D^- .

$$odds(T^-/D^+) = \frac{c}{a} = \frac{15}{200} = \frac{3}{40}$$

أي يوجد 3 نتائج سلبية مقابل كل 40 نتيجة إيجابية عند مجموعة المرضى D^+ .

$odds(T^-/D^-) = \frac{d}{b} = \frac{700}{85} = \frac{140}{7}$ يوجد 140 سلبية مقابل كل 7 إيجابية عند غير المرضى

$odds(D^+/T^+) = \frac{a}{b} = \frac{200}{85} = \frac{40}{17}$ يوجد 40 مريض مقابل كل 17 في التحاليل الإيجابية

$odds(D^+/T^-) = \frac{c}{d} = \frac{15}{700} = \frac{3}{140}$ يوجد 3 مريض مقابل كل 140 في التحاليل السلبية

$odds(D^-/T^+) = \frac{b}{a} = \frac{85}{200} = \frac{17}{40}$ يوجد 17 غير مريض مقابل كل 40 في التحاليل الإيجابية

$odds(D^-/T^-) = \frac{d}{c} = \frac{700}{15} = \frac{140}{3}$ يوجد 140 غير مريض مقابل كل 3 في التحاليل السلبية

وأخيراً نعرف المؤشرات الهامة التالية:

7- نسبة الأرجحية (odds Ratio): وتعرف للمرضى ككل ونرمز لها بـ OR وتحسب من العلاقة التالية:

$$OR = \frac{odds(D^+/T^+)}{odds(D^+/T^-)} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a*d}{b*c} = \frac{200*700}{85*15} = 109.80 \qquad (92-6)$$

أما نسبة الأرجحية لغير المرضى فتحسب من العلاقة :

$$\overline{OR} = \frac{odds(D^-/T^+)}{odds(D^-/T^-)} = \frac{\frac{b}{a}}{\frac{d}{c}} = \frac{b*c}{a*d} = \frac{1}{OR} = 0{,}009107 \qquad (93-6)$$

8-نسبة المخاطرة (Risk Ratio): وهي نسبة احتمال الحادث المطلوب على احتمال الحادث المتمم له ونرمز لها بـ RR وتحسب لاختبارات المرضى من العلاقة :

$$RR = \frac{P(D^+/T^+)}{P(D^+/T^-)} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{ac+ad}{ac+bc} = \frac{200*15+200*700}{200*15+85*15} = 33.45 \qquad (94-6)$$

وعندما يكون الحد $(a*c)$ صغيراً بالنسبة لـ $(a*d)$ فيمكن اهماله وتصبح نسبة المخاطرة المقربة كما يلي:

$$\widetilde{RR} \approx \frac{ad}{bc} \approx OR \qquad (95-6)$$

9- المخاطرة المطلقة لاختبارات المرضى وتحسب من العلاقة :

$$AR = P(D^+/T^+) - P(D^+/T^-) = \left[\frac{a}{a+b} - \frac{c}{c+d}\right] = 0.68078 \qquad (96-6)$$