Compare the classification of the quadratic function and the logistic function
مقارنة تصنيف الدالة التربيعية والدالة اللوجستية

أ. م. د. شروق عبد الرضا سعيد السباح مريم مهدي عناد الخزعلي

جامعة كربلاء/كلية الإدارة والاقتصاد/ قسم الإحصاء
(بحث مستل من رسالة ماجستير)

الخلاصة:-

تم تطبيق دالة التمييز لتشخيص ثلاث أنواع من الأورام السرطانية وهي سرطان الثدي وسرطان العظم وسرطان الرئة لكثرة شيوعها في مجتمعنا حالياً، ولغرض دراسة هذا الموضوع تم تسجيل قيم المشاهدات لخمس متغيرات وهي (الجنس، العمر، مهنة المريض، حالة خروج المريض، فترة بقاء المريض في المستشفى) من عينة عشوائية بسيطة بحجم (270 مريض) درست في نموذجين وهي دالة التمييز التربيعية ودالة التمييز اللوجستية، واستخرجت النتائج باستعمال البرنامج الاحصائي **Stata**، وتم استعمال معيار خطأ التصنيف كمعيار للمقارنة بين النماذج، بهدف الوصول إلى أفضل نموذج باقل خطأ تصنيفي ممكن، وتم الوصول الى ان الدالة اللوجستية ذات تفوق على الدالة التربيعية من حيث قلة نسبة التصنيف الخاطئ.

Abstract:-

The function of discrimination has been applied to diagnose three types of tumors, namely breast cancer, bone cancer and lung cancer, for their prevalence in our society. For the purpose of this study, the values of observations were recorded for five variables (sex, age, occupation, prognosis and duration of hospitalization) the random sample of size (270 patients) studied in two models: Quadratic discrimination function and Logistic discrimination function. The results were extracted using the statistical program **Stata**. With a view to reaching Li best model the lowest rank possible error, was reached that Logistic discrimination functionis superior to Quadratic discrimination function in terms of the lack of proportion of the wrong classification**.**

المقدمة ومشكلة البحث *Introduction and Problm*

يعد التمييز بين المشاهدات من الأساليب الشائعة الاستعمال وذلك لكثرة الظواهر التطبيقية التي يمكن أن يتم تحليلها من خلال أسلوب التمييز.

إذ إن دالة التمييز الخطية المسندة على تركيب خطي للمتغيرات لكي تكون مثلى يجب أن تنتج اصغر احتمال لخطأ التصنيف علما بان هناك افتراضات يجب بتوفرها حول البيانات المستخدمة في التحليل وهي أن يكون متجه المتغيرات التوضيحية ذا توزيع طبيعي متعدد المتغيرات ومصفوفات تباينات مشتركة متساوية ومتجهات متوسطات مختلفة في كل مجموعة من المجاميع؛ ولكن أحيانا نواجه اختلال في بعض الفرضيات وبالتالي تفقد دالة التمييز الخطية بعض خواصها المثلى. فعند عدم تحقق ثبات مصفوفة التباين والتباين المشترك يكون من الضروري استعمال دالة التمييز التربيعية (Quadratic Discriminate)، فضلاً عن ذلك ان دالة التمييز الخطية تفقد خواصها المثلى عندما تكون جميع او بعض المتغيرات الاضافية ذات طبيعة ثنائية او مصنفة فتكون دالة التمييز اللوجستية (Logistic Discrimination) البديل الامثل للتمييز الطبيعي. أما عملية التصنيف(Classification)هي العملية اللاحقة بعد تكوين الدالة المميزة حيث يتم الاعتماد على هذه الدالة بالتنبؤ وتصنيف المفردة الجديدة لإحدى المجموعات قيد الدراسة بأقل خطأ تصنيف ممكن.

فرضية البحث ***Search Hypothesis***

$\boldsymbol{H_0}$: عدم وجود مشاكل في النموذج الاحصائي

$\boldsymbol{H_1}$: وجود مشاكل في النموذج الاحصائي

هدف البحث ***Research Aim***

يهدف البحث إلى الوصول إلى أفضل تصنيف لبعض أنواع الأورام السرطانية على أساس معيار احتمال اقل خطأ تصنيفي ممكن باستعمال دالة التمييز التربيعية ودالة التمييز اللوجستية.

الجانب النظري

1. دالة التمييز التربيعية ***Quadratic Discriminant Function***

تعد هذه الطريقة من الطرق المعلمية إذ يستعمل هذا النوع من الدوال في حالة عدم تساوي مصفوفة التباين والتباين المشترك للمجموعات ويكون المجتمع قيد الدراسة ذا توزيع طبيعي متعدد المتغيرات وبمتجهات متوسطات مختلفة.

وعليه فان مقياس التمايز V سيكون كالآتي[10]:

$$V = \frac{f_1(x_i)}{f_2(x_j)} \neq 1 \qquad (1)$$

إذا كانت $V > 1$ المشاهدة $\boldsymbol{x}$ تعود إلى المجموعة الأولى و إذا كانت $V < 1$فان المشاهدة x تعود إلى المجموعة الثانية وعشوائيا عدا ذلك.

وبأخذ اللوغاريتم للطرفين نحصل على الآتي:

$$G = \ln V = \ln f_1(x_i) - \ln f_2(x_j) = 0 \qquad (2)$$

وفي حالة وجود P من المتغيرات $(X_1, X_2 \ldots\ldots, X_P)$ لكل مجموعة من المجموعتين فان مقياس التمييز يكون :

$$G = \ln f_1(x_1, \ldots., x_P) - \ln f_2(x_1, \ldots., x_P) \qquad (3)$$

وفي حالة إن المعالم الدالة الاحتمالية غير معلومة فسوف تكون دالة التمييز التربيعية التقديرية g والتي قدرت معالمها بطريقة الإمكان الأعظم كالآتي[13] :

$$g = \hat{G} = \frac{1}{2} log \frac{|S_1|}{|S_2|} - \frac{1}{2}\left(\overline{\underline{X}}_1' S_1^{-1} \overline{\underline{X}}_1 - \overline{\underline{X}}_2' S_2^{-1} \overline{\underline{X}}_2\right) + \underline{X}'\left(S_1^{-1}\overline{\underline{X}}_1 - S_2^{-1}\overline{\underline{X}}_2\right) - \frac{1}{2}\underline{X}'(S_1^{-1} - S_2^{-1})\underline{X} \qquad (4)$$

وسيتم تصنيف المشاهدات الجديدة بأنها تعود إلى المجموعة الأولى إذا كانت $g > 0$ وإلى المجموعة الثانية إذا كانت $g < 0$ وعشوائيا عدا ذلك.

2. اشتقاق دالة التمييز التربيعية [16], [15]

تكون الفرضية في حالة وجود مجموعتين من مجتمعين فرضية خاصة تمثل انتماء كل مشاهدة من المشاهدات إلى إحدى المجموعتين وهي:

$\boldsymbol{H_0}: \underline{X} \in f_2(\mathrm{x})$

$\boldsymbol{H_1}: \underline{X} \in f_1(\mathrm{x})$

وباستعمال أسلوب بيز لتقليل الخطورة التي تمثل نسبة الإمكان الأعظم والاحتمالات الأولية P_1و P_2 وكلف التصنيف الخاطئ c_{12} و c_{21} تكون :

$$\frac{f_1(\mathrm{x})}{f_2(\mathrm{x})} < \frac{c_{12}P_2}{c_{21}P_1} \quad (5)$$

وعند امتلاك كل من f_1 و f_2 توزيعاً طبيعياً و اخذ اللوغاريتم للطرفين ينتج:

$$ln\frac{|\Sigma_1|}{|\Sigma_2|} + \left(\underline{x}_{11} - \underline{\mu}_1\right)' \Sigma_1^{-1}\left(\underline{x}_{1i} - \underline{\mu}_1\right) - \left(\underline{x}_{2i} - \underline{\mu}_2\right)' \Sigma_2^{-1}\left(\underline{x}_{2i} - \underline{\mu}_2\right) < -2ln\frac{c_{12}P_2}{c_{21}P_1} \quad (6)$$

وبعد التبسيط تصبح دالة التمييز كما يأتي :

$$h(x) = \underline{x}'A\underline{x}b'x + c \ <> T \qquad (7)$$

إذ إن:

$$A = \Sigma_1^{-1} - \Sigma_2^{-1}$$
$$b = 2\left(\Sigma_2^{-1}\underline{\mu_2} - \Sigma_1^{-1}\underline{\mu_1}\right)$$
$$c = \left(\underline{\mu_1}'\Sigma^{-1}\underline{\mu_1} - \underline{\mu_2}'\Sigma_1^{-1}\underline{\mu_2} + ln\frac{|\Sigma_1|}{|\Sigma_2|}\right)$$
$$T = -2ln\frac{c_{12}P_2}{c_{21}P_1}$$

إذ تصنف المشاهدة x على إنها تنتمي للمجتمع f_1 إذا كانت $h(\underline{x}) < T$ وتنتمي إلى المجتمع f_2 إذا كانت $h(\underline{x}) > T$ وتعرف بدالة التمييز التربيعية.

أما عند تساوي مصفوفة التباين والتباين المشترك Σ_i تصبح الدالة بالشكل الآتي:

$$\mathrm{h}^*(x) = b'^*x + c^* <> T(8)$$

إذ أن:

$$b^* = 2\Sigma\left(\underline{\mu_2} - \underline{\mu_1}\right)$$
$$c^* = \underline{\mu_1}'\Sigma^{-1}\underline{\mu_1} - \underline{\mu_2}\Sigma^{-1}\underline{\mu_2}$$

وهي دالة التمييز الخطية.
الرمز <> يعني تصنيف المشاهدة x على أنها تنتمي إلى المجتمع المميز f_1 إذا كانت قيمتها اقل من T وتنتمي إلى المجتمع f_2 إذا كانت قيمتها اكبر من T.
وعليه تكون كلف التصنيف الخاطئ تساوي دائما واحد ، وان خطورة بيز تتحول إلى احتمال خطا التصنيف لذلك فان الحل الذي يقلل من خطورة بيز يقلل أيضاً من احتمال خطا التصنيف ولهذا فإن :

$$P_1f_1(\underline{x}) < P_2f_2(\underline{x}) \qquad (9)$$

وعند التعميم واخذ مجموعات متعددة تكون الفرضية:

$$\boldsymbol{H_0}: \underline{x} \in w_i \qquad i = 1, \ldots\ldots, g$$

وفي حالة الكلف تساوي واحد يكون الحل باختبار المجموعة التي تحقق ما يأتي كمجموعة تنتمي إليها المفردات$\underline{x}$:

$$P_kf_k > P_if_i \qquad (10)$$
$$for\ i \neq k \ \ , i = 1, \ldots\ldots, g$$

وبعبارة أخرى فان لدينا g من الدوال التمييزية $h_i(\underline{\mathrm{x}})$ والحل هو اختبار الدالة التي تعطي أكبر قيمة ، وكما يأتي:

$$h_{\mathrm{k}}(\underline{\mathrm{x}}) = \max h_i(\underline{\mathrm{x}}) \qquad (11)$$

إذ أن:

$$h_i(\underline{\mathrm{x}}) = P_if_i(\underline{\mathrm{x}})$$

وتوجد نظرية بهذا الخصوص تنص على إن دالة متزايدة رتيبة $h_i(\underline{\mathrm{x}})$ يمكن أن تحل بدلا من $h_i(\underline{\mathrm{x}})$ وفي حالة توزيع البيانات طبيعيا يكون [15]:

$$h_i(\underline{x}) = P_iN\left(\underline{x_i}\underline{\mu_1}, \Sigma_i\right) \qquad (12)$$

إذ إن :
$N(\underline{x_i}\underline{\mu_1}, \Sigma_i)$: هي دالة التوزيع الطبيعي متعدد المتغيرات
وفي حالة تحقق التوزيع الطبيعي فان الصيغة $h_i(\underline{x})$ تكون :

$$h_i(\underline{x}) = -\left(\underline{x} - \underline{\mu_i}\right)' \Sigma_1^{-1}\left(\underline{x} - \underline{\mu_i}\right) - ln|\Sigma_i| + 2lnP_i \qquad (13)$$

$i = 1, \ldots\ldots, g$

وهي دالة التمييز التربيعية المتعددة (*Multiple Quadratic Discriminant Function*)
وعند تساوي مصفوفة التباين والتباين المشترك تصبح الدالة بالشكل الآتي:

$$h_i^*(\underline{x}) = 2\underline{\mu}_i'\Sigma^{-1}\underline{x} - \underline{\mu}_i\Sigma^{-1}\underline{\mu}_i + 2lnP_i \ , i = 1, \ldots\ldots, g \qquad (14)$$

وهي دالة التمييز الخطية .

3. تقدير معالم النموذج [10],[12]

يتم تقدير Σ_i و$\hat{\Sigma}_i$ بطريقة الإمكان الأعظم لوجود خاصية الثبات إذ إن :

$$\hat{\mu}_i = \overline{\underline{x}}_i = \frac{1}{n_i}\sum_{j=i}^{ni} \underline{x}_{ij} \qquad (15)$$

$$\hat{\Sigma}_i = S_i^2 = \frac{1}{n_i}\sum_{j=i}^{ni} (\underline{x}_{ij} - \overline{x}_j)(\underline{x}_{ij} - \overline{x}_j)^2 \qquad (16)$$

ويمكن استبدال تقدير Σ_i المتحيز بتقدير Σ_1^{-1} الغير متحيز المشتق من توزيع ***(Wishert)*** وهو :

$$\hat{\Sigma}_1^{-1} = \left(\frac{ni - P - 3}{ni}\right) S_i^{-1} \qquad (17)$$

إذ إن S_i هي تقدير الإمكان الأعظم لـ Σ_1^{-1} *وكذلك بدلا من* Σ_i يمكن استعمال التقدير غير متحيز لمسافة مهالانوبس .

$$\hat{D}^2{}_{ij} = \frac{n_i + n_j - p - 3}{n_i + n_j - p - 2}\left(\overline{\underline{x}}_j - \overline{\underline{x}}_i\right)' S^{-1}\left(\overline{\underline{x}}_j - \overline{\underline{x}}_i\right) - \frac{p(n_i - n_j)}{n_i n_j} \qquad (18)$$

وإذا كانت $\hat{D}^2{}_{ij}$ سالبة فتؤخذ صفراً .
اما بالنسبة لتقدير الاحتمالات الأولية فتقديرها يكون:

$$\hat{P}_i = p_i = \frac{n_i}{n} \qquad (19)$$

إذ أن:
n : حجم العينة الكلي
n_i : حجم العينة للمجموعة*i*.

4. دالة التمييز اللوجستي ***Logistic discriminant Function***

يعد التمييز اللوجستي من بين الكثير من الأساليب المقترحة في التمييز الإحصائي للطرق المعلمية المستندة على التوزيع الطبيعي متعدد المتغيرات ، والطرق اللامعلمية حرة التوزيع كطريقة الجار الأقرب (*Nearest Neighbor*) لهذا السبب غالبا ما تدعى هذه الطريقة بالمعلمية الجزئية أو شبه معلمية . وفقاً لذلك فإن النموذج اللوجستي احد الأدوات الأكثر جاذبية واستعمالا في حل مسائل التمايز والانحدار اللذان يسيران نحو إحدى أو كلا الاتجاهين الآتيين :

1. تلخيص ووصف الفروقات بين المجتمعات.
2. تحديد موقع مفردات جديدة إلى مجاميعها أي التنبؤ بقيم متغير الاستجابة لمفردات جديدة.

ان هذه الطريقة من الطرائق المهمة وذلك لأنها تعالج مشاكل التمييز في حالة إن تكون المتغيرات (المقاييس) متقطعة أو مستمرة أو مختلطة فضلا عن سهولتها في الاستخدام إذ أن عدد حدودها يساوي عدد حدود الدالة الخطية. وهناك نوعان من النموذج اللوجستي وهي (1) النموذج اللوجستي ثنائي الاستجابة (2) النموذج اللوجستي متعدد الاستجابة وسيكون تركيزنا على النوع الثاني .

مميزات النموذج اللوجستي[5]

1. يتطلب القليل من الافتراضات حول التوزيع، فهو لا يفترض وجود علاقة خطية بين المتغير التابع والمتغيرات التوضيحية.
2. سهولة استخدامه، فعندما يتم تقدير معالم النموذج فان تحديد موقع مفردة جديدة يتطلب حساب دالة خطية واحدة فقط.
3. تكون المعالم المقدرة من النموذج ذات تفسير مباشر كدالة خطية للوغاريتم نسبة الارجحية .
4. المتغير التابع يجب ان يكون ثنائي التفرع بحيث يحتوي على فئتين (مصابين او غير مصابين على سبيل المثال)
5. يجب ان تكون الفئات محددة وشاملة بحيث ان كل مفردة تنتمي إلى فئة واحدة فقط .
6. النموذج اللوجستي لا يشترط ان تكون المتغيرات التوضيحية من النوع المستمر ولا تتبع التوزيع الطبيعي ولا ان تكون العلاقة بين المتغير التابع والمتغيرات التوضيحية خطية ولا يفترض تساوي التباين ضمن كل فئة . وهذا يجعل النموذج اللوجستي اكثر مرونة من بقية نماذج التنبؤ والتصنيف .
7. يجب ان يكون حجم العينة المستخدم في النموذج اللوجستي اكبر من حجم العينة المستخدم في النموذج الخطي، لأن معاملات نموذج اللوجستي يتم تقديرها باستخدام طريقة الامكان الاعظم وهي طريقة تحتاج إلى عينة كبيرة الحجم نسبياً .

5. نموذج اللوجستي ثنائي الاستجابة[8],[14]

يطبق هذا النموذج عندما يأخذ المتغير العشوائي**Y**قيمتين فقط قد تدل على نمط الاستجابة لمؤثر ما (كالأدوية والمنشطات والسموم) . في هذه الحالة تأخذ**Y** القيمتين0 و1 تشير إلى عدم حدوث وحدوث الاستجابة على التوالي . كذلك يستعمل في مسائل التشخيص الطبي والتمييز بين أنواع الإمراض. فالأورام السرطانية مثلا قد تكون خبيثة أو حميدة.

تعد هذه الطريقة من الطرائق المهمة وذلك لأنها تعالج مشاكل التمييز في حالة ان تكون المتغيرات مستمرة أو متقطعة أو مختلطة فضلا عن سهولتها في الاستعمال إذ إن عدد حدودها يساوي عدد حدود الدالة الخطية، يتكون نموذج اللوجستي من متغير الاستجابة Y_iيتأثر بمجموعة من المتغيرات الموضحة $(Explanatory\ variables)$ $(X_i's)$ على وفق علاقة تحتوي على مجموعة من المعالم $(b_i's)$ إذ أن متغير الاستجابة Y_i في التجارب الحياتية ثنائية الاستجابة الذي يعد نموذج اللوجستي احدها يتوزع حسب توزيع برنولي اي ان له مستويان هما الصفر والواحد.

$$Y_i \sim b(1, P_i)$$

إذ إن:

P_i :هو احتمال الاستجابة عندما $(Y = 1)$

$1 - P_i$:احتمال عدم الاستجابة عندما $(Y = 0)$

يعتمد P_i على مجموعة من المتغيرات التوضيحية$(X_i's)$ على وفق الصيغة الآتية التي يطلق عليه انموذج دالة الاستجابة اللوجستية :

$$P_i = \frac{e^{-x_i B}}{1 + e^{-x_i B}} \qquad (20)$$

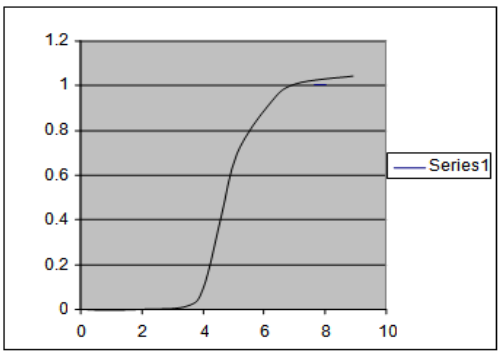
$$1 - P_i = \frac{1}{1 + e^{-x_i B}} \qquad (21)$$

اذ ان :

B: معالم مجهولة يتمتقديرها.

x_i : تمثل المتغيرات التوضيحية.

وان الصيغة (20) تسمى دالة الاستجابة اللوجستية ولها مخطط في **الشكل (1)** :



الشكل (1)يوضح العلاقة بين احتمال الاستجابة (P_i) والمتغير الموضح (x_i)

من **الشكل (1)**نلاحظ عدم خطية العلاقة بين المتغير الموضح x_i واحتمال الاستجابة P_i،ولغرض اجراء التحويل الخطي للدالة اللوجستية تمكن الباحث (***Berkson***) عام (1944) من تحويل العلاقة بين المتغيرات (x_i) ومتغير الاستجابة (P_i) إلى علاقة خطية وذلك برسم Log P بدلاً من Probit P مقابل (X) كالآتي [4].

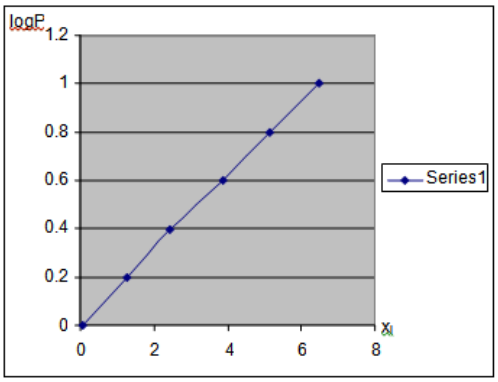
$$L_i = ln\left(\frac{P_i}{1 - P_i}\right) = log\ P_i = x_i b \qquad (22)$$

إذ إن :

$i = 1,2, \ldots\ldots, n$

L_i : يمثل تحويل (log) للاحتمال (P_i).

والصيغة (22) تبين التحويل الخطي (log) كما في **الشكل (2)** :



الشكل (2)يوضح العلاقة الخطية بين $logP_i$وx_i

$$L_i = ln\left(\frac{P_i}{1-P_i}\right) = b_0 + b_1x_1 + \ldots\ldots\ldots + b_Px_P \qquad (23)$$

اذ تكون$\underline{b}$ موجه عمودي من رتبة [(P+1)×1]للمعالم المطلوب تقديرها

$$\underline{b} = \begin{pmatrix} b_0 \\ b_1 \\ . \\ . \\ b_P \end{pmatrix}$$

$\underline{X}$ مصفوفة من رتبة [g×(p+1)] للمتغيرات التوضيحية تكون :

$$\underline{X} = \begin{pmatrix} 1 & X_{11} \cdots & X_{1P} \\ 1 & X_{21} \cdots & X_{2P} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & X_{g1} \cdots & X_{gP} \end{pmatrix}$$

وبما ان التوزيع هنا هو برنولي ولحالتين فقط عندئذ يكون [9].

$$\underline{L_i} = \underline{P_i} = \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 0 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}$$

$$\underline{L_i^*} = ln\left(\frac{P_i + \frac{1}{2n}}{(1-P_i) + \frac{1}{2n}}\right) \qquad (24)$$

إذ استعمل هنا معامل تصحيح الاستمرارية $\frac{1}{2n}$ حيث اعطى نتائج اكثر دقة وبخطأ اقل[2].

ونستعمل بعد ذلك طريقة المربعات الصغرى العامة *Generalized Least Squares* (GLS) التقديريةوهي:

$$\underline{b}^* = (X'X)^{-1}X'\underline{L_i^*} \qquad (25)$$

$$\underline{\hat{L}_i} = \hat{b}_0 + \hat{b}_1X_1 + \ldots\ldots\ldots + \hat{b}_PX_P \qquad (26)$$

وسوف يتم تصنيف المشاهدات الجديدة بأنها تعود إلى المجموعة الأولى إذا كانت$\underline{\hat{L}_i}$ اكبر من الصفر وإلى المجموعة الثانية اذا كانت $\underline{\hat{L}_i}$ اصغر من الصفر وعشوائياً عدا ذلك.

6. نموذج اللوجستي متعدد الاستجابة

يعد نموذج اللوجستي متعدد الاستجابة امتداداً طبيعياً لنموذج ثنائي الاستجابة . وهي الحالة التي يكون فيها المفردة أو المشاهدة محددة بواحدة من g من القيم الممكنة تدعى بفئات أو مجاميع الاستجابة . فقد تكون ذات طبيعة ترتيبية مثل درجة خطورة أو شدة مرض معين . كما يمكن أن تكون مجاميع الاستجابة ذات طبيعة اسمية كأصناف الدم مثلاً أو نوع واسطة النقل المفضلة في السفر (سيارة ، طائرة ، قطار ،.....).

نفرض وجود مجتمع كبير يحتوي على g من المجاميع المختلفة $(\pi_1, \dots, \pi_g)$ وسحبت عينة عشوائية منه . فأن عدد المفردات الداخلة ضمن العينة وتنتمي إلى المجموعة π_s يتبع توزيع متعدد الحدود *Multinomial Distribution* وان احتمال وقوع Y من المشاهدات ضمن العينة يكون كالآتي[11]:

$$P(Y_1 = y_1, \dots, Y_g = y_g) = \binom{n}{y} P_1^{y_1}, \dots \dots, P_g^{y_g} \qquad (27)$$

إذ إن:

$(P_1, \dots, P_g)$: تمثل نسبة المجاميع في المجتمع

$$\binom{n}{y} = \frac{n!}{y_1!\ y_2! \dots . y_p!} \qquad (28)$$

نستنتج من ذلك إن التوزيع متعدد الحدود يتولد كنتيجة لأسلوب المعاينة. فإذا رمزنا لاحتمال حدوث الاستجابة j بـ $P_j(j = 1, \dots, g)$ فعند قيم المشاهدة i ذات متجه الخصائص $x_i = (x_{i1}, \dots, x_{iP})$ يكون النموذج بالشكل الآتي:

$$P_j(x_i) = \frac{\exp(b_j' x_i)}{\sum_{j=1}^{g} \exp(b_j' x_i)}, j = 1, \dots, g-1 \qquad (29)$$

ويكون $P_g = 0$ إذ يمثل g مجموعة الأساس وb_jيمثل متجه المعالم الخاصة بالمجموعة j إذ يوجد g-1 من المتجهات المختلفة ضمن النموذج . وبذلك فان النموذج يحتوي على $(g\text{-}1)(p+1)$ من المعالم المجهولة و يمكن كتابة النموذج بصيغة لوغاريتم نسبة الإمكان الأعظم كما يلي :

$$log\left(\frac{P_j(x_i)}{P_g(x_i)}\right) = b_j' x_i \quad , j = 1, \dots, g-1 \qquad (30)$$

7. تقدير معلمات النموذج اللوجستي[11],[3]

لاحظ ***Anderson* (1972)** إن تطبيق النموذج اللوجستي يخضع للمعادلة الآتية :

$$log\left(\frac{P_j(x_i)}{P_g(x_i)}\right) = b_0^{\circ} + b'x \qquad (31)$$

وهو تركيب خطي للمتغير x يناسب المتغيرات الثنائية والثلاثية عندما x يأخذ القيم (0 ، 1 ، 2) ، بافتراض إن نسبة احتمال اللوغاريتمي هو تركيب خطي للمتغير x .

من اجل تطبيق النموذج اللوجستي يجب تحويل المتغير x إلى متغيرين ثنائيين x_1و x_2 ،

إذ إن $x = 0$إذا وفقط إذا كان$x_1 = 0$و$x_2 = 0$؛ $x = 1$إذا وفقط إذا كان$x_1 = 1$و$x_2 = 0$؛ و$x = 2$إذا وفقط إذا$x_1 = 0$ و$x_2 = 1$على نحو مماثل، يمكن استبدال كل متغير مع أكثر من$r \geq 2$مستويات من المتغيرات الثنائية $r - 1$.

وان النموذج اللوجستي في تحليل التمايز يكون :

$$\mathcal{T}_1(x) = \frac{\exp(b_0^{\circ} + b'x)}{1 + \exp(b_0^{\circ} + b'x)} \qquad (32)$$

إذ إن $\mathcal{T}_1(x)$ تمثل الاحتمالات اللاحقة لعضوية المجموعة من خلال نمذجة العلاقة بين متغير الاستجابة الفئوية الذي نرمز له بالرمز G وبين متجه المتغيرات التوضيحية X .

علماً أن متغير الاستجابة G هو متجه من المجموعات أو الفئات المختلفة $(G_1, \dots, G_g)$.

وعليه يكون نموذج لوغاريتمي خطي للاحتمالات اللاحقة للمتغير مستقل مع المتغير x مقارنة مع متغير أساس x_0 هو :

$$log\left[\frac{\mathcal{T}_1(x)}{\mathcal{T}_2(x)} \Big/ \frac{\mathcal{T}_1(x_0)}{\mathcal{T}_2(x_0)}\right] = b'(x - x_0) \qquad (33)$$

في كثير من الأحيان المرض قيد الدراسة يكون نادراً إلى حد ما ، بحيث $\mathcal{T}_2(x)$ و $\mathcal{T}_2(x_0)$ قريبة من الواحد ، وبالتالي نسبة الفرق تقريبا تكون :

$$\mathcal{T}_1(x)/\mathcal{T}_1(x_0) \quad (34)$$

ومن خلال التحويل اللوجستي نحصل على :

$$logit[\mathcal{T}_1(x)] = log[\mathcal{T}_1(x)/\mathcal{T}_2(x)] = b_0 + b'x \quad (35)$$

ولربط x بالاحتمال اللاحق $\mathcal{T}_1(x)$ يكون التمييز الاحتمالي بالصيغة الآتية :

$$probit[\mathcal{T}_1(x)] = \Phi^{-1}[\mathcal{T}_1(x)] \quad (36)$$

إذ إنΦ هي دالة التوزيع الطبيعي القياسي .
عملياً يمكن استعمال المعادلة الآتية للنماذج اللوجستية والاحتمالية وهي :

$$logit[\mathcal{T}_1(x)] \cong c\, \Phi^{-1}[\mathcal{T}_1(x)] \quad (37)$$

إذ إن $c = \sqrt{8/\pi} = 1.6$.

8. عملية التصنيف *Classification*

إن عملية التصنيف (Classification) هي العملية اللاحقة بعد تكوين دالة التمييز حيث يتم الاعتماد على هذه الدالة بالتنبؤ وتصنيف المفردة الجديدة لإحدى المجموعات قيد الدراسة بأقل خطأ تصنيف ممكن، ويشترط تساوي التباينات للمجموعات قيد البحث، وهناك تمييز خطي في حالة مجموعتين، وتمييز خطي في حالة أكثر من مجموعتين، أما التمييز غير الخطي فيستخدم في حالة عدم تساوي التباينات[1].

احتمال خطأ التصنيف

إن خطأ التصنيف هو عامل مهم لإثبات كفاءة الدالة التمييزية، أي أن الدالة التمييزية التي تعطي أقل خطأ تصنيف هي الدالة الأكثر كفاءة وتكون الأفضل من بين دوال التمييز.

وهناك نوعين من احتمال خطأ التصنيف[7]:

أولاً: احتمال خطأ التصنيف P_{ij} وهو احتمال تصنيف المشاهدة إلى المجموعة **j** وهي تعود إلى المجموعة i .

$$P_{ij} = \Phi\left(-\frac{\delta}{2}\right) \quad (38)$$

ثانياً: احتمال خطأ التصنيف P_{ji} وهو احتمال تصنيف المشاهدة إلى المجموعة i وهي تعود إلى المجموعة j.

$$P_{ji} = \Phi\left(-\frac{\delta}{2}\right) \quad (39)$$

إذ أن:
Φ : تمثل دالة التوزيع الطبيعي القياسي .
δ : تمثل جذر مقياس مسافة مهالانوبسD^2 .
إذن :

$$\delta^2 = D^2 = (\overline{X}_1 - \overline{X}_2)'\, S^{-1}(\overline{X}_1 - \overline{X}_2)$$

وسيكون احتمال خطأ التصنيف كالآتي:

$$P_{ij} = P_{ji} = \Phi\left(-\frac{D}{2}\right) \quad (40)$$

إذ إن D هو جذر مقياس مهالانوبس .

الجانب التطبيقي

تم جمع البيانات من خلال اخذ عينة عشوائية بسيطة حجمها **270** من طبلات المرضى الراقدين في مستشفى (مدينة مرجان الطبية) في محافظة بابل لإمراض السرطان للتمييز بين اشخاص مصابين بأورام الثدي والعمود الفقري (سرطان العظم) والجهاز التنفسي (سرطان الرئة) وذلك لكثرة المصابين بهذه الانواع خلال عام 2016، ولغرض تحليل البيانات تم استعمال البرنامج الاحصائي Stata وكذلك الحزمة الاحصائية SPSS، وتم دراسة المتغيرات التالية لكل مجموعة:(الجنس، العمر، مهنة المريض، تشخيص المرض، حالة خروج المريض، فترة بقاء المريض في المستشفى).

تعريف المتغيرات ***Definition of variables***

اعتمدنا في تكوين دالة التمييز على عدد من المتغيرات تم جمعها عن كل مشاهدة (مريض) من مشاهدات العينة ، وحيث إن جزء من المتغيرات هي متغيرات متقطعة وأخرى متصلة .

أ- المتغير المتعمد (Y) والذي يمثل نوع أو تشخيص المرض:

سرطان الثدي =1 سرطان العظم =2 سرطان الرئة =3

ب-المتغيرات التوضيحية تمثل ما يلي:

1- $\mathbf{X_1}$ = الجنس: شملت العينة كلا الجنسين وكانت نسبتهم إلى مجموع العينة متفاوتة ويمثل متغير ثنائي (ذكر=1 ، أنثى=2)

2- $\mathbf{X_2}$ = العمر: تتراوح أعمار المرضى المشمولين ضمن العينة بين (4 – 95) سنة.

3- $\mathbf{X_3}$= مهنة المريض: أعطينا رموز لكل مهنة وذلك لتسهيل تحليل البيانات

(طفل=1، ربة بيت أو كاسب=2 ، طالب =3 ، عاجز=4 ، متقاعد=5 ، موظف =6)

4- $\mathbf{X_4}$ =حالة خروج المريض :(وفاة=1 ، إحالة=2 ، خرج على مسؤوليته=3 ، تحسين=4)

5- $\mathbf{X_5}$= فترة بقاء المريض في المستشفى.

اختبار التوزيع الطبيعي لكل مجتمع

نختبر البيانات لمعرفتها فيما إن كانت المتغيرات التوضيحية للمجموعات الثلاث لإمراض السرطان تتوزع طبيعي أم لا باستخدام (***Kolmogorov-Smirnov***) وبموجب البرنامج الإحصائي SPSS وحسب الفرضية التالية:

$\mathbf{H_0}$: البيانات تتبع التوزيع الطبيعي

$\mathbf{H_1}$: البيانات لا تتبع التوزيع الطبيعي

الجدول (1) نتائج اختبار البيانات للتوزيع الطبيعي

Variables	Kolmogorov-Smirnov	
	statistic	Sig.
الجنس (X_1)	0.40	.000
العمر (X_2)	0.14	.000
المهنة(X_3)	0.34	.000
حالة خروج المريض (X_4)	0.48	.000
فترة بقاء المريض (X_5)	0.17	.000

في **الجدول(1)** أظهرت نتائج قيم اختبار (***Kolmogorov-Smirnov***)إن مستوى المعنوية لجميع المتغيرات اقل من **0.05**المستوى المعتمد في هذه الدراسة وعليه نرفض فرضية العدم القائلة إن البيانات تتبع التوزيع الطبيعي.ولكون حجم البيانات **270** مشاهدة يمكن إن نعّد إن البيانات تقترب من التوزيع الطبيعي حسب نظرية (الغاية المركزية).

حساب احتمال التصنيف الصحيح

The probability of correct classification

إن عملية التصنيف قد تؤدي الى الوقوع فيما يعرف بخطأ التصنيف (*misclassification*) وهو احتمال تصنيف مفردة معينة الى المجموعة الأولى بينما هي في الحقيقة تعود للمجموعة الثانية أو الثالثة وبالعكس وهنالك نوعان من الخطأ الأول يسمى بنسبة الخطأ الظاهري والثاني نسبه الخطأ الحقيقي.وفي هذه الدراسة تم التركيز على احتساب النوع الأول من الخطأ لجميع مشاهدات العينات الثلاث والجداول التالية تمثل التصنيف بحسب دالة التمييز التربيعية ودالة التمييز اللوجستية.

تصنيف دالة التمييز التربيعية

الجدول (2)يبين تصنيف الدالة التربيعية

التصنيف الصحيح (Y)	التـصـنـيـف			المجموع
	المجموعة الاولى	المجموعة الثانية	المجموعة الثالثة	
المجموعة الاولى	87	0	5	92
	94.57	0.00	5.43	%100
المجموعة الثانية	1	52	1	54
	1.85	96.30	1.85	%100
المجموعة الثالثة	47	2	75	124
	37.90	1.61	60.48	%100
المجموع	135	54	81	270
	50.00	20.00	30.00	

يبين**الجدول(2)** مدى دقة النتائج النهائية لتصنيف الدالة التربيعيةإذ يتبين إن (87) حالة من المجموعة الأولى وبنسبة 94.57 % قد تم تصنيفها بشكل صحيح وإن (5) حالات وبنسبة 5.43% تم تصنيفها بشكل خاطئ إذ صنفت بأنها ضمن المجموعة الأولى وهي تعود إلى المجموعة الثالثة.وفي المجموعة الثانية يتبين إن (52) حالة وبنسبة96.30% قد تم تصنيفها بشكل صحيح وبناءً عليه فإن هناك حالة واحدة وبنسبة (1.85) تم تصنيفها بشكل خاطئ إذ صنفت بأنها ضمن المجموعة الأولى وهي تعود إلى المجموعة الثانية و كذلك هناك حالة واحدة وبنسبة (1.85) صنفت بشكل خاطئ إذ تم تصنيفها ضمن المجموعة الثانية وهي تعود إلى المجموعة الثالثة. وفي المجموعة الثالثة يتبين إن (75) حالة وبنسبة60.81% تم تصنيفها بشكل صحيح وبناءً عليه فإن هناك (47) حالة ونسبة 37.90% تم تصنيفها بشكل خاطئ إذ صنفت ضمن المجموعة الثالثة وهي تعود إلى المجموعة الأولى وكذلك هناك حالتان وبنسبة 1.61% صنفت بشكل خاطئ إذ صنفت بأنها ضمن المجموعة الثالثة وهي تعود المجموعة الثانية، وكنتيجة عامة فقد دلت النتائج بأن الحالات المصنفة تصنيفا صحيحا كانت (214) حالة أي ما نسبته 79.26% من حالات العينة البالغة (270) .

تصنيف دالة التمييز اللوجستية

الجدول (3)يبين تصنيف الدالة اللوجستية

التصنيف الصحيح (Y)	التـصـنـيـف			المجموع
	المجموعة الاولى	المجموعة الثانية	المجموعة الثالثة	
المجموعة الاولى	83	1	8	92
	90.22	1.09	8.70	%100
المجموعة الثانية	1	52	1	54
	1.85	96.30	1.85	%100
المجموعة الثالثة	42	2	80	124
	33.87	1.61	64.52	%100
المجموع	126	55	89	270
	46.67	20.37	33.96	

يبين **الجدول (3)**إلى مدى دقة النتائج النهائية لتصنيف الدالة اللوجستية إذ يتبين إن (83) حالة من المجموعة الأولى وبنسبة 90.22 % قد تم تصنيفها بشكل صحيح وبناءً عليه فإن هناك حالة واحدة وبنسبة 1.09% تم تصنيفها بشكل خاطئ إذ صنفت ضمن المجموعة الأولى وهي تعود إلى المجموعة الثانية وكذلك يتبين إن هناك (8) حالات وبنسبة 8.70% تم تصنيفها بشكل خاطئ إذ صنفت بأنها ضمن المجموعة الأولى وهي تعود إلى المجموعة الثالثة.وفي المجموعة الثانية يتبين إن (52) حالة وبنسبة96.30% قد تم تصنيفها بشكل صحيح وبناءً عليه فإن هناك حالة واحدة وبنسبة (1.85) تم تصنيفها بشكل خاطئ إذ صنفت بأنها ضمن المجموعة الأولى وهي تعود إلى المجموعة الثانية و كذلك هناك حالة واحدة وبنسبة (1.85) صنفت بشكل خاطئ إذ تم تصنيفها ضمن المجموعة الثانية وهي تعود إلى المجموعة الثالثة. وفي المجموعة الثالثة يتبين إن (80) حالة وبنسبة64.52% تم تصنيفها بشكل صحيح وبناءً عليه فإن هناك (42) حالة ونسبة 33.87% تم تصنيفها بشكل خاطئ إذ صنفت ضمن المجموعة الثالثة وهي تعود إلى المجموعة الأولى وكذلك هناك حالتان وبنسبة 1.61% صنفت بشكل خاطئ إذ صنفت بأنها ضمن المجموعة الثالثة وهي تعود المجموعة الثانية، وكنتيجة عامة فقد دلت النتائج بأن الحالات المصنفة تصنيفا صحيحا كانت (215) حالة أي ما نسبته 79.62% من حالات العينة البالغة (270).

الاستنتاجات *Conclusions*

1. عند اختبار نسبة التصنيف الصحيح تبين ان الدالة اللوجستية افضل من الدالة التربيعية اذ بلغت نسبتها (79.62) في حين ان الدالة التربيعية فقد بلغت نسبتها (79.26).

2. عند المقارنة بين افضل نموذج للتصنيف تبين ان النموذج اللامعلمي افضل النماذج في تمثيل بيانات الدراسة كون البيانات كانت تعاني من مشاكل في التوزيع الطبيعي وتجانس التباينات.

التوصيات *Recommendations*

1. نوصي بتطبيق التصنيف على وفق دالة التمييز التربيعية ودالة التمييز اللوجستية في مجالات أخرى غير الطبية.
2. نوصي بأستخدام الدوال التمييزية الأخرى(دالة كيرنل، دالة الرتبة، الدالة اللوجستية التربيعية، نايف بيز للتصنيف).
3. توسيع عدد المتغيرات في الدراسة لتشمل (التدخين، الوراثة، تناول المشروبات الكحولية،التعرض المكثف للأشعة الضارة،الإصابة بعدوى فيروسية أو بكتيرية،....) من خلال استمارة شاملة كل المعلومات.
4. من الضروري تطوير مراكز الإحصاء في المستشفيات فضلاً عن معلومات جديدة للمرضى المصابين.
5. ضرورة تشخيص المرض وإجراء العمليات الجراحية اللازمة في وقت مبكر.

المصادر

1. الجبوري، شلال و حمزة،صلاح، "تحليل متعدد المتغيرات"،دار الكتب لجامعة بغداد، (2000).
2. Ager, J.W, JR and Brent, S.B, "An index of agreement between ahypothesized partial order and an empirical Rank order", American statistical Association, volume 73, number 364, Theory and method section, (1978).
3. Anderson, J.A., "Separate Sample Logistic Discrimination", Biometrika, 59, 19-35, (1972).
4. Berkson J., "Application of the Logistic function to Bioassay", JASA, vol. pp.357-365, (1944).
5. Burns, Robert and Burns, Richard, "Business Research Methods and Statistics using SPSS", Five extra advanced chapters, chapter 24 Logistic Regression: pp 568 -575, (2008).
6. Constanza, M.C. and Afifi, A.A., "Comparison of Stopping Rules in Forward Stepwise Discriminant Analysis", JASA. Vol. 74, No.368, pp777 – 785, (1979).
7. Hardle, W. and Simar, L., "Applied Multivariate Statistical Analysis", Berlin and Louvain-la-Neuve, Germany,p332, (2003).
8. He, X. and Fung, W.K., "High breakdown estimation for multiple with applications to discriminant analysis", Journal of Multivariate analysis, 72, 151-162, (2000).
9. Kiang, M.Y., "A Comparative Assessment of Classification Methods" .Decision Support Systems, 35, 441-454, (2003).
10. Kshlr Sagar M.A., "Multivariate Analysis", Marcel Dekker, Inc, New York, (1972).

11. McLachlan, G. J., "Discriminant Analysis and Statistical Pattern Recognition", New York: Wiley, (1992).

12. Nussbaumer, H. J., "Fast Fourier Transform and Convolution Algorithms" Springer – Verlag Berlin Heidelberg, (1982).

13. Rausch, J. R., "A comparison of Linear, quadratic, and mixture models for Discriminant Analysis under non Normal Continuous Predictors", http://www.stat.notre dame.edu/www/research/reports, (2005).

14. Schlotzhauer, D.C, "Some issues in using proc logistic for Binary logistic regression", the technical for SAS software users vol. 2. No. 4. (1993).

15. Therrien, C.W., "Decision Estimation and classification", John wiley & Sons New York, (1989).

16. Young, T.Y., "Classification, Estimation and Pattern Recognition" American Elsevier Publishing Company Inc, New York, (1974).