أستعمال بعض الطرائق التمييزية الحصينة لتشخيص أمراض أرتفاع نسبة الكولسترول في الدم

أ.م. د معاني احمد الحكيم **م.د.دريد حسين بدر**

قسم الاحصاء- كلية الادارة والاقتصاد- جامعة البصرة.

Using some robust discriminates estimation methods to diagnosis diseases height rate cholesterol in blood

Assist.prof.Dr.Maani Ahmed Alhakim

Assistant Lecture Duraid Hussein Badr

أستعمال بعض الطرائق التمييزية الحصينة لتشخيص أمراض أرتفاع نسبة الكولسترول في الدم

أ.م.د. معاني احمد الحكيم

م.د. دريد حسين بدر

المستخلص

في هذا البحث تمت المحاولة إلى إيجاد مقدرات حصينة وكفوءة لموجه المتوسط (معلمة الموقع) ومصفوفة التباين – التباين المشترك (معلمة القياس) لدوال التمييز الخطية والتربيعية والى إيجاد دوال تمييز خطية وتربيعية حصينة في حالة وجود قيم شاذة (Outliers) باستخدام عدد من طرائق التقدير الحصينة ، ومن هذه الطرائق طريقة مقدّر S (S – Estimators) الحصين ومقدّر (M) الحصين باستعمال دالة الوزن (Huber) ومقدّر (M) الحصين باستعمال دالة الوزن (Cambpe1l) ، ومقارنة دوال التمييز الحصينة مع دوال التمييز التقليدية للوصول إلى أفضل دالة تمييزية وذلك لتشخيص نوعين من أمراض أرتفاع نسبة الكولسترول في الدم (مرض الجلطة القلبية ومرض الجلطة الدماغية) .

ABSTRACT

In this research work an attempt has been made to find efficient estimations mean vector (location parameter) and variance - covariance matrix (scull parameter) for linear and quadratic discriminant function and to find the robust linear and quadratic discriminant functions when there is outliers by using a number of high quality robust methods estimation. of these methods are (S – Estimators) , M estimator by using weight (Huber) and M estimator by using weight (Cambpe1l) , moreover , comparing robust discriminant functions with classical discriminant functions to draw out the best discriminant function in order to diagnose two types of diagnosis diseases height rate cholesterol in blood which are : (Clot hearty and Clot stroke) .

المقـــدمـــة :

إن التحليل التمييزي هو أُسلوب إحصائي يتم بموجبه استخدام مجموعة من المتغيرات للتمييز بين مجموعتين أو أكثر عن طريق دالة تمييزية محددة .

فمقدرات المربعات الصغرى تتأثر بشدة بوجود عدد قليل من القيم الشاذة والنموذج المقدر سوف يعكس خصائص غير طبيعية لذا ظهرت دراسات عديدة حول معالجة مشكلة اختراق البيانات لشروط التوزيع الطبيعي من خلال أيجاد عدد من الطرائق الحصينة إذ تتصف بأن تقديراتها تكون قريبة من تقديرات المربعات الصغرى عند عدم وجود قيم شاذة وتكون أفضل عند وجود القيم الشاذة . .(Hubert, Driessen, 2004)[6]

مشكلة البحث (Objective of research)

إن أمراض أرتفاع نسبة الكورسترول في الدم من الأمراض التي تتسبب في معاناة المرضى وعدم أمكانية القيام بالأعمال الأعتيادية التي تكفل لهم العيش مما يتسبب بخسائر مادية كبيرة للمرضى وبالتالي للبلد خاصة ،وانطلاقاً من اهمية دراسة الجوانب الطبية وبخاصة امراض أرتفاع نسبة الكولسترول في الدم في العراق، فقد تم استخدام النماذج الاحصائية باستخدام عدد من طرائق التقديرالتقليدية الحصينة عند وجود القيم الشاذة .

هدف البحث :

ان هدف البحث يتمثل بالنقاط الاتية :

١. استخدام مقدرات حصينة مقاومة لتأثير الشواذ ومن تلك المقدرات مقدّر طريقة مقدّر S (S – Estimators) الحصين ومقدّر (M) الحصين باستعمال دالة الوزن (Huber) ومقدّر (M) الحصين باستعمال دالة الوزن (Cambpe1l) .

٢. حساب احتمال خطأ التصنيف لكل مقدر .

٣. المقارنة بين المقدّرات للتعرف على أفضل مقدّر والذي يعطي أقل احتمال لخطأ التصنيف من خلال التطبيق على بيانات حقيقية لمرضى ارتفاع نسبة الكولسترول في الدم .

٢. الجانب النظري

طرائق التقدير الحصينة

ان المعنى الإحصائي لمفهوم الحصانة (Robustness) كما عرفه (Huber,1981) بأنه عدم الحساسية للقيم الشاذة، وعدم الحساسية للانحرافات عن الافتراضات الواجب توافرها ، وعدم الحساسية للتوزيع المفترض، لذا فالطرائق الحصينة ينبغي أن تكون ذات كفاية قريبة من طرائق التقدير التقليدية في حالة تحقق الافتراضات، وأفضل منها في حالة الانحراف عن تلك المشاهدات. ومن طرائق التقدير الحصينة لمعلمتي الموقع، والقياس التي سنتناولها في دراستنا هذه هي طريقة مقدّر S (S – Estimators) الحصين ومقدّر (M) الحصين باستعمال دالة الوزن (Huber) ومقدّر (M) الحصين باستعمال دالة الوزن (Cambpe1l) التي تستخدم لتقدير موجه المتوسط $\underline{\mu}$ ومصفوفة التباين والتباين المشترك $\sum$. (Hubert,Driessen,2004)[6]

طريقة مقدّر S (S – Estimators) :

وتقوم هذه الطريقة بفكرة حساب تقديرات ابتدائية لموجه المتوسط ($\underline{\mu}$) ومصفوفة التباين والتباين المشترك ($\sum$) للمجموعات الجزئية بحجم (h) وتوصيف الطريقة يكون كما يلي :

[٢] (العلوي ، ٢٠٠٣) ، [10](Rousseeuw , Leroy, 1987)

إن مقدرات (S) معرفة كحلول جيدة للموقع والقياس (t_n , C_n) لمسألة تقليل محددة مصفوفة التباين المشترك (C) على وفق الأتي :

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left[d(\underline{X}i,\underline{t},C)\right]=bo \quad \ldots(2-1)$$

خلال كل (P) SPD , $\underline{t}$, $\underline{c}\epsilon R^p$

وعندما $b_o = P$ و $\rho(d) = d^2$ فأنة يتم الحصول على المقدرين التقليدين X , S كحلول وحيدة لمشكلة التصغير .

إذ أن d(xi , t , C) : تمثل مسافات مهلنوبس التربيعية العادية للمشاهدات في العينة وأن i=1 ,2, ……., n

SPD(P) : تمثل مجموعة كل المصفوفات المتماثلة الموجبة (Symmetric Positive definite) ذات البعد PxP .

وللدالة ρ شروط يجب توافرها وهي أن تكون متماثلة وتزداد في المدة R^+ أي أنها غير متناقصة في المدة $[0,\infty]$

وأخيراً أن $\rho(0) = 0$, $\rho(\infty) = 1$

وأن الأختيار الشائع إلى ρ هو دالة وزن توكي (Tokey Biweight Function) وهي معرفة بالصيغة الآتية :

$$(\rho(u)=\min\left[\frac{u^2}{2}-\frac{u^4}{2C^2}+\frac{u^6}{6C^4},\frac{C^2}{6}\right] \ldots 2-2)$$

إذ أن(C=1. 547 ثابت) يحقق قيمة مركزية لنقطة الانهيار وهي 50% [8](Pison,Aelst, 2002) .

وأن :

$$\rho(u)=\begin{cases}\dfrac{u^2}{2}-\dfrac{u^4}{2C^2}+\dfrac{u^6}{6C^4} & if\ |u|\le C\\ \dfrac{C^2}{6} & if\ |u|>C\end{cases}$$

ولتحقيق نقطة الانهيار الى 50% فأن ρ(u) تكون :

$$(\rho(u)=\frac{b}{\delta} \ldots 2-3)$$

إذ أن $\delta = 50\%$ (نقطة الانهيار).

خوارزمية حساب مقدر (S – Estimators) [٢] (العلوي ، ٢٠٠٣)

يمكن تلخيص خوارزمية مقدَر (S) بالخطوات الآتية :-

١- إن مقدرات(S) معرفة كحلول للموقع والتشتت (t_n , C_n) لمسألة تقليل محددة التباين المشترك C على وفق الأتي :

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left[d(\underline{X}i,\underline{t},C)\right]=bo$$

٢- للعينة المكونة من (n) من المشاهدات (الصفوف) و (P) من المتغيرات (الأعمدة) يتم أستخراج العينات الجزئية ذات الحجم (P+1) أي نختار C_{P+1}^{n} من العينات الجزئية .

٣- بعد أستخراج العينات الجزئية يجري أستخراج متجة الموقع (Location) ومصفوفة التشتت لكل عينة جزئية .

٤- يكون أختيار أفضل عينة جزئية من العينات المستخرجة على وفق محددة مصفوفة التباين المشترك إذ يتم أختيار العينة الجزئية التي تكون فيها محددة مصفوفة التباين المشترك لها أقل ما يمكن أي أن :

$$\tilde{J} = \arg_j \min \left| \sum_j \right| \qquad 2-4$$

بعد أختيار أفضل عينة جزئية من خلال موجة الموقع ومصفوفة التباين المشترك يتم استخراج مسافات مهلنوبس التربيعية العادية بين نقاط المشاهدات Xi والموقع t بالاعتماد على مصفوفة التباين المشترك إذ :

$$d(\underline{x}i, \underline{t}, C) = \left[(\underline{x}i - \underline{t})' C^{-1} (\underline{x}i - \underline{t}) \right]^{\frac{1}{2}} \quad \dots (2-5)$$

٥- يتم أستخراج ρ(u) حيث أن ρ هو دالة وزن (Biweight) بعد التعويض بقيمة (u=٠.٧) في المعادلة

$\rho(u) = \frac{b}{\delta}$ (2-2) نحصل على $\rho(u) = \frac{1}{6}$ وبتطبيق المعادلة $\rho(u) = \frac{b}{\delta}$ يتم تحديد قيمة b وبالشكل التالي :

$$\rho(u) = \frac{b}{\delta}$$

$$\frac{1}{6} = \frac{b}{0.50}$$

$$\therefore b = 0.083$$

٧- بعد أيجاد قيمة b أعلاة ندخل في عملية تكرار المعادلة ρ(u) ليتم فيها البحث عن قيمة u وذلك لتحقيق شرط المقدر في المعادلة (1 – 2).

٨- بأستخراج قيم الثوابت يحقق شرط المعادلة (1 – 2) ليتم بعدها أختيار متجه المتوسطات ومصفوفة التشتت التي تحقق الشرط مع متجه الموقع ليعدا مقدرات (S)الحصينة للموقع والتشتت والتي يتم بموجبها أستخراج مسافات مهلنوبس التربيعية الحصينة لتشخيص بعد ذلك نقاط البيانات الشاذة عن غيرها وتعد المشاهدة شاذة إذا كانت على وفق : $RDi > \sqrt{\chi^2_{(p,0.975)}}$

طريقة مقدّر(M) الحصين باستعمال دالة الوزن (Huber) : (M Estimator)

وهي إحدى طرائق التقدير لحصينة لمعلمتي الموقع والقياس و يعتمد هذا المقدّر على موازنة المشاهدات لتقليل تأثير الشواذ و إن عائلة M تعتمد على تقديرات الأوساط الحسابية و التباينات الموزونة و من هذه الطرائق طريقة

مقدر M باستعمال دالة الوزن (Huber) و يمكن توضيح هذه الطريقة بالخوارزمية الأتية كما يلي : [٣] (النداوي ، ٢٠٠٨) ، [4] (الياسين،٢٠٠٩) ، (Rousseeuw,1985)[9]

خوارزمية حساب مقدّر (M) الحصين باستعمال دالة الوزن (Huber)

1. تحسب مسافة مهلنوبس لكل مشاهدة و كالآتي:

$$D(\underline{x}_i, x) = \left[(\underline{x}_i - \underline{T}(x))' S^{-1}(x)(\underline{x}_i - \underline{T}(x))\right]^{\frac{1}{2}} \quad .....(2.6)$$

و إن :

$$T(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \quad .....(2.7)$$

$$S(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - T(x))(x_i - T(x))' \quad .....(2.8)$$

2. لتشذيب البيانات يعاد حساب المتوسطات و مصفوفة التباينات حيث تبدل $T(x)$ و $S(x)$ بأوزان تعتمد على d_i أي:

$$T^*(x) = \frac{\sum w_i x_i}{\sum w_i} \quad .....(2.9)$$

$$S^*(x) = \frac{\sum w_i (x_i - T^*(x))(x_i - T^*(x))'}{\sum w_i^2} \quad .....(2.10)$$

و إن :

$$w_i = \begin{cases} 2d_i & \text{if} \quad d_i > 2 \\ 1 & \text{if} \quad d_i \le 2 \end{cases}$$

ثم يعاد حساب الخطوة (2) بصورة تكرارية بالاعتماد على نتائج التكرار السابق و نتوقف عندما يكون الفرق بين تكرارين متعاقبين قليلاً على وفق مستوى معين من الدقة، إذ يتم الحصول على بيانات مشذبة للمجاميع كافة ، إذ يتم ضرب القيم بالتكرار الأخير بالمتغيرات كافة و للمجاميع كافة وعندها يتم تطبيق الدالة التمييزية الخطية و التربيعية .(Croux,Haesbroeck, 2001)[5] .

طريقة مقدّر(M) الحصين باستعمال دالة الوزن (Cambpe1l):(M Estimator)

وتقوم هذه الطريقة بفكرة حساب تقديرات ابتدائية لموجه المتوسط ($\underline{\mu}$) ومصفوفة التباين والتباين المشترك ($\sum$) للمجموعات الجزئية بحجم (h) وتوصيف الطريقة يكون كما يلي : (Rousseeuw , Leroy, 1987)[10]

لنفرض أن $X=\{x_{i1},\ldots\ldots,x_{iN}\}$ عينة عشوائية مكونة من (N) من المشاهدات في p من المتغيرات. يتم أولاً تقدير ($\overline{X}^0 \sum^0$) بواسطة طريقة التقدير كقيم أولية وهي محددة بواسطة اختيار المجموعة الجزئية ($X_{il},\ldots\ldots,$ X_{ih}) ذات (h) من المشاهدات (الصفات) والتي تقلل التباين العام (المقياس محدد بواسطة محدد مصفوفة التباين المشترك والتي تم الحصول عليها من المجموعة الجزئية) ، بعد ذلك يقدر متجه الأوساط الحسابية للمجموعة الجزئية ذات الحجم h وكالآتي : [٣] (النداوي ، ٢٠٠٨) :

$$\overline{X}_i^0 = \frac{1}{h}\sum_{j=1}^{h} X_{ij} \qquad \ldots(2\text{-}11)$$

وتقدير التباين– التباين المشترك كالآتي:

$$\hat{\Sigma}_{j,0} = dp \times \frac{1}{h}\sum_{i=1}^{h}(\underline{x}_{ij} - \overline{\underline{x}}_{j,0})(\underline{x}_{ij} - \overline{\underline{x}}_{j,0})' \qquad \ldots.(٢\text{-}12)$$

قيمة h مساوية إلى $h=[n(1-c)]$ حيث إن (c) تمثل نسبة القطع في تقديرات (M) باستعمال دالة الوزن (Cambpel1) وان هذه التقديرات تعطي كفاءة واطئة ضمن التوزيع الطبيعي متعدد المتغيرات.

للتغلب على هذه المسألة فأن خطوة إعادة الوزن يمكن أن تضاف للطريقة تقديرات (M) باستعمال دالة الوزن (Cambpe1l) إذ تحسب الأوزان كالآتي :-

$$W(d_i) = \begin{cases} d_i, & if \quad d_i \le d_0 \\ \dfrac{d_0 \exp\{-0.5(d_i - d_0)^2\}}{b_2{}^2}, & if \quad d_i > d_0 \end{cases} \qquad \ldots(2\text{-}13)$$

أذ إن :

$d_0 = \sqrt{p} + b_1\sqrt{2}$

$b_1 = (2, \infty)$

b2= 2, 1.25

حيث إن (c) هي قيمة القطع (Cut- Off Value) وقيمته تساوي القيمة الجدولية لمربع كاي بدرجة حرية p وبمستوى معنوية $1-\alpha$ إي إن $c = \chi^2{}_{p,1-\alpha}$

وبذلك يمكن أن تعرف مقدرات (M) باستعمال دالة الوزن (Cambpe1l) بالشكل الأتي :

$$\overline{X}_j = \frac{\sum_{i=1}^{n} W_i X_{ij}}{\sum_{i=1}^{n} W_i} \qquad \ldots\ (2\text{-}\ 14)$$

$$\sum = \frac{\sum_{i=1}^{n} W_i (X_{ij} - \overline{X}_j)(X_{ij} - \overline{X}_j)}{\sum_{i=1}^{n} W_i - 1} \qquad \ldots\ (2\text{-}\ 15)$$

أن التقديرات الموزونة الجديدة تقود إلى كفاءة إحصائية أفضل. وهذا ما يجعلنا نقول أن خطوة إعادة الوزن تزيد من جاذبية الطريقة لحساب مقدر (M) باستعمال دالة الوزن (Cambpe1l) .

. [11] (Rousseeuw,Katrin, 1999)، (الياسين ، ٢٠٠٩)[4]

خوارزمية حساب مقدّر (M) الحصين باستعمال دالة الوزن (Cambpe1l)

١. للعينة المكونة من (N) من المشاهدات و (p) عن الأبعاد يتم استخراج العينة الجزئية ذات الحجم p+1 .

٢. إيجاد متجه الأوساط الحسابية $\overline{X}_{P+1}$ للعينة الجزئية المستخرجة ومصفوفة التباين–التباين المشترك $\sum_{P+1}$.

٣. يتم اختيار العينة ذات الحجم (p+1) على أساس أنها تمتلك أصغر محدد لمصفوفة التباين–التباين المشترك $\sum_{P+1}$ من بقية العينات الجزئية الممكن اختيارها ذات الحجم (p+1)

٤. بعد أن يتم تحديد أفضل مجموعة جزئية التي تمتلك أصغر محدد لمصفوفة التباين – التباين المشترك يتم استخراج المسافات التربيعية للمشاهدات كافة على وفق اختيار $\sum_{P+1}$ و $\overline{X}_{P+1}$.

٥. نحسب قيمة $hp = \text{int}(\frac{N+p+1}{2})$ التي تمثل عدد المشاهدات التي سوف تعتمد لحساب مقدرات (M) باستعمال دالة الوزن (Cambpel1) .

٦. يتم اختبار (hp) من المشاهدات التي لها أصغر المسافات المستخرجة .

٧. للمصفوفة الجديدة التي تحتوي على (hp) من المشاهدات المحددة ويتم حساب متجه الأوساط الحسابية لها $\overline{X}_{hp}$ ومصفوفة التباين–التباين المشترك $\sum_{hp}$,

٨. تحسب المسافات التربيعية ثانية للعينة الكلية على وفق $\overline{X}_{hp}$ ومصفوفة التباين والتباين المشترك المتسقة $\sum_{hp}$.

٩. يتم اختيار (hp) من المشاهدات التي تمتلك أصغر قيم للمسافات المستخرجة إذ يتم لهذه المجموعة الجديدة المختارة حساب مقدر متجه الأوساط الحسابية معاد الأوزان Campbell – Estimate ($\overline{X}_{\text{Cambpell}}$) ومصفوفة التباين معاد الوزن Campbell – Covariance Matrix ($\sum_{\text{Cambpell}}$) .

ويتم بعد ذلك استخدام قواعد التمييز الخطية والتربيعية التقليدية والحصينة ولطرائق التقدير التقليدية والحصينة .

٢. قواعد التمييز الحصينة (Robust Discriminant Rules)

قواعد التمييز الخطية الحصينة (Robust Linear Discriminant Rules)

على وفق الطرائق يتم لكل مجموعة j إذ إن (j = 1,2,3,.....,L) تقدير موجه متوسط ($\underline{\mu}_j$)، ومصفوفة التباين والتباين المشترك(Σ_j) وخطأ التصنيف (p_j) للمجموعة الجزئية (h) التي تمتلك أصغرمحدد من بين (H) من المجموعات الجزئية الكلية لكل مجموعة اذ تكون المقدّرات على النحو الآتي :

$$\overline{\underline{x}}_{j,\text{Methed}} = \frac{\sum_{i=1}^{h} \underline{x}_{ij}}{h} \quad , j=1,2,...,L \quad \text{....(2-16)}$$

$$\hat{\Sigma}_{\mathrm{j,Methed}} = \mathrm{dp} \times \frac{\sum_{i=1}^{h} (\underline{x}_{ij} - \underline{\bar{x}}_{\mathrm{j,Methed}})(\underline{x}_{ij} - \underline{\bar{x}}_{\mathrm{j,Methed}})'}{h} \qquad \text{....(2-17)}$$

(الياسين ، ٢٠٠٩)[4] يمكن أن يقدّر. $(\hat{P}_j^R)$ وان مقدار احتمال خطأ التصنيف الحصين

$$\hat{P}_j^R = \frac{\tilde{n}_j}{\tilde{n}} \qquad \text{....(2-18)}$$

إذ إن:

$\tilde{n}_j$: عدد المشاهدات غير الشاذة المقدرة في المجموعة j .

$\tilde{n}$: عدد المشاهدات الكلية غير الشاذة المقدرة لكل المجتمعات.

وبعدها يتم حساب مصفوفة التباين والتباين المشترك المدمجة العامة (Pooled Var-Cov Matrix) على وفق الصيغة الآتية :

$$\hat{\Sigma}_{\mathrm{PMethod}} = \frac{\sum_{j=1}^{L} n_j \hat{\Sigma}_{\mathrm{j,Methed}}}{\sum_{j=1}^{L} n_j} \qquad \text{....(2-19)}$$

إذ إن قاعدة التمييز الخطية الحصينة(RLDR) على وفق المقدّرات في الصيغ (2-16) ، (2-17)، (2-19) تصبح :

بتعيين x إلى $\prod_k$ وإذا كانت :

$$\hat{d}_k^{RL}(\underline{x}) > \hat{d}_j^{RL}(\underline{x}) \quad , \forall j=1,\ldots,L \ , \ j \neq k \qquad \text{....(2-20)}$$

إذ إن:

$$\hat{\mathrm{d}}_{\mathrm{j}}^{\mathrm{RL}}(\underline{\mathrm{x}}) = \hat{\mathrm{d}}_{\mathrm{j}}^{\mathrm{RL}}(\underline{\mathrm{x}}, \underline{\hat{\mu}}_{\mathrm{j,Methed}}, \hat{\Sigma}_{\mathrm{PMethed}}) =$$

$$\underline{\bar{\mathrm{x}}}'_{\mathrm{j,Methed}} \hat{\Sigma}^{-1}_{\mathrm{PMethed}} \underline{\mathrm{x}} - \frac{1}{2} \underline{\bar{\mathrm{x}}}'_{\mathrm{j,Methed}} \hat{\Sigma}^{-1}_{\mathrm{PMethed}} \underline{\bar{\mathrm{x}}}_{\mathrm{j,Methed}} + \mathrm{Ln}(\hat{\mathrm{p}}_{\mathrm{j}}^{\mathrm{R}})$$

....(2-21)

و عندما $\pi_1 = \pi_2$ وان (L=2) فإن الصيغة (2-21) تعرف بقاعدة التمييز الحصينة لفشر التي توصف على النحو الآتي :

$$\begin{cases} x \in \pi_1 & \text{if} \quad (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' \hat{\Sigma}^{-1}_{\mathrm{PMCD}} (\underline{x} - (\underline{\bar{x}}_1 + \underline{\bar{x}}_2)/2) > 0 \\ x \in \pi_2 & \text{if} \quad \text{other wise} \end{cases}$$

....(2-22)

قواعد التمييز التربيعية الحصينة (Robust Quadratic Discriminant Rules)

وتقوم هذه الطريقة على إيجاد المقدّرات ($\underline{\bar{x}}_{j,Methed}$) و($\hat{\Sigma}_{j,Methed}$) و ($\hat{P}_j^R$) لكل مجموعة j إذ ان $j=1,2,3,...,L$ كما في الصيغ، (٢-١٦)،(2-17)، (2-19) . وبالتالي فان قاعدة التمييز التربيعية الحصينة (RQDR) وعلى وفق هذه الطريقة تصبح (Hubert,Driessen,2004)[٦] :

بتعيين x إلى $\prod_k$ وإذا كانت :

$$\hat{d}_k^{RQ}(\underline{x}) > \hat{d}_j^{RQ}(\underline{x}) \quad , \forall j=1,2,...,L \quad , j\neq k$$

إذ إن :

$$\hat{d}_j^{RQ}(\underline{x}) = \hat{d}_j^{RQ}(\underline{x}, \underline{\hat{\mu}}_{j,Methed}, \hat{\Sigma}_{j,Methed})$$

$$= -\frac{1}{2}Ln\left|\hat{\Sigma}_{j,Methed}\right| - \frac{1}{2}(\underline{x}-\underline{\bar{x}}_{j,Methed})'\hat{\Sigma}_{j,Methed}^{-1}(\underline{x}-\underline{\bar{x}}_{j,Methed}) + Ln(\hat{p}_j^R)$$

.(2-23)

[5](Croux, Haesbroeck, 2001)

٣ اختبار دالة التمييز :

اختبار المتغيرات :

لاختيار المتغيرات ذات القوة التمييزية المعنوية و التي تتبع أقل خطأ تصنيف، هناك عدة أساليب منها طرائق الاختيار المتدرج (Stepwise Selection) و فيها يكون اختيار دالة التمييز القياسي لمعنوية الفروق بين متوسطات المجاميع و بأخذ التباينات المشتركة و لـ P من المتغيرات : (Neil.H.Timm, 2002)[7].

$$\hat{P} = \left|W_{PP}\right| / \left|T_{PP}\right|$$

....(2-24)

إذأن :

W: مصفوفة مجموع مربعات الانحرافات داخل المجاميع.

T : مصفوفة مجموع مربعات الانحرافات الكلية.

و تكون إحصاءة F المناظرة المستندة إلى أساس هذا المقياس كالآتي :

$$F(\hat{P}) = \left[1-\hat{P}\right]\left[n-P-1\right] / \hat{P}P$$

....(2-25)

و التي تتوزع $F_{1,n-P-1}$.

اختبار دالة التمييز بعد حذف المتغيرات غير التمييزية :

ومن هذه الاختبارات :

اختبار معنوية الدالة المميزة (تساوي المجموعات) :

تم استخدام مقياس مربع كأي وتكون صيغته كالأتي :(Neil.H.Timm , 2002)[7].

$$\chi^2 = -\left[n-1-1/2\,(P+g)\right]\log \hat{e} \qquad \text{....(2-26)}$$

إذ أن:

P : المتغيرات المقدرة.

g :عدد المجاميع .

^: مقياس ولكس ويكون بين (الصفر) و(الواحد).

و الذي يتوزع تقريباً مربع كآي بدرجة حرية $P(g-1)$.

اختبار تساوي مصفوفة التباين و التباين المشترك لجميع المجاميع :

وتكون احصاءة هذا الاختبار كما يلي:(Neil.H.Timm, 2002)[7] ، (Pison,Aelst, 2002)[8] .

$$\mu = (\sum_{i=1}^{k} V_i)\,Ln\,|S| - \sum_{i=1}^{k}(V_i\,Ln\,|S|) \qquad \text{....(2-27)}$$

وقد أثبت Box عام (1949) أنه إذا ضرب μ في ثابت C^{-1} والذي يساوي :

$$C^{-1} = 1 - \frac{2m^2+3m-1}{6(m+1)(k-1)}\left[\sum_{i=1}^{k}\frac{1}{V_i} - \frac{1}{\sum_{i=1}^{k} V_i}\right]$$

....(2-28)

سوف نحصل على مقياس يتوزع بالتقريب توزيع χ^2 وبدرجة حرية $\frac{1}{2}(k-1)\,m\,(m+1)$

إذ أن :

m : عدد المتغيرات المدروسة في المصفوفة.

k : عدد المصفوفات المختارة .

٣. الجانب التطبيقي

أمراض أرتفاع نسبة الكولسترول في الدم أعراضها السريرية والعوامل المسببة للمرض وطرائق تشخيصها :

إن التصلب التدريجي لجدران الشرايين ينتهي بحدوث جلطة تسد الشريان الضيق تماماً وينقطع الدم عن جزء من عضلات القلب إذ تقل تغذية القلب نتيجة لانسداد الشرايين فيصبح أي عمل إضافي مجهوداً للقلب كعدم القدرة على المشي مسافة كان سابقاً يستطيع مشيها، وعدم القدرة على صعود مرتفع و عند تفاقم الحالة يكون صعود الدرج صعباً، كذلك يعاني من ضيق التنفس و عدم القدرة على النوم إذ تكون الآلام في الصدر ثم تنتشر إلى الكتف الأيسر عادة وتكون مصحوبة بتعرق و شحوب الوجه مع خفقان القلب مما يؤدي إلى عدم تشخيصه كجلطة قلبية و إذا كان الانسداد كبيراً فإن بطين القلب يبدأ بالارتجاف ثم يتوقف العمل ويموت المصاب ولما كان ارتفاع نسبة الكولسترول في الدم من أهم الأعراض إذ تلتصق الدهون بجدران الأوعية لتشكل ترسبات تسبب ضيق في الأوعية الدموية و تشكل في بعض الأحيان انتفاخاً يشبه البصلة داخل تجويف الأوعية و عند الجهد المفاجئ قد

تنفجر هذه البصلة و تتوزع محتوياتها على شكل كتل صغيرة داخل الوعاء الدموي و قد تسبب في حالات اغلاق شريان في القلب إذ تسبب جلطة في القلب أو اغلاق شريان في المخ و تسبب الجلطة الدماغية و حالات أخرى تسبب مضاعفات خطيرة كالشلل و السكتة القلبية. لذا يجب الاعتماد على الزيوت ذات المصدر النباتي و كذلك ممارسة الرياضة إذ تقوم بتخليص الدم من ترسبات الدهون العالقة و لتشخيص مثل هذه الأعراض يستند أولاً إلى الأعراض السريرية و هي تخطيط القلب، تحليل الدم وعند حدوث الجلطة عمل الايكو، عمل رنين للدماغ، وهنا تبرز أهمية التشخيص المبكر لأمراض ارتفاع نسبة الكولسترول في الدم . تشارلز، و.أيرتج وآخرون (1970)[1]

وصف العينة و المتغيرات المستخدمة وأسلوب جمع البيانات:

تم اعتماد ملفات المرضى (طبلات المصابين) في المستشفى التعليمي ومستشفى الجمهوري في محافظة البصرة للأعوام (٢٠١٥، ٢٠١٦ ،٢٠١٧)، واشتملت بيانات البحث على (١٤٠) مريضاً، وقد تم سحب عينة عشوائية من المجتمع الأول (الجلطة القلبية Clot hearty) بحجم (n_1=70) ومن المجتمع الثاني (الجلطة الدماغية Clot stroke) بحجم (n_1=70)، والتي تمثل مجموعتين من التشخيص لأمراض ارتفاع نسبة الكولسترول في الدم الذي يمثل المتغير المعتمد (متغير الاستجابة) الذي يتمثل بنوع المرض (مرض الجلطة القلبية =1) و(مرض الجلطة الدماغية =2) وعدد المتغيرات الإيضاحية (المستقلة) التي تم الاتفاق عليها مع الأطباء أصحاب الاختصاص هي (٧) متغيرات يمكن توضيحها بإيجاز على النحو الآتي :

1. العمر (age).
٢. الوزن (weight).
٣.الطول (height) .
٤. قياس ضغط الدم الواطئ (lbp).
5. قياس ضغط الدم العالي (hbp).
6. نسبة الكولسترول في الدم cholesterol.
7. الجنس(sex): و يمثل متغير ثنائي (ذكر = 1) ، (أنثى = 2).

و كان توزيع مفردات العينة بين المجموعتين و نسبة كل مجموعة كالآتي :

Group	No. of Cases	Prior	Group Name
1	70	0.5٠	الجلطة القلبية
2	٧٠	0.٥٠	الجلطة الدماغية
Total	1٤٠		

نتائج الاختبارات :

١. اختزال عدد المتغيرات:ولتقليص عدد المتغيرات لكي يشتمل النموذج على المتغيرات ذات الطابع التمييزي فقط تم استخدام الاختيار المتدرج (Stepwise Discriminant Analysis) وباستخدام إحصاءتي Wilk's Rao's V&Lambda كمعيار لاختيار المتغيرات ذات القيمة الأصغر لهذه الإحصاءة ، و أظهرت نتائج التحليل دخول ستة متغيرات و استبعاد متغير الجنس كما في الجدول رقم (١) ، (٢) ، (3) .

جدول (1)

نتائج اختيار المتغيرات

Variable in the Analysis (اختيار المتغيرات لتقدير الدوال المميزة)							
مستوى المعنوية	Rao's V	مستوى المعنوية	Wilk's Lambda	F to Remove	Tolerance	Variables	Step
		.000	.894	22.565	1.000	Cholesterol	1
.000	16.126	.000	.894	28.429	.961	Cholesterol	2
.000	22.555	.000	.858	21.802	.961	hbp	
.000	37.559	.000	.784	24.513	.960	Cholesterol	3
.000	40.346	.000	.771	22.008	.955	hbp	
.000	48.161	.000	.738	15.387	.994	age	
.000	44.234	.000	.755	28.956	.923	Cholesterol	4
.000	64.908	.000	.677	12.290	.899	hbp	
.000	59.471	.000	.696	16.332	.987	age	
.000	69.309	.000	.662	9.176	.878	lbp	
.000	50.318	.000	.730	30.930	.910	Cholesterol	5
.000	72.513	.000	.652	13.588	.889	hbp	
.000	67.21	.000	.669	17.383	.979	age	
.000	21.837	.000	.620	7.356	.872	lbp	
.000	83.474	.000	.591	6.317	.962	height	
.000	85.129	.000	.615	10.432	.676	Cholesterol	6
.000	76.006	.000	.641	16.578	.846	hbp	
.000	81.581	.000	.625	12.739	.951	age	
.000	88.473	.000	.606	8.325	.862	lbp	
.000	91.327	.000	.598	6.576	.960	height	
.000	93.976	.000	.591	4.991	.620	weight	

جدول (٢)

نتائج إحصاءة Wilk's Lambda

Exact F				df3	df2	df1	Lambda	Number of Variables	Step
Sig.	df2	df1	Statistic						
.000	136.000	1	22.555	136	1	1	.858	1	1
.000	135.000	2	23.904	136	1	2	.738	2	2
.000	134.000	3	22.763	136	1	3	.662	3	3

.000	133.000	4	20.408	136	1	4	.620	4	4
.000	132.000	5	18.242	136	1	5	.591	5	5
.000	131.000	6	16.494	136	1	6	.570	6	6

جدول (٣)

نتائج إحصاءة Rao's V

Change in v		Rao's V			Entered	Step
Sig.	Statistic	Approx. sig.	df	Statistic		
.000	22.555	.000	1	22.555	Cholesterol	1
.000	25.606	.000	2	46.161	hbp	2
.000	21.147	.000	3	69.309	age	3
.000	14.165	.000	4	83.474	lbp	4
.000	10.503	.000	5	93.976	height	5
.000	8.762	.000	6	102.739	weight	6

٢. تم اختبار البيانات لمعرفة هل أن المتغيرات تتوزع توزيعاً طبيعياً (Normal) أم لا من خلال أختبارجودة توفيق البيانات في برنامج (SPSS.16) وحسب الاختبارات أظهرت النتائج أن المتغيرات الوزن (weight)، الطول (height)، ضغط الدم الواطئ (lbp)، ضغط الدم العالي (hbp)، الكولسترول في الدم (cholesterol) لا تتوزع توزيعاً طبيعياً، فقط متغير العمر (age) كان يتوزع توزيعاً طبيعياً.

٣. تم الكشف عن وجود القيم الشاذة بطريقة (Box– and –Whisker Plot) إذ أظهرت النتائج أن المتغيرات (الوزن، الطول، ضغط الدم العالي و الواطئ، و نسبة الكولسترول في الدم) تحتوي على قيم شاذة و بأعداد متفاوتة من طرف واحد أو طرفين و لم تظهر قيم شاذة في متغير (العمر). و الرسوم معطاة في الملحق (١).

٤. اختبار معنوية الدالة المميزة و بالاعتماد على المعادلة (2.26) كانت قيمة مربع كاي $\chi^2 = 66.841$ وبمستوى معنوية (0.01) و هذا يدل على أن الدالة لها إمكانية جيدة على التمييز.

٥. اختبار تساوي التباينات و بالاعتماد على المعادلة (2.27) و (2.28) فإن Box'sM=142.488 و هذا يدل على عدم التجانس بين المجموعات .

نتائج تحليل طرائق التمييز لبيانات أمراض ارتفاع نسبة الكولسترول في الدم :

لغرض تطبيق طرائق التقدير الحصينة في الواقع العملي تم استخدام بيانات حقيقية عن أمراض ارتفاع نسبة الكولسترول في الدم وهي مرض الجلطة القلبية و مرض الجلطة الدماغية لأنها تحقق الغرض من هذه الدراسة، وذلك يعود إلى أن المجالات الطبية مليئة ببيانات تتصف بان بعض قياساتها أو التسجيلات الخاصة بمتغيراتها تكون متجاوزة لنسبها الطبيعية داخل جسم الإنسان، وهذه القياسات تعد حينئذ شاذة ، ونتيجة لعدم توافر برامج جاهزة بهذا الخصوص، فقد قام الباحث بكتابة برنامج بلغة (Q.Basic) لغرض تطبيق طرائق التقدير الحصينة ، (والبرنامج معطى في الملحق (٢)) ، وإن نتائج التقدير لجميع متغيرات الدراسة للمجموعتين الأولى والثانية

باستخدام دالة التمييز الخطية والتربيعية التقليدية والحصينة ولطرائق التقدير التقليدية والحصينة موضحة في الجدولين (٤) ، (5) على التوالي :

جدول (4)

نتائج تحليل طرائق التمييز لبيانات الجلطة القلبية

المجموعة الأولى باستعمال طرائق التمييز التقليدية والحصينة

RQ . S	RL .S	RQ. Huber	RL . Huber	RQ . Cambpe1l	RL . Cambpe1l	QDF	LDF	Variables
٢١١.٤٤	١١٥.٥٠	٥.٤٤١٩	- ١٣٨.٧٣	-٢٤٣. ٣٦	٥.٢٢٠٩	١٦١.٠٩	٢.٣٩٥٤	constant
٢.٢٢	٢.٢١	٠.٩٨	0.٩٧	٣.٠٢	٣.٣٠	٨٠.٣٤	٨١.٥٧	age
٢.٠٢	٢.٢٤	٤.٠٨	0.٧٠	٣.١٣	٥٢.45	٥.٧٤	٦.٨٢	weight
0.٥	0.٩٧	٢.1	٢.٠٧	٢.٩٨	٣.٦٢	١٩.٧٦	٢٢.٩٤	height
0.٨	0.٧	-٨.30	0.٩٩	٢.92	٦٢٥.23	١٤.٧٥	١٥.٤٢	lbp
0.٧٣	0.٨٨	٣.٧٨	٢.30	٣.٢٣	٦٨٣.٧٩	٠.٠٢	٠.٠٣	hbp
0.٨٧	٢.١٣	٢.٤٠	٢.٣٢	٢.٩٤	٤٢٠. ٣٥	٥.93	٦.٨٦	cholesterol

جدول (5)

نتائج تحليل طرائق التمييز لبيانات الجلطة الدماغية

المجموعة الثانية باستعمال طرائق التمييز التقليدية والحصينة

RQ . S	RL .S	RQ. Huber	RL . Huber	RQ . Cambpe1l	RL . Cambpe1l	QDF	LDF	Variables
- ٣٠٢.٦٦	١٠٣.٦٦	٧.55	-٤٠٤.٢٢	-٣٠٤.٥٣	٥. 34	١٨١.09	٥.21	constant
٢.٠٧	٢.05	١.٦٩	١.٠٤	٢.٥٧	٢.٣٥	٤٣.٦٣	٢٠.95	age
0.٧٨	٢.١٩	٣.٣٠	٣.٠٤	٣.١٧	٢.٩٥	0.٤٥	0.٣٠	weight
0.٧٧	١.٠	٢.٨٩	٢.٩٧	٢.٩٨	٢.٩٤	٥.٦٢	٤.٨٢	height
-٢.٠٧	0.٧٩	٥.١٥	٣.٢٧	٣.٥٧	٢.٧٨	٤.٠٢	٣.٥٧	lbp
٢.٤٠	٢.٢٩	٣.٩١	٢.٨٥	٢.٨٩	٢.٨٢	١.٠٢	١.٠٢	hbp
١.٧٥	٢.١٨	٢.٥٥	٣.٢٧	٢.٨٦	٣.١٠	٢.٠٩	٢.٣٠	cholesterol

وان نتائج احتمال خطأ التصنيف لجميع متغيرات الدراسة لبيانات (الجلطة القلبية والجلطة الدماغية) لأمراض ارتفاع نسبة الكولسترول في الدم للمجموعتين الأولى والثانية موضحة في الجداول **(6)** ، **(7)** على التوالي :

جدول (6)

احتمال خطأ التصنيف لبيانات الجلطة القلبية المجموعة الأولى

LDF	QDF	RL.S	RQ.S	RL.Huber	RQ.Huber	RL. Cambpe1l	RQ . Cambpe1l
0.٥٤	0.٤٤	0.٣٣	0.٣٠	0.١١	0.١٠	0.٣٧	0.35

جدول (7)

احتمال خطأ التصنيف لبيانات الجلطة الدماغية المجموعة الثانية

LDF	QDF	RL.S	RQ.S	RL.Huber	RQ.Huber	RL. Cambpe1l	RQ . Cambpe1l
0.٤٣	0.٣٣	0.٢٤	0.٢١	0.١٠	0.٠٥	0.٢٨	0.٢٥

ويلاحظ من الجدولين (٦،٧) إن مقدّر(M) الحصين باستعمال دالة الوزن (Huber) ، لبيانات (الجلطة القلبية والجلطة الدماغية) أمراض ارتفاع نسبة الكولسترول في الدم للمجموعتين الأولى والثانية تفوق بإعطائه اقل احتمال لخطأ التصنيف

الاستنتاجات :

في ظل تحليل نتائج الجانب التطبيقي تم التوصل إلى الاستنتاجات الآتية :

١. لوحظ أن مقدّر (M) الحصين باستعمال دالة الوزن (Huber) حقق نتائج كفوءة بدرجة عالية إذا أعطى أقل احتمال لخطأ التصنيف للدالة التمييزية التربيعية الحصينة (RQDF) فالدالة التمييزية الخطية الحصينة (RLDF) لأمراض ارتفاع نسبة الكولسترول في الدم للمجموعتين الأولى والثانية ، يليه مقدّر (S) الحصين ثم مقدّر (M) الحصين باستعمال دالة الوزن (Cambpe1l) .

٢. أن تقدير مقدّر (M) الحصين باستعمال دالة الوزن (Huber) كاف ويغني عن استخدام مقدري (S) الحصين ، مقدّر (M) الحصين باستعمال دالة الوزن (Cambpe1l) .

التوصيات:

١. نوصي باستخدام مقدّر (M) الحصين باستعمال دالة الوزن (Huber) الحصين وفي حالة تعدد المجتمعات.

٢. نوصي بتطبيق الأساليب الأخرى في الدوال التمييزية كالدالة التمييزية اللوجستية .

المصـــادر

١. تشارلز، وأيرتج وآخرون، (1970). "الموسوعة الطبية الحديثة "الطبعة الثانية ،مؤسسة سجل العرب ،القاهرة .

٢ . العلوي ، لقاء علي محمد ، (2003) ، "مقارنة مقدرات التباين المشترك الحصينة في تحليل المركبات الرئسية " ، أطروحة دكتوراه في الاحصاء ، كلية الأدارة والأقتصاد ، جامعة بغداد.

٢. النداوي ، سرى صباح كيتب ، (2008) ." مقارنة بعض المقدرات الحصينة في الدوال التمييزية مع تطبيق عملي " رسالة ماجستير في الإحصاء ، كلية الإدارة والاقتصاد، جامعة بغداد .

٤. الياسين ، دريد حسين بدر ، (٢٠٠٩) ." أستخدام بعض طرائق التمييز الحصينة لتشخيص أمراض سرطان الدم (اللوكيميا) " رسالة ماجستير في الإحصاء ، كلية الإدارة والاقتصاد ، جامعة المستنصرية .

5. Croux,C.& Hasbrouck, G.,(2001) . "A Note on finite–sample efficiencies of Estimators for the Minimum Volume Ellipsoid ", Submitted.

6. Hubert, M., Driessen, K.V., (2004)."Fast and Robust Discriminant Analysis", Computational statistics and Data Analysis, vol .45, 301-20.

7. Neil.H.Timm.(2002) .Applied Multivariate analysis ,John Wiley & sons.

8. Pison, G., Van Aelst, S., and Willems, G. (2002b) . "Small Sample Corrections for LTS and MCD," Metrika, 55, 111-123.

9. Rousseeuw,P.J., (1985) ."Multivariate Estimation with High Breakdown Point", Mathematical Statistics and Application, B., pp. 283-397.

10. Rousseeuw,P.J.&Leroy,A.M.,(1987). "Robust Regression and Outlier Detection", John Wiley & Sons, New York.

11. Rousseeuw, P.J., &KatrinV.D.,(1999) ."A fast Algorithm for the Minimum Covariance Determinant Estimator", Techno metrics, 41, pp. 212-223 .