مقارنه بين بعض طرائق التصنيف الخطية مع تطبيق عملي

أ.م.د. حمزة اسماعيل شاهين
قسم الاحصاء/ كلية الادارة والاقتصاد

المستخلص

يعد تحليل التمييز الخطي والانحدار اللوجستي من أهم طرائق التحليل الإحصائي متعدد المتغيرات استخداماً في مجال تحليل البيانات المصنفه (Categorical data) واللذين يمكن عدهما تطويراً لنماذج التصنيف الخطية .ان استخدام تحليل التمييز الخطي يتطلب ان تكون بيانات المتغيرات التوضيحية ذات توزيع طبيعي متعدد المتغيرات.في حين الانحدار اللوجستي لايشترط ايه افتراضات تخص توزيع المتغيرات التوضيحيه، اذ يعد الانحدار اللوجستي أكثر مرونه وأكثر حصانه في حاله عدم تحقق الافتراضات الأساسية للتحليل التمييزي الخطي.

في هذا البحث تم التركيز على أجراء مقارنات بين ثلاث صيغ لتصنيف بيانات تعود لمجموعتين في حاله متغير الاستجابة مؤلف من مستتويين (تصنيفين) فقط ، الصيغه الأولى كانت دالة التمييز الخطي في حالة المجموعتيين تتبعان التوزيع الطبيعي متعدد المتغيرات والثانيه فكانت الصيغه الاحتماليه التي اشتقت كبديل لدالة التمييز الخطي ،اما الثالثة فكانت الصيغه الاحتمالية لنموذج الانحدار اللوجستي ثنائي الاستجابه وتمت المقارنة بين هذه الصيغ الثلاثة على وفق معيار احتمال خطأ التصنيف (Misclassification) . واثبتت النتائج ان التصنيف بحسب الصيغة الاحتمالية لنموذج الانحدار اللوجستي تعطي أقل احتمال لخطأ التصنيف من خلال التطبيق على بيانات تخص نوعين من امراض سرطان الدم(اللوكيميا) .

المصطلحات الرئيسية للبحث/ الدالة التمييزية الخطية – الانحدار اللوجستي ثنائي الاستجابة – احتمال خطأ التصنيف.

مجلة العلوم
الاقتصادية والإدارية
المجلد 20
العدد ٨٠
لسنة 2014
الصفحات٣٩٣-٤١٠

١- المقدمة وهدف البحث

ازدادت في السنوات الاخيرة أهمية استخدامات التصنيف ولاسيما في المجالات الطبية والاجتماعية والزراعية وعلوم الارض وغيرها من المجالات الاخرى .

ويعتبر تحليل التمييز الخطي Linear discriminant analysis والانحدار اللوجستي (logistic regression) من اهم طرائق التحليل الاحصائي متعدد المتغيرات استخداماً في مجال تحليل البيانات المصنفة واللذين يمكن عدهما تطويراً لنماذج التصنيف الخطية[١٦] .

أن استخدام التحليل التمييزي الخطي يتطلب توفر عدد من الافتراضات اهمها ان تكون بيانات المتغيرات التوضيحية ذات توزيع طبيعي متعدد المتغيرات في حين الانحدار اللوجستي لا يشترط ايه فروض تخص توزيع المتغيرات التوضيحيه ،اذ يعد الانحدار اللوجستي طريقه اكثر مرونه ، واكثر حصانه في حالة عدم تحقق الافتراضات الاساسيه للتحليل التمييزي الخطي[٧].

وتم التركيز على مسألة الاختيار ما بين طريقتين لتصنيف البيانات في حالة متغير الاستجابة مؤلف من مستتويين (تصنيفين) فقط هما طريقة دالة التمييز الخطي وطريقة نموذج الانحدار اللوجستي ثنائي الاستجابة .

ان هذا البحث يهدف بالدرجة الاساس الى اجراء المقارنه بين ثلاث صيغ لتصنيف البيانات تعود لمجموعتين الصيغه الاولى تتمثل بدالة التمييز الخطي في حال كانت المجموعتان تتبعان التوزيع الطبيعي متعدد المتغيرات والثانية الصيغه الاحتمالية التي تم اشتقاقها كبديل لدالة التمييز الخطي اما الصيغه الثالثة فكانت الصيغة الاحتمالية لنموذج الانحدار اللوجستي ثنائي الاستجابة . واجراء المقارنه بين هذه الصيغ الثلاثة على وفق معيار احتمال خطأ التصنيف (misclassification)

ويهدف البحث كذلك الى الوصول الى افضل نموذج خطي للتشخيص وللتمييز بين نوعين من امراض سرطان الدم (اللوكيميا)، سرطان الدم (النخاعي) الحاد وسرطان الدم (اللمفاوي) الحاد .

٢- التحليل التمييزي Discriminant analysis

يعد التحليل التمييزي (Discriminant Analysis) من إساليب تحليل الاحصائي متعدد المتغيرات التي تهتم بفصل مجموعات مختلفة من المفردات (أو المشاهدات) وبتوزيع المفردات (أو المشاهدات) الجديدة على مجموعات سبق تعريفها[٨].

(2-1) دالة التمييز الخطيه لمجتمعين يتبعان التوزيع الطبيعي متعدد المتغيرات [٩]

لنفترض لدينا مجتمعان( ) ولكل منهما داله كثافه احتماليه $f(\underline{x}|G_0)$ ، $f(\underline{x}|G_1)$ على الترتيب G_1, G_0

ولنفرض ان (π_1, π_0) هي احتمالات اوليه (Prior probability) ان المشاهده $\underline{x}$ تاتي من (G_1, G_0) على الترتيب وعلى افتراض ان دوال الكثافه تتبع توزيعاً طبيعياً متعدد المتغيرات تحت حاله تساوي مصفوفتا التباين والتباين المشترك $(\sum_0 = \sum_1)$

اذ أن :-

النسبه لدالتي الكثافه هي

$$h(\underline{x}) = \frac{\frac{\pi_0}{(2\pi)^{n/2}\,|\Sigma|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_0)'\Sigma^{-1}(\underline{x}-\underline{\mu}_0)\right]}{\frac{\pi_1}{(2\pi)^{n/2}\,|\Sigma|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)\right]} \qquad (1)$$

$$h(\underline{x}) = exp\left\{-\frac{1}{2}\left[-2\underline{\mu}'_0\,\Sigma^{-1}\underline{x} + 2\underline{\mu}'_1\,\Sigma^{-1}\underline{x} + \underline{\mu}'_0\,\Sigma^{-1}\underline{\mu}_0 - \underline{\mu}'_1\,\Sigma^{-1}\underline{\mu}_1\right]\right\} * \frac{\pi_0}{\pi_1} \qquad (2)$$

باضافة وطرح ($\underline{\mu}'_1\,\Sigma^{-1}\underline{\mu}_1$) للطرف الايمن للمعادله (2) ينتـج:

$$h(\underline{x}) = exp\left[(\underline{\mu}_0 - \underline{\mu}_1)'\Sigma^{-1}\underline{x} - \frac{1}{2}(\underline{\mu}_0 + \underline{\mu}_1)'\,\Sigma^{-1}(\underline{\mu}_0 - \underline{\mu}_1)\right] * \frac{\pi_0}{\pi_1} \qquad (3)$$

وحيث ان لوغاريتم الداله المذكور انفاً يعطي نتيجه الدالة نفسها فللبساطة تكتب المعادله (3) كالاتي:-

$$\ln h(\underline{x}) = \left[(\underline{\mu}_0 - \underline{\mu}_1)'\Sigma^{-1}\underline{x} - \frac{1}{2}\left(\underline{\mu}_0 + \underline{\mu}_1\right)'\,\Sigma^{-1}\left(\underline{\mu}_0 - \underline{\mu}_1\right) + \ln\frac{\pi_0}{\pi_1}\right] \qquad (4)$$

يمثل الحد الأول من المعادله (4) داله التمييز الخطي لفشر ، إما الحد الثاني فيمثل نقطة الفصل بين المجموعتين G_1, G_0.

ان المعالم المجهوله $\underline{\mu}_1 , \underline{\mu}_0$, Σ في المعادله (4) يتم تقديرها من بيانات العينة بطريقة الامكان الاعظم (Maximum Likelihood) حيث افضل تقدير لـ $\underline{\mu}_0$ هو $\overline{\underline{x}}_0$ ولـ$\underline{\mu}_1$ هو $\overline{\underline{x}}_1$ اذ[٣]:-

$$\overline{\underline{x}}_0 = \frac{\sum_{i=1}^{n_0}\underline{x}_{0i}}{n_0} \qquad (5)$$

$$\overline{\underline{x}}_1 = \frac{\sum_{i=1}^{n_1}\underline{x}_{1i}}{n_1} \qquad (6)$$

اما مصفوفه التباين والتباين المشترك المدمجه Σ فيكون تقديرها بمصفوفه التباين والتباين المشترك المدمجه للعينه S_p وبحسب الصيغه الاتية.

$$S_p = \frac{(n_0 - 1)\,S_0 + (n_1 - 1)\,S_1}{n_0 + n_1 - 2} \qquad (7)$$

حيث ان S_1, S_0 مصفوفتا التباين والتباين المشترك التقديريه للمجموعتين G_1, G_0

اما π_1, π_0 فيتم تقديرها باسلوبين اما بافتراضهما متساويين اي $(\pi_0 = \pi_1)$ للمجتمعين او انهما يقدران بحسب الاتي:-

$$\hat{\pi}_0 = \frac{n_0}{n}$$
$$\hat{\pi}_1 = \frac{n_1}{n} \qquad (8)$$

اذ ان

n_0 :حجم العينه المسحوبه من المجتمع G_0

n_1 :حجم العينه المسحوبه من المجتمع G_1

وان

$$n = n_0 + n_1$$

وبالتعويض عن المعالم المجهوله في المعادله (4) نحصل على الاحصاءة الاتية:-

$$W = \left(\bar{x}_0 - \bar{x}_1\right)' S_p^{-1} \underline{x} - \frac{1}{2}\left(\bar{x}_0 + \bar{x}_1\right)' S_p^{-1}\left(\bar{x}_0 - \bar{x}_1\right) + \ln\frac{\hat{\pi}_0}{\hat{\pi}_1} \qquad (9)$$

وفي حالة ($\pi_0 = \pi_1$) فان المعادله (9) تصبح

$$W^* = \left(\bar{x}_0 - \bar{x}_1\right)' S_p^{-1} \underline{x} - \frac{1}{2}\left(\bar{x}_0 + \bar{x}_1\right)' S_p^{-1}\left(\bar{x}_0 - \bar{x}_1\right) \qquad (10)$$

وان قاعده التصنيف ستكون بأن تصنف المشاهده $\underline{x}$ الى المجموعه G_0 اذا كان $W > 0$ او $W^* > 0$

والى المجموعه G_1 اذا كان $W \le 0$ او $W^* \le 0$

(٢-2) **دالة التمييزية الخطية وفقاً لصيغه مسافه (مقياس) مهالونوبيس**[٨]

بالامكان ايجاد دوال تمييزية خطية في حالة لدينا K من المجاميع من خلال التوسع في استخدام صيغه مسافه مهالونوبيس D_i^2 اذ ان:

$$D_i^2 = \left(\underline{x} - \bar{x}_i\right)' S_p^{-1} \left(\underline{x} - \bar{x}_i\right) \qquad (11)$$

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

والصيغه (١١) يمكن ان تكتب كالاتي

$$D_i^2(\underline{x}) = \underline{x}' S_P^{-1} \underline{x} - \underline{x}' S_P^{-1} \bar{x}_i - \bar{x}_i S_P^{-1} \underline{x} + \bar{x}_i S_P^{-1} \bar{x}_i$$
$$= \underline{x}' S_P^{-1} \underline{x} - 2\bar{x}_i S_P^{-1} \underline{x} + \bar{x}_i S_P^{-1} \bar{x}_i \qquad (12)$$

الحد الاول من الطرف الايمن للصيغه (12) يكون ثابتاً لكل المجاميع ، والحد الثاني عباره عن دالة خطيه في الموجه $\underline{x}$ ،اما الحد الاخير فانه لا يعتمد على الموجه $\underline{x}$ وبالتالي يمكن اهمال الحد الاول من الصيغه (12) والحصول على دالة تصنيف خطيه ويعبر عنها $d_i(\underline{x})$ ولكي تكون هذه الصيغه متوافقه مع داله التمييز الخطي في الفقره (2-1) يتم ضرب هذه الداله بالمقدار ($-\frac{1}{2}$) فتكون داله التمييز الخطي كالاتي :-

$$d_i(\underline{x}) = \bar{x}_i' S_P^{-1} \underline{x} - \frac{1}{2}\bar{x}_i' S_P^{-1} \bar{x}_i \qquad (13)$$

وفي حالة افتراض دوال الكثافة الاحتمالية لكل مجموعه تتبع التوزيع الطبيعي متعدد المتغيرات تحت حالة تساوي مصفوفات التباين والتباين المشترك وباحتمالات اوليه $\pi_1, \pi_2, \ldots\ldots, \pi_k$.

$$d_i^*(\underline{x}) = \ln \hat{\pi}_i + \bar{x}_i' S_P^{-1} \underline{x} - \frac{1}{2}\bar{x}_i' S_P^{-1} \bar{x}_i \qquad (14)$$

وتكون قاعده التصنيف بان تصنف المشاهده $\underline{x}$ الى المجموعه i اذا كانت قيمته الدالة $d_i(\underline{x})$ اكبر ما يمكن من بقيه المجاميع الاخرى .

وبالاسلوب نفسه في حالة دوال الكثافة الاحتمالية تتبع التوزيع الطبيعي متعدد المتغيرات فان المشاهده $\underline{x}$ تصنف بانها تعود للمجموعه i اذا كانت قيمته $d_i^*(\underline{x})$ اكبر ما يمكن من بقيه المجاميع الاخرى.

(٣-2)**الصيغه الاحتمالية لدالة التمييز الخطي**

يمكن اشتقاق صيغه بديله لدالة التمييز الخطي تعرف بالصيغه الاحتماليه وحسب الاتي:-

لنفرض لدينا مجموعتين G_1, G_0 ولكل منها دالة كثافة احتمالية (p.d.f) هي $f(\underline{x}|G_1)$ ، $f(\underline{x}|G_0)$ ولنفترض ان (π_1, π_0) هي احتمالات اوليه ان متجه المشاهدة $\underline{x}$ تأتي من المجتمع (G_1, G_0) على الترتيب فأن التوزيع اللاحق ان $\underline{x}$ تأتي من المجموعة G_1 يكون .

$$P_r(G_1/\underline{x}) = \frac{\pi_1 f(\underline{x}|G_1)}{\pi_0 f(\underline{x}|G_0) + \pi_1 f(\underline{x}|G_1)} \qquad (15)$$

والذي يمكن كتابته كالآتي:

بافتراض ان دوال الكثافة الاحتمالية $f(\underline{x}|G_0)$ ، $f(\underline{x}|G_1)$ دوال تتبع التوزيع الطبيعي متعدد المتغيرات تحت حالة افتراض تساوي مصفوفتا التباين والتباين المشترك $(\Sigma_1 = \Sigma_0)$

عندئذ يمكن كتابة.

$$P_r(G_1/x) = \frac{1}{1 + \frac{\pi_0}{\pi_1} * \frac{exp\left(-\frac{1}{2}\left(\underline{x} - \underline{\mu}_0\right)' \Sigma^{-1}\left(\underline{x} - \underline{\mu}_0\right)\right)}{exp\left(-\frac{1}{2}\left(\underline{x} - \underline{\mu}_1\right)' \Sigma^{-1}\left(\underline{x} - \underline{\mu}_1\right)\right)}}$$

$$= \frac{1}{1 + \frac{\pi_0}{\pi_1} * exp\left[-\frac{1}{2}\left(\underline{x} - \underline{\mu}_0\right)' \Sigma^{-1}\left(\underline{x} - \underline{\mu}_0\right) + \frac{1}{2}\left(\underline{x} - \underline{\mu}_1\right)' \Sigma^{-1}\left(\underline{x} - \underline{\mu}_1\right)\right]}$$

باستخدام الطرق الجبرية تتوصل الى ان

$$P_r(G_1/\underline{x}) = \frac{1}{1 + \left\{e^{Log\frac{\pi_0}{\pi_1} - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_0)' \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_0) + (\underline{\mu}_1 - \underline{\mu}_0)' \Sigma^{-1} \underline{x}}\right\}^{-1}} \qquad (16)$$

والمعادلة(16) يمكن تكتب بالشكل الاتي

$$P_r(G_1/\underline{x}) = \frac{1}{1 + \left\{e^{\alpha + \underline{\beta}\,\underline{x}}\right\}^{-1}} \qquad (17)$$

إذ أن:

$$\alpha = Log\frac{\pi_0}{\pi_1} - \frac{1}{2}\left(\underline{\mu}_1 + \underline{\mu}_0\right)' \Sigma^{-1}\left(\underline{\mu}_1 - \underline{\mu}_0\right) \qquad (18)$$

$$\underline{\beta} = \left(\underline{\mu}_1 - \underline{\mu}_0\right)' \Sigma^{-1} \qquad (19)$$

حيث ان المعالم $(\pi_1, \pi_0, \sum, \underline{\mu}_1, \underline{\mu}_0)$ تكون غير معلومة ويجب تقديرها من بيانات العينة وفق الصيغ(5)،(6)، (7) ،(8) على الترتيب.

وبالتالي فأن تقدير احتمال المشاهده $\underline{x}$ تنتمي للمجموعه G_1 هو

$$\hat{P}_r(G_1/\underline{x}) = \frac{1}{1+\left\{e^{\hat{\alpha}+\hat{\underline{\beta}}\,\underline{x}}\right\}^{-1}} \qquad (20)$$

وتقدير احتمال المشاهدات $\underline{x}$ تنتمي للمجموعة G_0 هو:

$$\hat{P}_r(G_0/\underline{x}) = 1 - \hat{P}_r(G_1/\underline{x}) \qquad (21)$$

وان قاعده التصنيف على وفق صيغه الدالة الاحتمالية في حالة لدينا مجموعتين تكون :

ان المشاهده $\underline{x}$ تعود للمجموعه G_0 اذا كان

$$\hat{P}(G_0/\underline{x}) > \hat{P}(G_1/\underline{x})$$

وتعود المشاهده $\underline{x}$ للمجموعه G_1 اذا كان

$$\hat{P}(G_0/\underline{x}) < \hat{P}(G_1/\underline{x})$$

(3) التحليل اللوجستي Logistic Analysis

في السنوات الاخيرة ازدادت اهمية التحليل اللوجستي في مجال تحليل البيانات المصنفه وخاصة في مجالات البحوث الطبية والاجتماعية والزراعية وغيرها، وذلك لكونه يهتم بتحليل البيانات ذات الاستجابة الثنائية والتي يكون فيها متغير الاستجابة (Response Variable) ثنائياً (Binary) ، حيث حالة النجاح (Success) يكون فيها متغير الاستجابة يأخذ القيمة (1) وحالة الفشل (Failure) يأخذ القيمة (0)[٦] .

(3-١): نموذج الانحدار اللوجستي ثنائي الاستجابه

يبنى نموذج الانحدار اللوجستي على فرض اساسي هو ان المتغير التابع (y) متغير الاستجابة الذي تهتم بدراسته هو متغير ثنائي يتبع توزيع بيرنولي (Bernoulli)، يأخذ القيمة (1) باحتمال مقداره (π) والقيمة (0) باحتمال ($1-\pi$) اي الى حدوث الاستجابة وعدم حدوثها[١٢] وكما نعلم في الانحدار الخطي الذي نأخذ متغيراته المستقلة والمتغير التابع قيماً مستمرة فأن النموذج الذي يربط بين المتغيرات هو على النحو الآتي :

$$y = B_0 + B_1 x + \epsilon \qquad (22)$$

إذ أن (y) يمثل متغيراً مشاهداً مستمراً وبفرض ان متوسط قيم (y) المشاهدة أو الفعلية عند قيمة معينة للمتغير (x) هي E(y) فأنه يمكن كتابة النموذج على النحو الآتي :

$$E(y/x) = \beta_0 + \beta_1 x \qquad (23)$$

ومن المعروف في الانحدار ان الطرف الايمن لهذا النموذج يأخذ قيماً من ($-\infty$) الى (∞) ولكن عندما يكون لدينا متغيران احدهما ثنائي (y) فأن نموذج الانحدار الخطي البسيط لايكون ملائماً لأن:

$$E(y/x) = P_r(y=1) = \pi \qquad (24)$$

وبذلك تكون قيمة الطرف الايمن محصورة مابين الرقمين (1,0) وبذلك يكون النموذج غير قابل للتطبيق من وجهة نظر الانحدار .وأن احدى طرائق حل هذه المشكلة هو ادخال تحويلة رياضية مناسبة على المتغير التابع .

(y) ومن المعروف ان ($0 \le \pi \le 1$) ومن ثم النسبة ($\frac{\pi}{1-\pi}$) هي عبارة عن مقدار موجب محصور بين (0 و∞) اي ($0 \le \frac{\pi}{1-\pi} \le \infty$) وبأخذ اللوغارتيم الطبيعي للاساس (e) للتحويل ($\frac{\pi}{1-\pi}$) فأن مجال قيمه تصبح محصوره ($-\infty \le log_e\left(\frac{\pi}{1-\pi}\right) \le \infty$) وعليه يمكن كتابة نموذج الانحدار في حالة متغير مستقل واحد على النحو الآتي:

$$log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x \qquad (25)$$

واذا كان لدينا P من متغيرات المستقلة فأن النموذج يكون

$$log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \qquad (26)$$

اذ ان $i=1,2,\ldots,n$

$j=1,2,\ldots,p$

ويسمى هذا النموذج بنموذج الانحدار اللوجستي ثنائي الاستجابة وتسمى التحويلة $log_e\left(\frac{\pi}{1-\pi}\right)$ او $ln\left(\frac{\pi}{1-\pi}\right)$ بتحويلة لوجيت (Logit transformation) وان الدالة الوجستية هي دالة مستمرة تأخذ القيم ما بين (0 , 1) وتقترب y من الصفر كلما اقترب الطرف الايمن للدالة الوجستية من ($-\infty$) وتقترب y من الواحد كلما اقترب الطرف الايمن لهذه الدالة من (∞) وهي دالة متماثلة عندما يكون الطرف الايمن لهذه الدالة مساوياً للصفر[١٣].

تسمى النسبه ($\frac{\pi}{1-\pi}$) بنسبه الافضليه أو أفضلية النجاح (odds of success) أو نسبة الافضلية للحدث المرغوب والنسبة ($\frac{\pi}{1-\pi}$) يمكن ان تسمى ايضاً نسبة افضلية الفشل (odds of failure) وان المقدار $log_e\left(\frac{\pi}{1-\pi}\right)$ يسمى لوغارتيم نسبة الا فضلية (Log odds ratio) أو اللوجيت (Logit)[١٦].

(2-3)الصيغة الاحتمالية لنموذج الانحدار اللوجستي ثنائي الاستجابة

يمكن كتابة نموذج الانحدار اللوجستي ثنائي الاستجابة بالصيغه الاحتمالية وذلك برفع طرفي المعادلة (26) للاساس (e) ونحصل

$$\left(\frac{\pi}{1-\pi}\right) = e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}} \qquad (27)$$

$$\frac{1}{\frac{1}{\pi} - 1} = \frac{1}{e^{(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})^{-1}}} \qquad (28)$$

وباستخدام الطرائق الجبرية فان المعادلة (٢٨) يمكن ان تكتب كالاتي:-

$$\pi = \frac{1}{1 + e^{(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})^{-1}}} \qquad (29)$$

وبالتالي فان احتمال ان متغير الاستجابة y يأخذ القيمه (١) يكون

$$p(y = 1/\underline{x}) = \frac{1}{1 + e^{(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})^{-1}}} \qquad (30)$$

واحتمال ان متغير الاستجابة y ياخذ القيمه (0) يكون

$$p(y = 0/x) = 1 - p(y = 1/x) \qquad (31)$$

وان الصيغ التقديرية للمعادلتين (31,30) هي

$$\hat{p}(y = 1/\underline{x}) = \frac{1}{1 + e^{(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij})^{-1}}} \qquad (32)$$

وان

$$\hat{p}(y = 0/\underline{x}) = 1 - \hat{p}(y = 1/\underline{x}) \qquad (33)$$

حيث ان $\hat{\beta}_0, \hat{\beta}_1, \ldots\ldots\ldots\ldots, \hat{\beta}_p$ تمثل مقدرات الامكان الاعظم للمعالم المجهولة $\beta_0, \beta_1, \ldots\ldots\ldots\ldots, \beta_p$.[٢]

وان قاعده التصنيف وفق الصيغه الاحتمالية لنموذج الانحدار اللوجستي ثنائي الاستجابة تكون:-

ان المشاهده $\underline{x}$ تعود للمجموعه G_0 اذا كان

$$\hat{p}(y = 0/\underline{x}) > \hat{p}(y = 1/\underline{x})$$

وتعود المشاهده $\underline{x}$ تعود للمجموعه G_1 اذا كان

$$\hat{p}(y = 0/\underline{x}) < \hat{p}(y = 1/\underline{x})$$

وتجدد الاشارة الى ان نموذج الانحدار اللوجستي لايشترط توفر الافتراضات الآتيـة[١٢]

١- وجود علاقة خطية مابين المتغيرات التوضيحية والمتغير المعتمد .

٢- ان يكون توزيع المتغيرات توزيعاً طبيعياً .

٣- تحقق خصيصة ثبات التباين (تجانس التباينات) .

٤- الجانب التطبيقي

اعتمدت في هذا الجانب بيانات جمعت من المستشفى التعليمي والمستشفى الجمهوري في محافظة البصرة تعود الى (160) مريض مصابين بسرطان الدم (اللوكيميا) الحاد للسنوات (2007,2006,2005) والتي تمثلت بطبلات المصابين التي كانت تحتوي معلومات كاملة وتضمنت الدراسة سحب عينتان عشوائيتان متكونة من (80) مريض لكل عينـه، تمثل هاتين العينتين مجموعتين من المرضى المصابين بسرطان الدم الحاد ،حيث تمثل المجموعه الاولى المرضى المصابين بسرطان الدم النخاعي الحاد (AML) والمجوعة الثانية تمثل المرضى المصابين بسرطان الدم اللمفاوي الحاد (ALL) .

وقد اعتبرت الاصابة بهذه الامراض تمثل متغير الاستجابة (Y) (response variable) حيث اعطيت القيمة (0) للمريض المصاب بسرطان الدم اللنخاعي الحاد والقيمة (١) للمريض المصاب بسرطان الدم اللمفاوي الحاد. وقد حددت المتغيرات التوضيحيه التي لها علاقة في الاصابة بهذا المرض بالاتفاق مع الاطباء اصحاب الاختصاص والتي هي كالاتي :

١- الجنس (Sex) وصنف الى (ذكر =١ ، انثى=٢).

٢- العمر (age) وصنف الى $(1 = 50 \leq , 2 = 50 \succ)$.

٣- الوزن(Weight) صنف الى $(1 = 40 \leq , 2 = 40 \succ)$.

٤- نسبة كريات الدم الحمراء(P.C.V) وصنفت الى $(1 = 35 \leq , 2 = 35 \succ)$.

٥- هيموجلوبين الدم(H.B) وصنف الى $(1 = 11.5 \leq , 2 = 11.5 \succ)$.

٦- معدل عدد كريات الدم البيضاء (W.B.C) وصنف الى $(1 = 7.5 \leq , 2 = 7.5 \succ)$.

٧-نسبة سرعة ترسب كريات الدم الحمراء (E.S.R) وصنف الى $(1 = 22.5 \leq , 2 = 22.5 \succ)$.

٨-عدد الصفائح الدمويه (P.C) وصنفت الى $(1 = 150 \leq , 2 = 400 \succ)$.

وقد عرضت بيانات العينتين لمجموعتين المرضى المصابين بسرطان الدم النخاعي الحاد (AML) والمصابين بسرطان الدم اللمفاوي الحاد (ALL) في الملحق (الجدولين (١)،(٢))

(٤-١) التحليل الاحصائي للدالة التمييزيه الخطيه

ان من اهم الشروط الاساسيه لاستخدام الدالة التمييزيه الخطيه تحقق (فرضية التوزيع الطبيعي،معنويه الدالة التمييزيه ، تجانس التباينات).

وبأستخدام البرنامج الاحصائي (SPSS.V20) تم اختبار البيانات الخاصه بالمتغيرات التوضيحيه الثمانيه باعتماد اختبار كولمكروف_ سميرنوف واظهرت النتائج كما في الجدول (١) ان اغلب متغيرات الدراسة تتوزع طبيعياً ماعدا متغيرات ((الجنس x_1) ، الوزن (x_3)،عدد الصفائح الدموية (x_8)).

وبالنظر لكون حجم البيانات لتلك المتغيرات تجاوز (30) مشاهده فانه يمكن اعتبار بياناتها تتوزع بالتقريب التوزيع الطبيعي وذلك حسب نظرية الغاية المركزية.

جدول رقم(١) نتائج اختبار البيانات للتوزيع الطبيعي

Variables	Kolmogorov -Smirnov		
	Statistic	df	Sig.
(x_1) sex	٢.٢١٤	١٦٠	0.000
age (x_2)	٠.١٥٨	١٦٠	1.000
Weight (x_3)	1.581	١٦٠	0.013
P.C.V (x_4)	٠.١٥٨	١٦٠	1.000
H.B (x_5)	٠.٠٠٠	١٦٠	1.000
W.B.C (x_6)	٠.٤٧٤	١٦٠	0.978
E.S.R (x_7)	٠.٣١٦	١٦٠	1.000
P.C (x_8)	١.٦٦٠	١٦٠	0.008

ولتكوين دالة تمييزية مقبولة احصائياً بمستوى معنوية تم اختبار معنويه الفروق بين متوسطي المجموعتين قيد الدراسة وفق الفرضيه الاتية :

$$H_0 : \underline{\mu}_0 = \underline{\mu}_1$$

$$H_1 : \underline{\mu}_0 \neq \underline{\mu}_1$$

واظهرت النتائج كما في الجدول (٢) وبالاعتماد على مقياس (Wilks) واحصاءة (χ^2) وجود فروق معنوية بين المتوسطين وهذا يعني لدينا مجموعتين مختلفتين من المرضى وبالتالي يكون للدالة التمييزية القدرة على التفريق (الفصل) ما بين تلك المجموعتين وتصنيف ايه مفرده جديدة لاحدى هاتين المجموعتين.

جدول (٢) اختبار معنويه الداله التمييزيه

Teast fuctions	Wilks' Lamda	Chi-Squar	df	Sig.
1	0.670	61.611	٨	0.000

ولاختبار تجانس التباين بين المجموعتين فقد تم اختبار الفرضية الاتية :-

$$H_0 : \Sigma_0 = \Sigma_1$$

$$H_1 : \Sigma_0 \neq \Sigma_1$$

ولوحظ من نتائج جدول (٣) باستخدام احصاءة (Box's M) ان قيمة Sig> 0.05 وهذا يشير الى قبول فرضية العدم H_0 ويدل ذلك على تجانس التباينات للمجموعتين وهذا يعني تحقق شرط استخدام الدالة التمييزية الخطية

جدول (٣) اختبار تجانس التباينات بين المجموعتين

Box's M	29.784
F Approx.	1.389
Df1	36
Sig	0.114

وللحصول على دالتي تمييز مثلى للمجموعتين تم اختبار معنوية المتغيرات التوضيحية الثمانيه لمعرفة اهمية كل متغير بشكل منفرد ومدى تاثيره على متغير الاستجابة ومن خلال نتائج جدول (٤) يتضح ان متغيرات (الجنس (x_1) ، الوزن (x_3)،عدد الصفائح الدموية (x_8))لها تاثير واهميه في تكوين وبناء الدالة التمييزيه حيث كانت قيم المعنويه في العمود السادس من جدول رقم (٤) اقل من (0.05) في حين لم تظهر بقيه المتغيرات اي تأثير معنوي.

جدول (٤) اختبار معنويه المتغيرات التوضيحيه

Variables	Wilks' Lamda	F	df1	Df2	Sig.
(x_1) sex	0.٧٩٩	39.754	1	158	0.000
age (x_2)	0.999	0.107	1	158	0.744
Weight (x_3)	0.927	12.392	1	158	0.001
P.C.V (x_4)	0.999	0.101	1	158	1.751
H.B (x_5)	1.000	0.000	1	158	1.000
W.B.C (x_6)	0.993	1.122	1	158	0.291
E.S.R (x_7)	0.998	0.396	1	158	0.530
P.C (x_8)	0.927	12.447	1	158	0.001

وبعد تقدير موجهي متوسطي العينتين ومصفوفة التباين والتباين المشترك المدمجه باستخدام مقدرات الامكان الاعظم وفق الصيغ (5) ،(6) ،(7) فقد تم تقدير دالتي التمييز الخطيه للمجموعتين وفق الصيغه (14) وعرضت النتائج في جدول رقم (5) .

جدول (٥) الدوال التمييزية الخطية التقديرية

Variables	الداله التمييزية الاولى	الداله التمييزية الثانية
Constant	-57.733	-66.876
(x_1) sex	18.630	21.972
age (x_2)	9.658	9.963
Weight (x_3)	7.809	9.092
P.C.V (x_4)	12.608	14.210
H.B (x_5)	-1.172	-1.992
W.B.C (x_6)	11.304	11.833
E.S.R (x_7)	8.996	9.283
P.C (x_8)	6.371	4.975

يلاحظ من جدول (٥)ان متغيرات (العمر (x_2) ، هيموجلوبين الدم H.B (x_5) ، معدل عدد كريات الدم البيضاء W.B.C (x_6)) كانت لها نفس قيم المعاملات في الدالتين وهذا يدل على ان هذه المتغيرات سوف لن يكون لها اي دور مهم في عملية تصنيف البيانات .

وكانت الخطوة الاخيره لهذا التحليل تصنيف مفردات (مشاهدات) مجموعتين المرضى المصابين بسرطان الدم وفق دوال التمييز التقديريه المذكوره في جدول (٥) وكما هو معلوم ان عمليه التصنيف (Classification) قد تؤدي الى الوقوع فيما يعرف بخطأ التصنيف (Misclassification) وهو احتمال تصنيف مفردة معينة الى المجموعة الاولى بينما هي في الحقيقه تعود للمجموعه الثانية او بالعكس وقد كانت نتائج تصنيف البيانات حسب دالتي التمييز الخطية ملخصة في الجدول (٦).

جدول (٦) تصنيف المشاهدات حسب دالتي التمييز الخطية

الحالة	التصنيف		
	اصبح المريض عائد الى المجموعه الاولى (٠)	اصبح المريض عائد الى المجموعه الثانيه (١)	نسبة التصنيف الصحيح
المريض عائد الى المجموعه الاولى (٠)	53	27	66.2%
المريض عائد الى المجموعه الثانيه (١)	19	61	76.3%
نسبه التصنيف الكلي	45.6%	54.4%	71.2%

يلاحظ من الجدول (٦) ان احتمال التصنيف الصحيح لمريض يعود الى المجموعة الاولى (٠) كانت 66.2% ويعود للمجموعه الثانية (١) هي 76.3% في حين احتمال خطأ التصنيف للمجموعة الاولى 33.8% وللمجموعة الثانية 23.7% وقد بلغت نسبة التصنيف الصحيح الكلية % 71.2 ونسبة التصنيف الخاطىء الكلية 28.8% .

كذلك جرى تصنيف البيانات بأستخدام الصيغ الاحتماليه التقديريه (٢٠) و(٢١) كطريقة بديلة للتصنيف وكانت النتائج ملخصة في جدول رقم (٧)

جدول (٧) تصنيف المشاهدات حسب الصيغة الاحتمالية لدالة التمييز الخطي

	التصنيف		
الحالة	اصبح المريض عائد الى المجموعه الاولى (٠)	اصبح المريض عائد الى المجموعه الثانيه (١)	نسبة التصنيف الصحيح
المريض عائد الى المجموعه الاولى (٠)	54	26	67.5%
المريض عائد الى المجموعه الثانيه(١)	19	61	76.3%
نسبه التصنيف الكلي	45.6%	54.4%	71.9%

اظهرت نتائج جدول (٧) ان احتمال التصنيف الصحيح لمريض يعود الى المجموعة الاولى (٠) كانت 67.5% ولمريض يعود للمجموعة الثانية (١) كانت 76.3% في حين خطأ التصنيف للمجموعة الاولى 32.5% وللمجموعة الثانية 23.7% اما نسبة التصنيف الكلية هي 71.9% ونسبة التصنيف الخاطىء الكلية 28.1% وفق الصيغة الاحتمالية لتصنيف المشاهدات .

(٤-٢) التحليل الاحصائي لنموذج الانحدار اللوجستي ثنائي الاستجابة

بأستخدام البرنامج الاحصائي (SPSS.V20) وبأستخدام طريقة Enter حصلنا على المعلومات الوصفية لعينة الدراسة الملخصة في جدول (٨)

جدول رقم (٨) المعلومات الوصفيه لعينه الدراسه

Unweighted cases	N	Percent
Selected cases Included in analysis	١٦٠	100.0
Missing cases	٠	0
Total	١٦٠	100.0
Unselected cases	٠	0
Total	١٦٠	100.0

يلخص الجدول (٨) البيانات المدخلة في التحليل وحجم العينة المدروسة والبيانات المفقودة (Missing data) . ويتضمن الجدول (٩) عدد الدورات التكرارية لمشتقات دالة الامكان للحصول على اقل قيمه لسالب ضعف لوغارتيم دالة الامكان (−2log likelihood) للحصول على التقدير الامثل لمعالم نموذج الانحدار اللوجستي ثنائي الاستجابة .

جدول (٩) الدورات التكرارية للحصول على التقدير الامثل لمعالم نموذج الانحدار اللوجستي ثنائي الاستجابة

Iteration	−2log likelihood	Coefficients								
		Constant	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
١	162.878	0.758	−2.268	−0.207	−0.871	−0.108	0.557	−0.359	−0.195	0.948
٢	156.984	0.901	−3.387	−0.236	−1.184	−1.37	0.669	−0.474	−0.194	1.246
٣	156.243	0.913	−3.998	−0.228	−1.258	−1.13	0.679	−0.497	−0.185	1.312
٤	156.200	0.912	−4.188	−0.227	−1.264	−1.416	0.679	−0.499	−0.183	1.316
٥	156.200	0.912	−4.206	−0.227	−1.264	−1,416	0.679	−0.499	−0.183	1.316
٦	156.200	0.912	−4.206	−0.227	−1.264	−1.416	0.679	−0.449	−0.183	1.316

وقد حصلنا في الدوره السادسة لمشتق سالب ضعف دالة الامكان على اقل قيمة لها وهي مساوية (156.200) اي (−2log L =156.200) وتوقفنا عند هذه الدوره لان التغير في قيم المعاملات $(\hat{\beta}_8,.....,\hat{\beta}_1)$ اصبح اقل من 0.001 وفي حقيقة الامر فان التغير في المعالم المقدرة اصبح بطيئاً بعد الدوره الرابعه وكما نلاحظ ذلك في الجدول (٩) لذلك يمكن القول ان مقدرات المعالم في الدورات (6,5,4) هي متشابهة مع فروقات بسيطة جداً وتوقفنا عند الدوره السادسة واعتبرنا معالمها افضل نتيجة ويلخص جدول رقم (١٠) معالم النموذج الامثل والخطأ المعياري لكل معلمه .

جدول (١٠) تقدير معالم نموذج الانحدار اللوجستي ثنائي الاستجابة

المتغيرات	B	S.E	Wald	df	Sig.	Exp(B)	95% C.I FOR exp(B)	95% C.I FOR exp(B)
							LOWER	UPPER
$X_1(1)$	−4.206	1.071	15.426	١	0.000	0.015	0.002	0.122
$X_2(1)$	−0.227	0.430	0.278	١	0.598	0.797	0.343	1.853
$X_3(1)$	−1.264	0.435	8.461	١	0.004	0.283	0.121	0.662
$X_4(1)$	−1.416	0.878	2.603	١	0.107	0.243	0.043	1.356
$X_5(1)$	0.679	0.852	0.635	١	0.425	1.973	0.371	10.489
$X_6(1)$	−0.499	0.459	1.179	١	0.278	0.607	0.247	1.494
$X_7(1)$	−0.183	0.407	0.202	١	0.653	0.833	0.375	1.850
$X_8(1)$	1.316	0.404	10.620	١	0.001	3.730	1.690	8.233
Constant	0.912	0.625	2.133	١	0.144	2.490		

ولأختبار جودة توفيق النموذج تم استخدام احصاءة اختبار نسبة الامكان (Likelihood ratio test) والتي تتبع توزيع مربع كاي (χ^2) وفق العلاقة الاتية .

$$\chi^2 = 2[\log_e L_o - \log_e L_1]$$

اذ:

L_1 : قيمة دالة الامكان الاعظم عند الفرضية H_1

L_0 : قيمة دالة الامكان الاعظم عند الفرضية H_0

حيث ان قيمة $\chi^2 = 65.607$ وهي معنويه عند مستوى دلاله اقل من 0.05 كما يلاحظ ذلك في الجدول رقم (١١) اذ Sig=0.000 مما يؤكد معنوية النموذج الموفق كما هو موضح في الجدول رقم(١١) عند درجة حرية d.f=8 والتي تمثل عدد المتغيرات المدخلة باسلوب النموذج الكامل (Enter) .

جدول (١١)

قيمة χ^2 للنموذج الموفق

	χ^2	d.f	Sig.
Model	65.607	٨	0.000

وهناك اختبار لامعلمي لجودة توفيق النموذج يعرف باختبار (Hosmer and lemeshow) حيث كما يلاحظ من الجدول (١٢) ان قيمة $\chi^2 = 11.210$ وان قيمة المعنوية $P> 0.05$ وهذا يؤكد ملائمة النموذج للبيانات .

جدول (١٢)

اختبار (Hosmer and lemeshow)

Step	χ^2	d.f	Sig.
١	11.210	٨	0.190

وبموجب الجدول رقم (١٠) نلاحظ ان عمود (B) يحتوي على القيم التقديريه لمعاملات نموذج الانحدار اللوجستي ثنائي الاستجابة للمتغيرات التوضيحية الثمانية بوحدات (Log-odds) وان معادلة النموذج يمكن كتابتها على النحو الاتي:

$$Log(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}) = 0.912 - 4.206X_{1(1)} - 0.227X_{2(1)} - 1.264X_{3(1)} - 1.416X_{4(1)} + 0.679X_{5(1)} - 0.499X_{6(1)} - 0.183X_{7(1)} + 1.316X_{8(1)}$$

اذ $\hat{\pi}$ هي احتمال الاصابة بمرض سرطان الدم اللمفاوي الحاد (ALL) والدالة المذكورة توضح العلاقة ما بين المتغيرات التوضيحية ومتغير الاستجابة (التابع) بوحدات (Logit) .

والعمود الرابع من الجدول رقم (١٠) يمثل احصاءة (Wald) التي تتوزع مربع كاي (χ^2) بدرجه حريه ١ والتي صيغتها .

$$wald = (\frac{\hat{B}_i}{S.E(\hat{B}_i)})^2$$

اذ يلاحظ بموجب احصاءة (Wald) ان متغيرات (الجنس (X_1)،الوزن (X_3)،عدد الصفائح (X_8)) ، كانت ذات تاثير معنوي على متغير الاستجابة (الاصابة بمرض سرطان الدم الحاد) اما بقية المتغيرات التوضيحية فلم يكن لها اي تأثير معنوي وكما هو موضح في العمود السادس من الجدول (١٠).

اما العمود السابع من الجدول المذكور exp(B) فيوضح قيمة الدالة الاسية لمعاملات الانحدار ، وهو يعبر عن المضاعف الذي تتغير فيه نسبة الترجيح (احتمال وقوع الحدث π الى احتمال عدم وقوعه $(1-\pi)$) وبالنسبة للمتغير الاول(X_1) مثلاً فأن السطر الاول من الجدول (١٠)

$$Exp(-4.206)=e^{-4.206}=0.015$$

وهذه القيمة تعني ان تغير الجنس من ذكر الى انثى سيقلل من احتمال الاصابة بمرض سرطان الدم اللمفاوي الحاد (ALL) بمقدار $(\hat{\beta}_1 = -4.206)$في لوغارتيم الافضلية لمتغير الاستجابة Y .وبموجب النموذج المقدر تم تقدير احتمال متغير الاستجابة (y) ياخذ القيمة (١) حسب الصيغة (32) وكذلك تقدير احتمال متغير الاستجابة (y) ياخذ القيمة (٠) وتم تصنيف بيانات مجموعتي المرضى المصابين بنوعي سرطان الدم النخاعي الحاد (AML) واللمفاوي الحاد (ALL) ،وكانت نتائج تصنيف البيانات بموجب الصيغة الاحتمالية لنموذج الانحدار اللوجستي ثنائي الاستجابة ملخصة في جدول (١٣).

جدول (١٣) تصنيف المشاهدات حسب الصيغة الاحتمالية لنموذج الانحدار اللوجستي ثنائي الاستجابة

الحالة	التصنيف		
	اصبح المريض عائد الى المجموعه الاولى (٠)	اصبح المريض عائد الى المجموعه الثانيه (١)	نسبة التصنيف الصحيح
المريض عائد الى المجموعه الاولى (٠)	57	23	71.3%
المريض عائد الى المجموعه الثانيه (١)	21	59	73.8%
نسبه التصنيف الكلي	48.75%	51.25%	72.5%

يلاحظ من الجدول (١٣) ان احتمال التصنيف الصحيح لمريض يعود الى المجموعة الاولى 71.3% ولمريض يعود للمجموعة الاولى 28.7% وللمجموعة الثانية 26.2% .وان نسبة التصنيف الصحيح الكلية 72.5% في حين نسبة التصنيف الخاطىء الكلية 27.5% .

٥- الاستنتاجات

١- اظهرت الدراسة ان تصنيف البيانات بطريقة الصيغة الاحتمالية لنموذج الانحدار اللوجستي ثنائي الاستجابة تفوقت على طريقتي دالة التمييز الخطي والصيغه الاحتمالية لها باعطائها اقل احتمال لخطأ التصنيف.وكذلك كانت الصيغة الاحتمالية لدالة التمييز الخطي افضل في تصنيف البيانات من دالة التمييز الخطي.

٢- امكانية استخدام تحليل الانحدار اللوجستي كنموذج تصنيف خطي للتمييز (للفصل) بين مجموعتين من المشاهدات وخاصة في حالة البيانات المصنفة ذات الاستجابة الثنائية وكذلك في حالة عدم تحقق الافتراضات الاساسية للدالة التمييزية الخطية .

٣- من خلال نتائج التحليل لطريقتي الدالة التمييزية الخطية ونموذج الانحدار اللوجستي ثنائي الاستجابة لوحظ ان كلا الطريقتين قد حددت نفس المتغيرات التوضيحية ذات الاكثر أهمية في تشخيص امراض سرطان الدم النخاعي الحاد (AML) واللمفاوي الحاد (ALL) .

٦- التوصيات

١- التوسع في استخدام الصيغة الاحتمالية للدالة التمييزية الخطية كطريقة لتصنيف البيانات في حالة لدينا ثلاث مجموعات فأكثر.

٢-التوسع في استخدام تحليل التمييز الخطي والانحدار اللوجستي في حالة البيانات المصنفة لاكثر من مستويين والتي يكون فيها المتغير التابع متعدد الاستجابة(Multirespose) .

المصادر

اولاً: المصادر العربية

١. الباجلان ،عباس كول مرادبك مراد(٢٠٠٩)"استخدام نموذجي (Cox) و(Logistic) في تحليل البقاء مع تطبيق عملي"رساله ماجستير في الاحصاء ،كليه الادارة والاقتصاد ،الجامعه المستنصريه.

٢. البياتي .هبه ابراهيم صالح(٢٠٠٥) "تحليل المسار في أنموذج الانحدار اللوجستي مع تطبيق عملي "رساله ماجستير في الاحصاء،كليه الادارة والاقتصاد ،الجامعه المستنصريه.

٣. الياسين .دريد حسين بدر(٢٠٠٩) " استخدام بعض طرائق التمييز الحصينه لتشخيص امراض سرطان الدم(اللوكيميا)"رساله ماجستير احصاء.كليه الادارة والاقتصاد ،الجامعه المستنصريه

٤. حميد ،رند سليم (١٩٩١)" استخدام الداله المميزه في تشخيص بعض الاورام السرطانيه "رساله ماجستير في الاحصاء مقدمه الى كليه الادارة والاقتصاد ،جامعه بغداد .

ثانياً :- المصادر الأجنبية :-

5. **Abbas, F. M. Azhar, M. E (2008)" Comparing Discriminant Analysis and Logistic Regression Model"as a Statistical Assessment Tools of Arsenic and Heavy Metal Contents in Cockles"School of Industrial Technology, Environmental Technology Division Universiti Sains Malaysia, 11800 Penang, Malaysia.**
6. **AL- AFIFI ,R .M*(2010)"The Use of Multinomial Logistic Regression Model on Physical Violence Data"* degree of Master of Applied Statistics, Al- Azhar University – Gaza.**
7. **Al-Thabhawee ,G .D (2012) "A Comparison between Discriminant Analysis and Logistic Regression on the Classification of Cancer Patients" for the Degree of Master of Science in Mathematics (Mathematical Statistic) University of Kufa.**
8. **Alvin C.Rencher(2002)"Methods of Multivariate Analysis" John Widy and Sons,puplication.**
9. **Anderson,T.W.(1918)"An introduction to multivariate statistical analysis"by John Wiley &Sons.Inc**
10. **Dillon,W.R and Goldistein,M(1984)"Multivariate Analysis Methods and Application".John Wily &Sons,Inc,New York,USA.**
11. **Feighner, J. P., & Sverdlov, L. (2002)" The use of discriminant analysis to separate a study population by treatment subgroups in a clinical trial with a new pentapeptide antidepressant". *Journal of Applied Research, 2*;17 – 18.**
12. **Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression (2nd Edition)*. New York: Wiley**
13. **Kemp, G.C.R.(2000). " Semi – Parametric Estimation of a Logit Model ", University of Essex. http://www.econometricsociety.org/meetings/wcoo/pdf 0879.pdf**
14. **Krieng kitbumrungrat (2012)"Comparison logistic regreesion and Discriminant analysis in classification groups for breast cancer"Faculty of Information Technology, Rangsit University, Thailand**
15. **Lei, P., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: a comparison of classification errors in the two-group case. *Journal of Experimental Education, 72*; 25 – 49.**
16. **Pohar,M&Blas,M&Turk,S.(2004)"Comparison of logistic regression and linear Discriminant analysis"Asimulation study Metodolski zvezki,Vol 1,No.1(143-161).**

Comparison between some of linear classification models with practical application

Abstract

Linear discriminant analysis and logistic regression are the most widely used in multivariate statistical methods for analysis of data with categorical outcome variables .Both of them are appropriate for the development of linear classification models .linear discriminant analysis has been that the data of explanatory variables must be distributed multivariate normal distribution. While logistic regression no assumptions on the distribution of the explanatory data. Hence ,It is assumed that logistic regression is the more flexible and more robust method in case of violations of these assumptions.

In this paper we have been focus for the comparison between three forms for classification data belongs two groups when the response variable with tow categorise only.

The first form is the linear discriminant function ,The second is the probability form which it is derivative as alternative for the linear discriminant function while the third form is the probability function model. Of the logistic regression the comparison between these methods is based on measure of the probability of misclassification .We show that the results of the probability form of the logistic regression has minimum probability of misclassification through the application on the data of two types of (leukemia).

Key words/ Linear discriminant analysis ,binary response logistic regression and misclassification probability.