AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

دراسة احصائية لبعض المقدرات اللامعلمية لأنموذج الانحدار اللوجستي

لقاء علي محمد	نرجس باسم خلف
lekaa.a@coadec.uobaghdad.edu.iq	Narjes.bassim@alayen.edu.iq
قسم الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد، بغداد، العراق	قسم المختبرات الطبية، كلية التقنيات الصحية والطبية، جامعة العين العراقية، ذي قار، العراق

المستخلص

تقدم الدراسة تقديرا لنموذج الانحدار اللوجستي الثنائي اللامعلمي مع احتمالية الإصابة بمرض لسرطان الدم الليمفاوي المزمن اعتمادا على العوامل المؤثرة على الإصابة (جنس المريض، عمر المريض، خلايا الدم البيضاء، الهيموغلوبين، الخلايا الدموية ، الصفائح الدموية) ويعد سرطان الدم الليمفاوي من اخطر الامراض على حياة الانسان، فقد ينتج عنه العديد من المضاعفات التي قد تؤدي الى التعرض لأنواع أخرى من الدم . لا يمكن السيطرة عليها . وتهدف الدراسة الى التعرف على اهم الأساليب الإحصائية في تحليل بيانات مرضى السرطان والتعرف على اهم المؤثرات عليها وهي المتغيرات الفئوية، وتكمن أهميتها في كيفية تقدير التنبؤات بعد الحصول على افضل نموذج لوجستي باستعمال الطريقة المختلطة للمقدر اللبي والشريحة الممهد لتقدير معلمة التمهيد المثلى باستعمال العبور الشرعي والعبور الشرعي المعمم، تم التوصيل الى ان الطريقة المختلطة للمقدر اللبي والشريحة الممهدة (KAS) هي الأفضل من خلال معيار المقارنة متوسط مربع الخطأ.

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث: 16/2/2024
تاريخ قبول البحث: 12/4/2024
تاريخ رفع البحث على الموقع: 31/12/2024

الكلمات المفتاحية:

معلمة التمهيد ،طريقة العبور الشرعي CV، طريقة العبور الشرعي المعمم GCV، الطريقة المختلطة للمقدر اللبي والشريحة الممهدة (KAS خوارزمية Logt الموضعية LL، خوارزمية التهديف الموضعي (LSA حيث يتم في هذا البحث استعمال الطريقة المختلطة للمقدر اللبي والشريحة الممهدة

للمراسلة:

نرجس باسم خلف

Narjes.bassim@alayen.edu.iq

DOI: https://doi.org/10.55562/jrucs.v56i1.14

1. المقدمة (Introduction)

ان انموذج الانحدار اللوجستي يعد من النماذج الاحصائية ذات الاهمية الكبيرة التي تستعمل في تحليل البيانات الثنائية والتي تأخذ القيم الاحتمالية [1,0] والتي تتبع توزيع برنولي ،حيث الانحدار اللوجستي يستعمل لوصف العلاقة بين متغير الاستجابة (Y) والمتغيرات التوضيحية (x_i) وكذلك للتنبؤ و للتقدير

2. نموذج الانحدار اللوجستي الثنائي اللامعلمي:

يتم تحديد انموذج الانحدار اللوجستي الثنائي بواسطة متغير الاستجابة Y الذي يأخذ قيمة (0.1) والتي (تفسر إما بنعم أو لا، مريض وغير مريض، مصاب وغير مصاب، وهكذا) .تم استخدام نموذج الانحدار اللوجستي الثنائي لمعرفة العوامل التي تؤثر على حدوث سرطان الدم الليمفاوي، متغير الاستجابة هو مؤشر على ما إذا كان المريض سيصاب بالسرطان أم لا، في حين أن المتغيرات التفسيرية (العوامل المؤثرة) هي جنس المريض وعمر المريض وخلايا الدم البيضاء، دم الهيموجلوبين، الخلايا الدموية والصفائح الدموية، وبالتالي فهي دالة كثافة الاحتمال هي كما يلي:

$$y_i \sim Ber(\pi_{x_i}) \qquad i = 1,2,\dots,n \tag{1}$$

$$\mathrm{p}(y_i/x_i) = [\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i} \quad ; y_i = 0,1 \tag{2}$$

$$p\ (y = 1) = \pi(x_i) \quad (3)$$

$$p\ (y = 0) = 1 - \pi(x_i) \quad (4)$$

حيث ان :

y_i متغير تابع ثنائي الاستجابة (0,1)

$\pi(x_i)$: احتمال حدوث استجابة النجاح (Success) عندما $y_i = 1$

$1 - \pi(x_i)$: احتمال عدم حدوث استجابة الفشل (Failure) عندما $y_i = 0$

وأن توقع متغير الاستجابة يمثل احتمال النجاح

$$E(y_i) = \pi(x_i) \quad (5)$$

وتباين متغير الاستجابة يمثل احتمال حاصل ضرب النجاح مع احتمال الفشل

$$v(y_i) = \pi(x_i)\left(1 - \pi(x_i)\right) \quad (6)$$

اذا كان لدينا $[y_1, y_2, y_3, \dots\dots, y_n]$ تمثل عينة عشوائية من المتغير ثنائي الاستجابة

$y_i \in [0,1]$، ولتكن $x_{i1}, x_{i2}, x_{i3}, \dots\dots\dots, x_{ip}$ تمثل مجموعة المتغيرات التوضيحية

وان n تمثل عدد المشاهدات لهذه المتغيرات التي تكون المصفوفة الآتية:

$$X = (x_{ij})_{n+k}$$

X : تمثل مصفوفة المتغيرات التوضيحية بدرجة $n + 1 \times p$

n : حجم العينة، $i = 1,2, \dots, n$

k :تمثل عدد المتغيرات التوضيحية $(j = 1,2, \dots, k)$

p: عدد المعلمات والذي $k + 1$

اذ ان احتمال الاستجابة لأنموذج الانحدار اللوجستي الثنائي يأخذ القيمة (1) وحسب الصيغة الآتية:

$$p(y = 1|\ x) = \frac{1}{1 + (e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}})^{-1}} \quad (7)$$

وأن احتمال الاستجابة عند القيمة (0) يكون كالاتي:

$$p(y = 0|\ x) = \frac{1}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}} \quad (8)$$

وبما أن مجموع الاحتمالات يساوي(1) وحسب الصيغة الآتية:

$$p(y = 1|\ x) + p(y = 0|\ x)\ = 1 \quad (9)$$

فأنه يمكن التعبير عن أنموذج الانحدار اللوجستي بالصيغة الاتية:

$$y_i = \pi(x_i) + \varepsilon_i \quad , i = 1,2, \dots . n \quad (10)$$

اذ ان

$\pi(x_i)$ تمثل دالة الانحدار اللوجستي أو تمثل دالة الاستجابة اللوجستية

εi يمثل الخطأ العشوائي حيث يكون له متوسط يساوي صفر وكالآتي:

$$\varepsilon_i = y_i - \pi(x_i)$$
$$E(\varepsilon_i) = E(y_i) - E(\pi(x_i)) = \pi(x_i) - \pi(x_i) = 0$$

اما تباين حد الخطأ العشوائي فأنه يكون مساوياً الى تباين المتغير y_i ثنائي الاستجابة وكالآتي:

$$v(\varepsilon_i) = v(\ y_i) = \pi(x_i)(1 - \pi(x_i)) \quad (11)$$

حيث تكتب دالة الانحدار اللوجستي (احتمال الاستجابة) حسب الصيغة الآتية:

$$\pi(x_i) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}} \quad (12)$$

وأن دالة الانحدار اللوجستي التقديرية تكون كالآتي:

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}}}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}}} \quad (13)$$

نلاحظ من المعادلة (12) ان شكل العلاقة بين المتغيرات التوضيحية (x_{ij}) واحتمال الاستجابة $\pi(x_i)$ لا يمكن ان يكون خطياً وانما تأخذ شكلاً منحنياً أي على شكل حرف (S) .

$$1 - \pi(x_i) = 1 - \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}} \quad (14)$$

$$\frac{\pi(x_i)}{1 - \pi(x_i)} = \frac{\pi(x_i)(y=1)}{1 - \pi(x_i)(y=1)} = e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}} \quad (15)$$

حيث أن

$\beta_0, \beta_1, \beta_2, \dots \dots, \beta_j$: تمثل معلمات مجهولة يراد تقديرها.

x_{ij} : متغيرات التوضيحية.

3. الطريقة المختلطة للمقدر اللبي والشريحة الممهدة (KAS)

(Mixed method kernel estimator and spline smoothing)

يمكن تعريف البيانات لمتغير الاستجابة (y) والمتغيرات التوضيحية (يمكن تعريف البيانات لمتغير الاستجابة (y) والمتغيرات التوضيحية (x_i) حيث في هذه الطريقة تكون بعض المتغيرات التوضيحية تتغير بفترات معينة ومحددة اما المتغيرات التوضيحية الأخرى فهي متغيرات عشوائيا وبذلك تصبح دالة الانحدار اللوجستي الثنائي، كما في صيغة (2-27) ، وهو عبارة عن دالة تجميعية ويمكن استعمال العديد من النماذج اللامعلمية لتقدير دالة الانحدار اللوجستي الثنائي، بما فيه Spline Smoothing و Kernel ، حيث يتم تطبيق مقدر Spline بشكل مناسب إذا كان نمط البيانات يميل إلى التغيير في فترات فرعية معينة، فضلاً عن ذلك يمكن أن ينتج مقدر الشريحة مرونة و منحنيات ممهدة، ويتم تحديد ممهد منحنى الشريحة من خلال معامل التمهيد، ويُفترض أن تكون دالة الانحدار اللوجستي الثنائي(Spline Smoothing) ممهدة ومضمنة في فضاء (Sobolev) وكتابتها كالاتي:

$$\beta \epsilon W_2^m[a,b] \; ; W_2^m[a,b] = \left\{ \beta : \int_{a_j}^{b_j} \left[\beta^{(m)}(x)\right]^2 dx < \infty \right\} \quad (16)$$

حيث m هي شريحة متعددة الحدود.

يمكن الحصول على ممهد الشريحة (Smoothing Spline) باستعمال تقليل مربعات صغرى جزائية (Penalized Least Squares(PLS)) وصيغة (PLS) هي معيار تقدير يجمع بين دالة حسن المطابقة والجزاء.

$$Min_{\beta \in W_2^m(a,b)} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta(x_i)^2 + \alpha \int_a^b \left(\beta^{(m)}(x_i)\right)^2 dx \right\} \quad (17)$$

الجزء الأول في معادلة (17) هو الدالة التي تقيس ملائمة البيانات حيث تمثل المربعات الصغرى العادية، والجزء الثاني هو مركبة حد الجزاء لدالة ($\beta(x_i)$)، حد الجزاء هو تحديد الجزء غير الممهد (خشونة المنحنى) مع (α) وهي معلمة التمهيد، وتمثل معلمة التمهيد دورًا حيويًا في تحديد تمهيد الدالة، وبناء عليه يعرف هذا المعيار باسم (PLS).

على عكس مقدر Smoothing Spline، يستعمل مقدر Kernel بشكل شائع لنمذجة البيانات بنمط عشوائي، حيث يعتمد نهج Kernel على قيمة معلمة التمهيد، والتي تعمل على التحكم في تمهيد الدالة المقدرة، وان اختيار معلمة التمهيد المثلى أمر بالغ الأهمية لمقدر Kernel ،إذا كانت دالة الانحدار اللوجستي الثنائي $\beta(x_i)$ غير معروفة يقترب من مقدر Kernel ، فيمكن تقريب دالة $\beta(x_i)$ بواسطة سلسلة Taylor مع رتبة m على النحو الآتي:

$$\beta(x) \approx \beta(x_0) + \frac{(x-x_0)^1 \beta^1(x_0)}{1!} + \frac{(x-x_0)^2 \beta^2(x_0)}{2!} + \cdots + \frac{(x-x_0)^m \beta^m(x_0)}{m!} \quad (18)$$

اذا كانت

$$\tau_v(x_0) = \frac{\beta^v(x_0)}{v!} \quad ; v = 1,2,\dots,m \tag{19}$$

ويمكن كتابة صيغة (18) بالشكل الاتي:

$$\beta(x) \approx \tau_0(x_0) + (x - x_0)^1\tau_1(x_0) + (x - x_0)^2\tau_2(x_0) + \cdots + (x - x_0)^m\tau_m(x_0) \tag{20}$$

يتم الحصول على مقدر *Kernel* عندما يكون ترتيب متعدد الحدود $m = 0$،وبذلك تصبح صيغة (81-2) كالاتي:

$$\beta(x) \approx \tau_0(x_0) \tag{21}$$

ويمكن الحصول على مقدر *Kernel* من خلال طريقة تقدير *WLS* كالاتي:

$$Min_\lambda\{(y - \beta)^T K_\lambda(x_0)(y - \beta)\} \tag{22}$$

حيث $K_\lambda(x_0)$ هو مصفوفة الاوزان لدالة الالب (kernel)، و(λ) هي معلمة التمهيد Kernel

الان سوف نتبع طريقة التقدير للمرحلتين التي اقترحها (33,Rahmawati,et al) وهو خليط من Spline Smoothing و Kernel المقدر للانحدار اللوجستي اللامعلمي ثنائي الاستجابة، ويتم الحصول على تقدير الخليط من خلال مرحلتين من التقدير، المرحلة الأولى باستخدام طريقة تقدير PWLS والمرحلة الثانية باستخدام طريقة تقدير WLS، حيث يتم وصف طريقة التقدير ذات المرحلتين للحصول على مقدر الخليط المقترح على النحو الاتي.

اذا أعطيت البيانات بشكل ازواج$(y_i, x_{1i,}x_{2i,}\dots\dots x_{pi}, t_{1i,}, t_{2i,}\dots,, t_{qi})$، وعليهِ يكون انموذج الانحدار اللوجستي اللامعلمي ثنائي الاستجابة كالاتي:

$$y_i = \omega_i(x_{ji}, t_{ki}) + \varepsilon_i \quad , i = 1,2,\dots,n;\ j = 1,2,\dots,p;\ \ k = 1,2,\dots,q \tag{23}$$

حيث ان $\omega_i(x_{ji}, t_{ki})$: هو منحنى الانحدار اللوجستي الثنائي يُفترض فيه أن شكل الدالة غير معروف، مع وجود بعض المتغيرات التوضيحية التي لها أنماط تتغير في فترات جزئية معينة والمتغيرات المتبقية لها نمط عشوائي، منحنى الانحدار اللوجستي الثنائي هذا هو دالة مضافة، لذلك يمكن كتابته كلاتي:

$$\omega_i = \sum_{j=1}^{p} g_j(x_{ji}) + \sum_{k=1}^{q} f_k(t_{ki}) \ , \quad i = 1,2,\dots\dots,n \tag{24}$$

حيث ان $\sum_{j=1}^{p} g_j(x_{ji})$: يمثل منحى الانحدار اللوجستي الثنائي مع p من المتغيرات التوضيحية و:$\sum_{k=1}^{q} f_k(t_{ki})$ تمثل منحى الانحدار اللوجستي الثنائي مع q من المتغيرات التوضيحية.

يمكن كتابة صيغة (24-2) على شكل متجه كلاتي:

$$\omega = g + f \tag{25}$$

ويمكن كتابة الصيغة بشكل متجه كالاتي:

$$y = g + f + \varepsilon \tag{26}$$

منحنى الانحدار اللوجستي الثنائي g يقترب من خلال دالة Smoothing Spline ، ومن المفترض أن يكون منحنى الانحدار اللوجستي الثنائي ممهد ومحتوى في فضاء (*Sobolev*) $g \in W_2^m(a,b)$، وفي الوقت ذاته يتم تقريب منحنى الانحدار f بواسطة دالة Kernel ، يمكن الحصول على خليط ثنائي الاستجابة من Smoothing Spline و Kernel المقدر من خلال طريقة تقدير على مرحلتين. يتم تقدير *(g)* Spline بواسطة PWLS في المرحلة الأولى من التقدير، ويتم تقدير Kernel *(f)* بطريقة WLS في المرحلة الثانية من إجراء التقدير، قبل تقدير جزء Smoothing Spline ، نحتاج إلى تعديل صيغة (26) في الشكل الاتي:

$$h = g + \varepsilon \tag{27}$$

وان :

$$h = y - f$$

يمكن الحصول على تقدير جزء Smoothing Spline لـ $(\hat{g})$ باستعمال صيغة مربعات الصغرى الموزونة الجزائية PWLS لـصيغة (27) كلاتي:

$$Min_{g \in W_2^m(a,b)} \left\{ \frac{1}{n}(h - g)^2\sigma^2 + \alpha \int_a^b \left(g^{(m)}(x)\right)^2 dx \right\} \tag{28}$$

$$0 < \alpha < \infty$$

حيث (α) هي معلمة التمهيد لمقدر Smoothing Spline لاستجابة الثنائية، ويتم استبدال نتائج التقدير من المرحلة الأولى في (25)، في المرحلة الثانية من التقدير، ويمكن الحصول على تقدير جزء Kernel لـ $(\hat{f})$ باستعمال WLS على النحو الاتي:

$$Min_{\alpha}\{(y-\hat{g}-f)^{T}\Sigma^{2}K_{\lambda}(t_{0})(y-\hat{g}-f)\} \quad (29)$$

من خلال استبدال النتائج في التقدير على مرحلتين لصيغة (25)، يمكن الحصول على خليط من Spline Smoothing و Kernel في الانحدار اللوجستي اللامعلمي ثنائي الاستجابة.

4. ممهد ناداريا – واتسون(NW) (Nadaraya- Watson smoothing)

ان هذا الممهد تم اقتراحه من قبل الباحثان (Nadaraya and Watson) عام (1964) م والذي عن طريقه يتم الحصول على دالة الانحدار اللامعلمي التقديرية ($\widehat{m}(x)$)ويتم ذلك من خلال الاعتماد على متسلسلة الأوزان، وان ممهد (Nadaraya-Watson) يُعد من المقدرات الاكثر استعمالاً في تقدير نماذج الانحدار اللامعلمي.

من اجل تقدير طريقة المقدر اللبي (Nadaraya-Watson) في أنموذج الانحدار اللوجستي ثنائي اللامعلمي، على النحو الاتي :

$$\text{Logit}(\pi_{x_i}) = m(x) \quad (30)$$

حيث ان :

$m(x)$ **:** تمثل دالة الانحدار اللامعلمية بطريقة المقدر اللبي (kernel)

وباستعمال مقدر (Nadaraya-Watson) وهو احد مقدرات (kernel) الاكثر شيوعاً ، ويسمى ايضا (Local Constant) ثابت موضعي وذلك بسبب استعمال معلمة تمهيد واحدة فقط لتقدير دالة الانحدار .

وعليه فأن صيغة المقدر اللبي (Nadaraya-Watson) لدالة الانحدار ($\widehat{m}(x)$) تكون على النحو الاتي:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{(X-x_i)}{\lambda}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{(X-x_i)}{\lambda}\right)}$$

$$\widehat{m}(x) = \sum_{i=1}^{n} W_{\lambda i}(x) Y_i \quad (31)$$

ويمكن إعادة صياغة المعادلة (31) بصيغة المصفوفات وكالاتي:

$$\widehat{m}(x) = WY \quad (32)$$

اذ ان:

$$W = (\acute{w}_1, \acute{w}_2, \dots, \acute{w}_n) \quad , Y = (Y_1, Y_2, \dots, Y_n)$$

5. المحاكاة (simulation)

يعد مفهوم التحليل باستخدام المحاكاة (simulation) امتداداً طبيعياً ومنطقياً للنماذج التحليلية والرياضية حيث يوجد العديد من الحالات التي لا يمكن تمثيلها رياضياً وقد يكون السبب الطبيعة العشوائية للمشكلة المدروسة او بسبب تعقيدها او نظراً لعدم وصف المشكلة , فان الاسلوب الذي غالبا ما يستخدم عند فشل جميع الأساليب في الحصول على البيانات هو اسلوب المحاكاة فالتقدم الذي حصل مؤخراً وتوافر البرمجيات والتطورات التقنية في مجال البرمجة والحاسوب جعلت من أسلوب المحاكاة ان تحقق حل افتراض حجم العينة المطلوبة n=(50,100,150).

1. افتراض قيم الانحراف المعياري σ=(0.1,1,3)
2. افتراض الدالة اللبية وهي (Gaussian)
3. تحديد تكرار التجربة المحاكاة (500) مرة

يتم في هذه المرحلة توليد المتغيرات كالاتي.

1. توليد المتغيرات التوضيحية : ان توليد المتغيرات التوضيحية p=(4,6,8) حيث تتبع التوزيع الطبيعي باستعمال الاوامر و packges الخاصة في برنامج R باستعمال طريقة (Monte Carlo).

$$x_i \sim N(0, \sigma^2) \qquad i = 1,2, \dots, n \quad (33)$$

2. توليد الاخطاء العشوائية : يتم توليد الاخطاء العشوائية بنفس اسلوب توليد المتغيرات التوضيحية وذلك باستعمال الاوامر في برنامج R ، حيث يتم الحصول على الاخطاء عشوائية تتوزع توزيع طبيعي بمتوسط صفر وتباين وحصلنا على متجه للأخطاء العشوائية ذي بعد (n*1)
3. تم ربط المتغيرات التوضيحية في الفقرة أعلاه وحسب عدد المتغيرات التوضيحية تم كتابة الدوال التمهيد كالاتي :

1. $m_{(x)} = x + 2\exp(-16x^2)$
2. $m_{(x)} = \sin(\pi x)$
3. $m_{(x)} = 9\sin(\pi x^2) + x^2$
4. $m_{(x)} = \frac{1}{(2x-0.07)^2+0.08} + \frac{2}{(0.25x-0.25)^2+0.5} - 3$
5. $m_{(x)} = 0.5\sin(2\pi x)$
6. $m_{(x)} = \frac{\sin(2.5\pi x)}{1+3x^2}$

4. يتم حساب القيمة الاحتمالية $\pi(x_i)$ وفق الصيغة الاتية:

$$\pi(x_i) = 1/(1 + \exp(-m_{(x)})) \quad (34)$$

حيث ان:

$(m_{(x)})$ تساوي الانموذج (34) أعلاه استنادا الى عدد المتغيرات الداخلة في الانموذج

5. توليد قيم المتغير المعتمد y وفق توزيع ثنائي الاستجابة بحجم عينة مساوي الى n=(100,150,50)

6. نتائج تجربة المحاكاة The results of The simulation experiment

اجريت تجارب المحاكاة باستعمال ثلاثة احجام للعينات n=(50,100,150) وبتكرار (500) . اذ سيتم تحليل وتفسير النتائج من خلال مقارنة الطرائق بمعيار (MASE) حسب كل طريقة من الطرائق الموضحة في الجانب النظري وعند كل حجم عينة وقيمة تباين ومعلمة تمهيد لعرض الاطلاع على اداء الطرائق باختلاف المواقف المتوقع مواجهتها في الجانب التطبيقي وكما يلي:

جدول (1): يبين قيم معلمة التمهيد λ لكل قيمة من قيم المتغيرات التوضيحية

Var / λ	X1	X2	X3	X4	X5	X6
CV	0.04677712	5.491377	0.4377533	0.4130487	0.4603479	0.3617489
GCV	0.24790830	3.649883	0.4245050	0.4386346	0.1062034	0.3601639

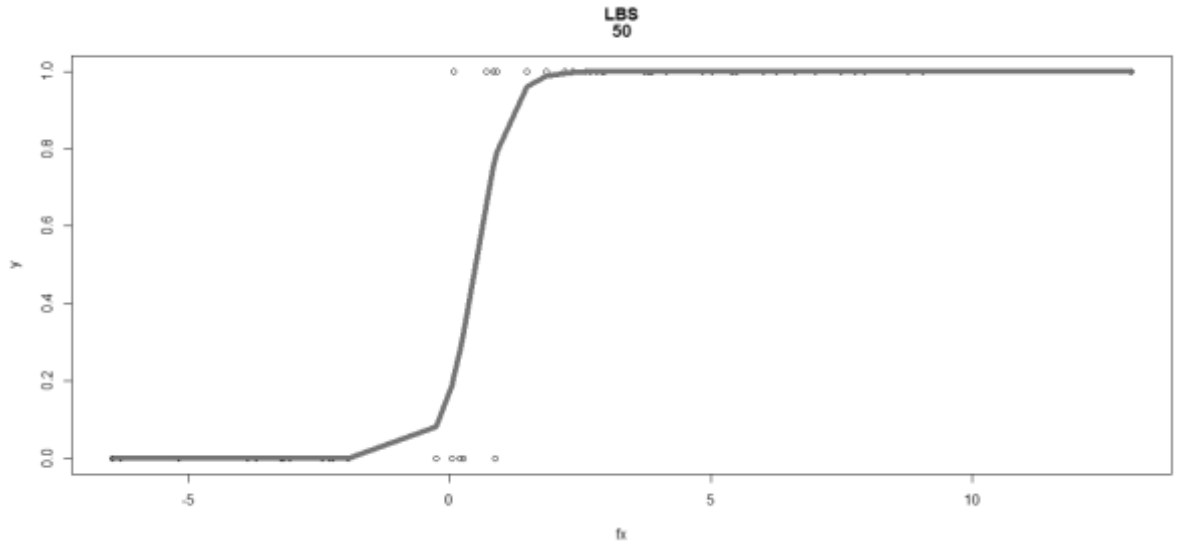
جدول (2): يبين معيارMASE **لمقارنة التقديرات لدالة الانحدار اللوجستي لحجوم عينات** n=150، n=100، n=50, **و** $\sigma = 0.1, \sigma = 1, \sigma = 3$ **ومعلمة تمهيد** CV

n=50		
σ	**NW**	**KAS**
0.1	0.4373236	0.4597410
1	0.3082588	0.3172990
3	0.4986625	0.4546600
n=100		
σ	**NW**	**KAS**
0.1	0.3924750	0.3856103
1	0.3903860	0.3921511
3	0.3696678	0.3547555
n=150		
σ	**NW**	**KAS**
0.1	0.4131742	0.3928202
1	0.3242449	0.3343259
3	0.4343203	0.4371999

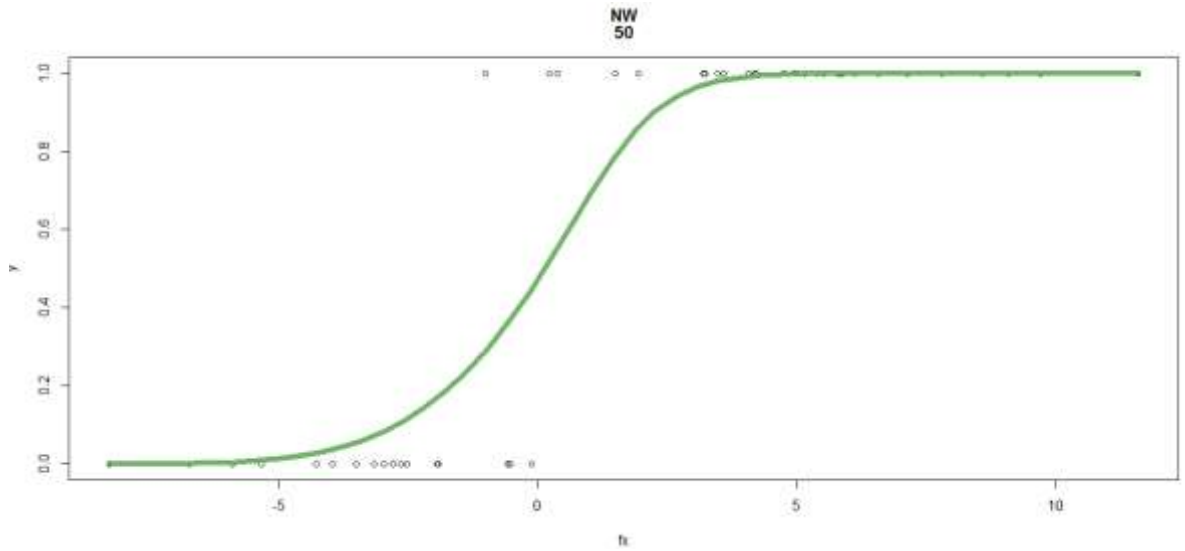
من الجدول رقم (2) يتضح ما يلي :

1. عند حجم العينة **n=50,100,150** افضل طريقة هي **NW** عند **σ=1** في حين كانت افضل الطرائق عند **σ=0.1** هي طريقة **KAS**

2. نلاحظ انخفاض قيمة (**MASE**) عند زيادة حجم العينة اذ كانت افضل اداء للطرائق عند حجم عينة **n=100** وباختلاف قيم σ مع ارتفاع بسيط في قيمته عند حجم العينة **n=150** الا انه يقل عند حجم عينة **n=50** .



شكل (1): القيم التقديرية للمتغير الاستجابة Y لا فضل الطرائق باستعمال قيمة معلمة التمهيد (CV) وحجم عينة 50 وقيمة تباين (0.1)



شكل (2): القيم التقديرية للمتغير الاستجابة Y لا فضل الطرائق باستعمال قيمة معلمة التمهيد (GCV) وحجم عينة 50 وقيمة تباين (0.1)

العوامل المؤثرة في المرض

ذكرت اغلب الدراسات الطبية التي تناولت مرض سرطان الدم الليمفاوي المزمن بأن هناك عدة عوامل تأثر على الإصابة بالمرض فضلا عن المتغيرات التي تم اقتراحها وهي كالاتي :

1. Y (متغير الاستجابة) : وهو متغير معتمد يأخذ احتمالين فقط هما (0) احتمال الشخص غير مصاب بالمرض، (1) احتمال الشخص المصاب بالمرض .
2. اما المتغيرات التوضيحية فهي :

X_1 : يمثل نوع الجنس (Sex) حيث يرمز (1) للذكور و (2) للإناث.
X_2 : يمثل العمر (age) .
X_3 : يمثل خلايا الدم البيضاء WBC(White Blood Cells).
X_4 : يمثل نسبة هيموجلوبين الدم HGB (Hemoglobin Blood)
X_5 : يمثل نسبة الهيماتوكريت HCT (Haematocrit)
X_6 : يمثل نسبة الصفائح الدموية PLT (Blood Platelets) .

تصنيف البيانات

إن أسلوب جمع البيانات تم عن طريق الاعتماد على طبلات المرضى (المصابين وغير المصابين) والتي تضمنت دراسة حول مرض (سرطان الدم الليمفاوي المزمن) في مستشفى الديوانية التعليمي في محافظة القادسية (وحدة استشارية الأورام) وقد شملت الدراسة 100 مشاهدة من الذكور والإناث (المصابين وغير المصابين) وقد قسمت المشاهدات إلى مجموعتين كالاتي:

1. المجموعة الأولى : قد شملت الأشخاص المصابين بالمرض بحجم (53)عينة ورمزنا لهم بالرمز(1).
2. المجموعة الثانية: قد شملت الأشخاص غير المصابين بالمرض بحجم(47)عينة ورمزنا لهم بالرمز(0).

بناء الانموذج

من اهم مراحل بناء الانموذج هو تحديد متغير الاستجابة (Response Variable) والمتغيرات التوضيحية (Variable Exploratory) بناءً على طبيعة ونوع البيانات قيد الدراسة والتي سيتم توضيحها كما في ادناه :

1. متغير الاستجابة Y (يمثل معدلات الاصابة بمرض سرطان الدم اللمفاوي من عدمها) : وكون المتغير يمثل الاصابة من عدمها لذلك سيكون Y=1 ويمثل المصابين بالمرض وY=0 يمثل غير المصابين بالمرض.
2. المتغيرات التوضيحي x_i تم توضيحها انفاً .

تحليل النتائج (Results Analysis) :

قبل البدء بعملية التقدير تم اجراء اختبار Variance inflation factor (VIF) لاختبار وجود مشكلة ارتباط خطي بين المتغيرات التوضيحية اذ تبين انه لا توجد علاقة ارتباط خطية بين المتغيرات التوضيحية اذ ان جميع متغيرات التوضيحية كانت قيمة VIF اقل من 5 وتكتب صيغة عامل تضخم التباين كالاتي :

$$VIF = \frac{1}{1-R_j^2}$$

حيث ان :

R_j^2 : معامل التحديد، $j = 1,2,\dots,k$

وكما مبين في الجدول ادناه :

جدول (3): يبين عامل التضخم التباين للبيانات الحقيقية التي تمثل الاصابة بسرطان الدم اللمفاوي.

المتغير	SEX	Age	WBC	HGB	HCT	PLT
VIF	1.164	1.484	1.128	4.57	4.11	1.049

جدول (4): يبين قيم متوسطات مربعات الخطأ MSE **لمعدل الاصابة بسرطان الدم اللمفاوي.**

Estimation Method	NW	KAS
MSE	0.2962321	0.23750245

تم احتساب قيم متوسط مربعات الخطأ (MSE) حيث بين النتائج ان طريقة (KAS) هي افضل الطرائق

7. الخاتمة

من خلال النتائج التي تم الحصول عليها من المحاكاة استنتج ان الطريقة المختلطة للمقدر الأساسي والشريحة الممهدة (LBS) هي الطريقة المفضلة عند أحجام العينات الثلاث وقيم التباين المختلفة وكانت تعطي اقل خطأ عند التباين 0.1وطرائق معلمة التمهيد العبور الشرعي CV ، العبور الشرعي المعمم) GCV باستعمال معيار المقارنة. MSE

المصادر

[1] الزركاني، اياد حبيب شمال، (2015) ، "تقدير معلمات انموذج الانحدار اللوجستي المتعدد في حالة وجود مشكلة الفصل والتعدد الخطي مع تطبيق عملي في المجال الطبي"، رسالة ماجستير، قسم الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

[2] العزاوي، احمد ذياب احمد، (2005) ، "المقارنة بين بعض طرائق تقدير انموذج انحدار اللوجستك والطرائق الحصينة للتجارب الحياتية ذات الاستجابة الثنائية باستخدام اسلوب المحاكاة" رسالة ماجستير، كلية الادارة والاقتصاد، جامعة بغداد.

[3] جاسم، مصطفى نصيف، (2018)، "مقارنة مقدرات الامكان الاعظم والدالة اللبية باعتماد اسلوب البوتستراب لأنموذج الانحدار اللوجستي"، رسالة ماجستير، قسم الاحصاء التطبيقي، كلية الادارة والاقتصاد، جامعة المستنصرية.

[4] حسين، شيرين علي، (2009)، "مقدرات الامكان الاعظم الموزونة الحصينة ومقارنتها مع طرائق اخرى لأنموذج اللوجستك مع تطبيق عملي"، رسالة ماجستير، كلية الادارة والاقتصاد، جامعة بغداد.

[5] حمزة، حنان عباس، (2013)، "تحليل أثر بعض المتغيرات المؤثرة على مرض سرطان الدم الليمفاوي المزمن باستخدام التحليل التميزي والشبكات العصبية الاصطناعية" مجلة العلوم الإحصائية، كلية الإدارة والاقتصاد، جامعه كربلاء، العدد 28

[6] حمزة، سعد كاظم، (2009)، "مقارنة بعض الطرائق اللبية في تقدير نماذج الانحدار اللامعلمي بوجود بيانات تامة وغير تامة "، قسم الاحصاء، رسالة ماجستير، كلية الادارة والاقتصاد، جامعة بغداد.

[7] حمود، مناف يوسف، (2000) ،"مقارنة مقدرات (kernel) اللامعلمي لتقدير دوال الانحدار رسالة ماجستير، قسم الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

[8] حمود، مناف يوسف، (2010)، " مقارنة مقدرات معلمة التمهيد متغيرة باستخدام مقدر انحدار خطي موضعي مع تطبيق عملي " المجلة العراقية للعلوم الاحصائية ، العدد 3

[9] خمو، خلود يوسف، (2004)،"مقارنة اساليب بيز مع طرائق اخرى لتقدير منحنى الانحدار اللامعلمي"، اطروحة دكتوراه ، قسم الاحصاء ، كلية الادارة والاقتصاد، جامعة بغداد.

[10] خمو، خلود يوسف، (2005)،"مقارنة الطرائق الشرائحية لتقدير منحنى الانحدار اللامعلمي"، المجلة العراقية للعلوم الاحصائية ، العدد 8.

[11] رشيد، حسام عبد الرزاق، (2014)،" الممهدات اللامعلميه لأنموذج المعاملات المتغيرة والمتغيرة جزئيا"، اطروحة دكتوراه ، قسم الاحصاء ، كلية الادارة والاقتصاد ، جامعة بغداد.

[12] حمود، مناف يوسف، كاطع ، مياسة محمد، (2014)، " مقارنة مقدرات انموذج شبه معلمي باستعمال طرائق تمهيد مختلفة ، قسم الاحصاء، كلية الادارة والاقتصاد ، جامعة بغداد.

[13] Aydin, D., 2007. A comparison of the nonparametric regression models using smoothing spline and kernel regression. World Academy of Science, Engineering and Technology, 36, pp.253-257.

[14] Berkson, J. (1994),"Application of the logistic to Bio-Assay", JASA Vol. 39,No.227,pp.357-465.

[15] Chen, W., Chen, Y., Mao, Y. and Guo, B., 2013, August. Density-based logistic regression. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 140-148).

[16] Rahmawati, D.P., Budiantara, I.N., Prastyo, D.D. and Octavanny, M.A.D., 2021. Mixture Spline Smoothing and Kernel Estimator in Multi-Response Nonparametric Regression. IAENG International Journal of Applied Mathematics, 51(3), pp.1-12.

[17] Rifada , Suliyanto, M. and Tjahjono, E., 2020, September. Estimation of nonparametric binary logistic regression model with local likelihood logit estimation method (case study of diabetes mellitus patients at Surabaya Hajj General Hospital). In AIP Conference Proceedings (Vol. 2264, No. 1, p. 030007). AIP Publishing LLC.

[18] Shim, J. Y., and Seok, k. H.(2012) "Semi parametric kernel logistic regression with longitudinal datac".Journal of the Korean Data and information Science Society, 23 (2) ,385-392.

[19] Signorini, D. F., and Jones, M. C. , 2004 . Kernel estimators for univariate binary regression. Journal of the American Statistical Association, vol 99 no 465, pp 119-126

[20] Sylvain, A. and Alain, C., (2010) ,"Asurvey of cross _validation procedures for model selection". Universit' e Lille l,

[21] Wu,H.Zhang,.T.,(2006) "Nonparametric Logistic Regression with application to Assessing Drug Safety lssues" Minnesota, USA.

[22] Wibowo, W., Amelia, R., Octavia, F. A., and Wilantari, R. N., 2021,. Classification using nonparametric logistic regression for predicting working status. In AIP Conference Proceedings Vol. 2329, No. 1, p. 060032. AIP Publishing LLC.

Journal of AL-Rafidain University College for Sciences (2024); Issue 56; pp. 156-165

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

A Statistical Study of Some Nonparametric Estimators of the Logistic Regression Model

Lekaa A. Mohammed	Narjes B. Khalaf
lekaa.a@coadec.uobaghdad.edu.iq	Narjes.bassim@alayen.edu.iq
Department of Statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq	Department of Medical Laboratories, College of Health and Medical Technologies, Al-Ain University of Iraq, Dhi Qar, Iraq

Article Information

Article History:
Received: February, 16, 2024
Accepted: April, 12, 2024
Available Online: December, 31, 2024

Keywords:
CV boot parameter, GCV mixed method, plug in chip. Grader Nadaria – Watson NW, mixed method of core estimator and smoothed slide (LBS, Logt algorithm positional LL) (Positional Scoring Algorithm (LSA)

Correspondence:
Narjes B. Khalaf
Narjes.bassim@alayen.edu.iq

Abstract

The study provides an estimate of the nonparametric binary logistic regression model with the likelihood of chronic lymphocytic leukemia depending on the factors affecting the injury (patient sex, patient's age, white blood cells, hemoglobinemia, hematoglocytes, platelets) and lymphocytic leukemia is one of the most dangerous diseases for human life, it may result in many complications that may lead to exposure to other types of blood Irrigation of cancer, uncontrollable. The study aims to identify the most important statistical methods in analyzing cancer patient data and identifying the most important influences on it, which are categorical variables, and their importance lies in how to implement predictions after obtaining the best logistical model. Using the mixed method of core estimator and smoothed slide to estimate an optimal smoothing parameter using legitimate transit and generalized legitimate transit, it was concluded that the mixed method of core estimator and smoothed slide (LBS) is the best through the comparison criterion average squared error.

DOI: https://doi.org/10.55562/jrucs.v56i1.14