https://doi.org/10.31272/jae.i148.1426
https://admics.uomustansiriyah.edu.iq
P-ISSN: 1813–6729 E-ISSN: 2707-1359

JAE

تقدير معلمات أنموذج الانحدار اللوجستي باستعمال طريقة المقدرات المعدلة

حسنين جليل نعمه الساعدي
قسم كلية معلوماتية الاعمال جامعة تكنولوجيا المعلومات والاتصالات، بغداد، العراق.
Email: hasanien.1975@uoitc.edu.iq , ORCID: https://orcid.org/0009-0009-8128-3068

فراس منذر جاسم
قسم الاحصاء، كلية الادارة والاقتصاد، الجامعة المستنصرية، بغداد، العراق.
Email: firasm@uomustansiriyah.edu.iq, ORCID: https://orcid.org/0000-0002-4948-8583

المستخلص

تعد مشكلة الفصل بين مشاهدات المتغير التابع ثنائي الاستجابة الذي يعتمد على حجم العينة، ومشكلة التعدد الخطي بين المتغيرات التوضيحية، من اهم المشاكل التي تظهر في أنموذج الانحدار اللوجستي. وتتطلب معالجة هذه المشاكلات تطبيق طرائق فعالة، حيث تم اعتماد طرائق التقدير مقدرات الإمكان غير الأعظم الجزائية المزدوجة (DPMLE) وطريقة المقدرات المعدلة (Adjusted) في هذا البحث. ومن اهم النتائج التي تم التوصل اليها بعد تشخيص مشكلة الفصل والتعدد الخطي للبيانات الحقيقية المتمثلة بالإصابة بفقر الدم، والتي تم الحصول عليها من المختبرات الخاصة بأمراض الدم في مستشفى مدينة الطب، ان طريقة المقدرات المعدلة هي الافضل في معالجة مشكلتي الفصل والتعدد الخطي، وذلك بالاعتماد على متوسط مربعات الخطأ (Mse) كمعيار للمقارنة.

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث: 28/ 4 / 2025
تاريخ قبول البحث: 26/ 5 / 2025
تاريخ نشر الكتروني 01/ 06 / 2025
عدد صفحات البحث 88 - 98

الكلمات المفتاحية:
المقدرات المعدلة، التعدد الخطي، فصل البيانات، دالة اللوجستيك، التباين الموزون.

المراسلة:
أسم الباحث: فراس منذر جاسم
Email: firasm@uomustansiriyah.edu.iq

1. المقدمة

تحليل الانحدار هو التحليل الذي يختص بدراسة اعتماد متغير واحد يعرف بمتغير الاستجابة على متغير واحد أو أكثر تعرف بالمتغيرات التوضيحية، وذلك لغرض التقدير والتنبؤ بقيمة المتغير التابع باعتماد معلومات المتغيرات التوضيحية. ويعتبر أنموذج الانحدار اللوجستي من نماذج الانحدار غير الخطية يوضح العلاقة بين المتغير التابع ثنائي الاستجابة والمتغيرات التوضيحية او التفسيرية. ان اول من استعمل أنموذج الانحدار اللوجستي هو الباحث (Verhulst)، حيث استخدم دالة اللوجستيك (logistic function) لوصف نمو المجتمع وكانت تسمى هذه الدالة بدالة النمو (growth function). وفي عام (1920) قام الباحثان (pearl and reed) باستخدام دالة لحساب نمو السكان وأطلق عليها فيما بعد بدالة اللوجستيك، وتبرز العديد من استخدامات هذا الأنموذج في الدراسات المتعلقة بعلوم الحياة والعلوم الزراعية والطبية وبشكل عام في الدراسات ذات الطابع التجريبي، لكونه من النماذج الملائمة للبيانات الثنائية (binary data). وظهرت العديد من الابحاث التي تهتم بدراسة بعض المشكلات في أنموذج الانحدار اللوجستي مثل مشكلة التعدد الخطي (Multicollinearity) بين المتغيرات التوضيحية، ومشكلة الفصل (Separation) بين مشاهدات المتغير التابع ثنائي الاستجابة.

تعتمد مشكلة التعدد الخطي على وجود علاقة خطية قوية بين المتغيرات التوضيحية أي وجود ارتباط قوي بين هذه المتغيرات فتصبح مقدرات انموذج الانحدار اللوجستي غير موثوقة. اما مشكلة الفصل (Separation) فإنها تعتمد على شكل انتشار البيانات أي عندما تكون بيانات المتغير التابع ثنائية الاستجابة وعلى حجم العينة أي عندما تكون العينات ذات حجوم صغيرة، بمعنى كلما ازدادت عدد المشاهدات تقل فرصة الحصول على مشكلة الفصل، وأن امكانية حصول مشكلة الفصل شائع أكثر مما هو متوقع في مجال التطبيقات الطبية.

من اهم الدراسات في هذا السياق كانت في عام (2008) عندما اقترح الباحثان (Urgan and Tez) مقدر(Liu) لأنموذج الانحدار اللوجستي للتنبؤ بالمتغير ثنائي الاستجابة في حالة وجود التعدد الخطي بين المتغيرات التوضيحية الذي سوف يكون له متوسط مربعات الخطأ اقل من مقدرات الإمكان الأعظم (MLE)، وتمت مقارنة مقدر (Liu) مع المقدرات البديلة في الانحدار اللوجستي مثل مقدر الحرف ومقدر المركبات الرئيسية ومقدر شتاين وتحت معيار متوسط مربعات الخطأ اذا تبين ان مقدر ليو(Liu) هو الأفضل على الرغم من كونه مقدرا متحيزا لأنه صاحب اقل متوسط مربعات الخطأ (MSE). في عام (2011) درس الباحثان (Miller and David) مشكلة عدم التقارب في تحليل البيانات لأنموذج الانحدار اللوجستي عندما تكون البيانات تعاني من الفصل شبه التام، حيث جرى استعراض مقدرات أنموذج الانحدار اللوجستي من وجهة نظر النماذج الخطية المعممة وتم تطبيق هذا الأجراء لبيانات البحوث التربوية لإثبات نجاحها في القضاء على المشكلة. وفي عام (2013) قدم الباحث (Ogoke) واخرون طريقة لتعديل

مقدرات انحدار الحرف اللوجستي المعمم بوساطة دالة الاستجابة الأسية لمصفوفة الأوزان مما يقلل من التباين، حيث تم استخدام المقدر المعدل (Jackknife) لتخفيض التباين، وتوصلوا الى ان مقدر (Jackknife) متفوق على كل من مقدر الحرف اللوجستي ومقدر الحرف اللوجستي المعدل من خلال التباين والحد من التحيز.

إن طرائق التقدير التقليدية لمعلمات أنموذج الانحدار اللوجستي عند تحليل البيانات ثنائية الاستجابة(Binary data response) ضعيفة في معالجة المشكلات التي تظهر في انموذج الانحدار اللوجستي مثل الفصل بين بيانات المتغير التابع ثنائية الاستجابة والتعدد الخطي بين المتغيرات التوضيحية ومشكلة عدم تجانس التباينات.

2. هدف البحث

يهدف البحث الى تقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلتي الفصل والتعدد الخطي معاً، حيث يتم ذلك من خلال تطبيق طرائق التقدير على بيانات فقر الدم التي تعاني مشكلة الفصل والتعدد الخطي في نفس الوقت، بغية التوصل الى اهم العوامل المؤثرة على الإصابة بفقر الدم. كما يهدف البحث الى المقارنة بين طرائق التقدير من خلال اعتماد معيار (Mse) كمعيار للمقارنة.

3. أنموذج الانحدار اللوجستي

أحد نماذج الانحدار الذي تكون فيه العلاقة بين المتغير التابع (y) والمتغيرات التوضيحية ($x_1, x_1, x_2 \dots x_p$) غير خطية حيث يكون المتغير التابع ثنائي الاستجابة مفترضاً إحدى القيمتين (1,0)، أما النجاح أي حدوث الاستجابة باحتمال (π_i) أو الفشل أي عدم حدوث الاستجابة باحتمال ($1-\pi_i$)، لذلك يكون المتغير التابع يتبع توزيع برنولي وسوف تكون دالة الكثافة الاحتمالية بالصيغة (1) الأتية: [17]

$$p(Y=y_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad ,\forall\, y=0,1 \tag{1}$$

حيث يمثل (y_i) متغير تابع ثنائي الاستجابة. ويمثل (π_i) احتمال حدوث الاستجابة (دالة الانحدار اللوجستي) عندما ($y_i=1$). لذلك فأن توقع المتغير التابع يمثل احتمال حدوث الاستجابة، أي ان:

$$E(y_i)=p(Y=1)=\pi_i \tag{2}$$

أما تباين المتغير التابع فيكون بالصيغة (3) الاتية:

$$Var(y_i)=\pi_i(1-\pi_i) \tag{3}$$

ولتكن ($x_1, x_2, \dots \dots x_k$) مجموعة من المتغيرات التوضيحية ولتكن (n) تمثل عدد المشاهدات لهذه المتغيرات التي تكون المصفوفة X. بالتالي فأن نموذج الانحدار اللوجستي يكتب بالصيغة (4) الاتية:

$$Y_i=\pi_i+\varepsilon_i \tag{4}$$

اذ يمثل $Y_i=[y_1, y_2, \dots \dots y_n]$ يمثل عينه عشوائية من المتغير ثنائي الاستجابة، وأن (π_i) يمكن التعبير عنها بالصيغة (5) الاتية:

$$\pi_i=p(y=1)=\frac{e^{\underline{x}_i\underline{\beta}}}{1+e^{\underline{x}_i\underline{\beta}}} \tag{5}$$

حيث يمثل ($\underline{\beta}$) متجه من المعلمات ذي الابعاد (p×1)، وان $\underline{x}_i=\{x_{i1}, x_{i2}, \dots x_{ip}\}$ متجه صفي من المتغيرات التوضيحية ذي الابعاد (1×p). ويمثل الرمز ε_i حد الخطأ العشوائي بمتوسط صفر وتباين يساوي تباين المتغير المعتمد ثنائي الاستجابة، أي ان:

$$Var(\varepsilon_i)=Var(y_i)=\pi_i(1-\pi_i) \tag{6}$$

ان تباين حد الخطأ يعتمد على قيم احتمال الاستجابة π_i أي على قيم المتجه $\underline{x}_i$ وبالتالي سوف يكون تباين حد الخطأ غير متجانس[14]

4. مشكلة الازدواج الخطي

تحصل مشكلة التعدد الخطي عندما تكون هناك علاقة خطية قوية بين بعض أو كل المتغيرات التوضيحية وأن الارتباط بين هذه المتغيرات يعرف بالتعدد الخطي، أي تظهر مشكلة التعدد الخطي عندما تكون قيمة أحد المتغيرات التوضيحية متساوية لجميع المشاهدات، أو عند اعتماد قيمة أحد المتغيرات التوضيحية على قيمة واحدة أو أكثر من المتغيرات التوضيحية في الأنموذج علما بان هذه المشكلة قد تواجه الباحث في حالة تجانس أو عدم تجانس التباين وفي حالة تكون البيانات على شكل سلاسل زمنية أو على شكل بيانات مقطعية [13] أن أحد الشروط الواجب توفرها في نموذج الانحدار عند عدم وجود مشكلة التعدد الخطي بصورة عامة هو شرط الرتبة (rank condition).

$$rank(X)=k+1<n \tag{7}$$

حيث تمثل X مصفوفة المتغيرات التوضيحية ذات الابعاد (n×)k+1()، وعليه عندما تكون المتغيرات التوضيحية مستقلة خطيا يمكن أيجاد معكوس المصفوفة ($\acute{X}X$) وبالتالي يمكن أيجاد تقديرات المعلمات، أما أذا كان هناك علاقة خطية بين اثنين أو أكثر من المتغيرات فان ذلك سيؤدي الى انتهاك شرط الرتبة، لذا لا يمكن أيجاد معكوس مصفوفة المعلومات ($\acute{X}X$) وبالتالي لا توجد مقدرات المعلمات وهذه الحالة تسمى بالتعدد الخطي التام حيث أن $|\acute{X}X|=0$ أي إن محدد مصفوفة المعلومات بين المعلمات تساوي صفراً [8]

أما أذا كان محدد المصفوفة قريب من الصفر $|\acute{X}X| \approx 0$ اي عندما تميل المتغيرات للتحرك سوية بالزيادة أو الانخفاض أو في حالة استخدام المتغيرات المرتدة زمنياً (lagged variables)، فأن تقدير معلمات الأنموذج ستكون غير دقيقة وغير ممثلة لواقع المشكلة، وتباينات المعلمات المقدرة ستكون كبيرة جدا [1]

5. تشخيص مشكلة الازدواج الخطي الموزون في الانحدار اللوجستي

من اجل الحصول على التشخيص المناسب لمشكلة التعدد الخطي الموزون يفضل استعمال مصفوفة المعلومات الموزونة. لتكن $(y_1,y_2,...y_n)$ تمثل (n) من قيم المتغيرات العشوائية ثنائية الاستجابة المستقلة فيما بينها والتي تتبع توزيع برنولي، فأن المصفوفة القطرية $(\widehat{W})$ تمثل التباين المقدر لقيم (y) أي تباين حد الخطأ العشوائي (مصفوفة الاوزان) بالصيغة (8) الاتية: [9]

$$\widehat{W} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_{1)} & \dots & \cdots \\ \vdots & \ddots & \vdots \\ \vdots & \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix} \tag{8}$$

ويمكن التعبير عن مصفوفة المعلومات الموزونة المقدرة بالصيغة (9) الاتية.

$$\widehat{\Phi} = \acute{X}\widehat{W}X \tag{9}$$

وتوجد طريقتان لتشخيص مشكلة التعدد الخطي الموزون في أنموذج الانحدار اللوجستي وهي كالاتي:

1.5. العدد الشرطي الموزون

تعتمد قيم اعداد الشرط الموزون على مصفوفة الارتباطات الموزونة المقدرة $(\widehat{\Phi}^*)$ في تشخيص مشكلة التعدد الخطي الموزون، اذ تشير الحالة التي تكون فيها قيم اعداد الشرط أكبر من 30، أي ان $(k_j > 30)$ إلى حالة التشخيص أي وجود مشكلة التعدد الخطي الموزون بين المتغيرات التوضيحية في أنموذج الانحدار اللوجستي. ولتكن $(\lambda_0^*, \lambda_1^*, \dots, \lambda_p^*)$ الجذور المميزة لمصفوفة الارتباطات الموزونة المقدرة $(\widehat{\Phi}^*)$ التي يعبر عنها بالصيغة (10) الاتية:

$$\widehat{\Phi}^* = \acute{\hat{S}}^*\hat{S}^* \tag{10}$$

إذا ان حيث إن $(\hat{S} = \widehat{W}^{0.5}X)$، بالتالي فأن قيم اعداد الشرط تعرف كالاتي:

$$k_j = \left(\frac{\lambda^* \max}{\lambda_j^*}\right)^{0.5} \tag{11}$$

اذ إن $(\lambda^*\max)$ هو أكبر الجذور المميزة، وان (λ_j^*) هو الجذر المميز (j) في المصفوفة $(\widehat{\Phi}^*)$. [17]

2.5. نسبة التباين الموزون

يعتمد تشخيص مشكلة التعدد الخطي في هذه الحالة على قيم نسبة التباين الموزونة التي تقع بين الصفر والواحد الصحيح، لذا ففي حالة وجود قيمتين من قيم نسبة التباين الموزونة قريبتان من الواحد الصحيح تحدث مشكلة التعدد الخطي الموزون بين متغيرات هذه النسب. ولتكن (M) مصفوفة المتجهات المتعامدة التي يتم الحصول من المصفوفة $(\widehat{\Phi}^*)$، وان (Λ^*) هي مصفوفة قطرية من الجذور المميزة للمصفوفة $(\widehat{\Phi}^*)$ التي تحقق الشرط الاتي:

$$\acute{M}\widehat{\Phi}^*M = \Lambda^* \tag{12}$$

فان أي عنصر من نسبة التباين الموزونة وليكن (ω_{uj}) يمكن التعبير عنها بالصيغة (13) الاتية:

$$\omega_{uj} = \frac{{m_{ju}}^2/\lambda_u^*}{C_{jj}} \tag{13}$$

إذا يمثل (m_{ju}) عنصر من مصفوفة المتجهات المميزة (M) من الرتبة $(j\times u)$. وان الرمز (λ_u^*) يمثل الجذر المميز المحسوب من مصفوفة الارتباطات الموزونة المقدرة $(\widehat{\Phi}^*)$. ويمثل الرمز (C_{jj}) الجذور المميزة الصغيرة (القيم الذاتية الصغيرة نسبة الى أكبر جذر مميز) ويمكن أن يحسب بالصيغة (14) الاتية:

$$c_{jj} = \sum_{u=1}^{p} {\lambda_u^*}^{-1} {m^2}_{ju} \tag{14}$$

إذا يمثل (u) عدد الجذور المميزة. [9]

6. فصل البيانات في الانحدار اللوجستي

أن مجموعة نقاط العينة (n) يمكن تصنيفها في واحد من ثلاث تكوينات متنافية للمتغير التابع (الفصل التام، الفصل شبه التام، التداخل)، اذ يستخدم مصطلح الفصل لوصف مجموعة من نقاط العينة التي تنتمي الى تكوين الفصل التام أو الفصل شبه التام، وعندما يكون هناك فصل تام أو فصل شبه تام بين مشاهدات العينة للبيانات فان مقدرات الإمكان الأعظم تكون غير دقيقة في حدود فضاء المعلمات.[7]

تعتمد مشكلة الفصل على كل من البيانات والانموذج أي أن مشكلة الفصل تحدث بالدرجة الأولى مع العينات الصغيرة ويمكن أن تتحقق من خلال زيادة عدد المتغيرات التوضيحية الواردة في الأنموذج، وفي أنموذج الانحدار اللوجستي تظهر مشكلة الفصل

بشكل واسع عندما تكون هناك (g) من اصناف الاستجابة، بالتالي وجود الفصل قد يشير الى أن هذا الانموذج هو أكثر تركيب للبيانات لذلك عند اختيار أنموذج مختلف على سبيل المثال تحليل التمايز يمكن تجنب المشاكل الناجمة عن الفصل تماما. [3]

لتكن عينة بحجم (n) من المشاهدات عندما $\underline{x_i} = (x_{i0}, x_{i1} \dots \dots x_{ik})$ هو متجه من المتغيرات التوضيحية، حيث ان $(x_{i0} = 1)$، وان (y) هو متغير الاستجابة الذي يأخذ القيم $(y_i = y_1, y_2, \dots, y_n)$ حيث أن كل مشاهدة من y_i محددة بواحدة من (g) من مجاميع الاستجابة $(y_i = y_{i1}, y_{i2}, \dots, y_{ig})$ للتعبير عن الدالة اللوجستية متعددة الاستجابة يكون بالشكل التالي. [2]

$$\pi_{is^{**}} = p_r(y_{s^{**}}/x) = exp\left(\underline{x_i}\underline{\beta_{s^{**}}}\right).p_r\left(\frac{y_g}{x}\right) \quad (15)$$

إذا ان

$$p_r(y_g/x) = \frac{1}{\{\sum_{s=1}^{g} exp(\underline{x_i}\underline{\beta_{s^{**}}})\}} \quad (16)$$

$$\pi_{s^{**}i} = \frac{exp(\underline{x_i}\underline{\beta_{s^{**}}})}{\{\sum_{s=1}^{g} exp(\underline{x_i}\underline{\beta_{s^{**}}})\}} \quad (17)$$

إذا يمثل الرمز $(s^{**} = 1,2,3 \dots\dots g)$ عدد مجاميع الاستجابة، وان $\beta_{s^{**}}^{T} = (\beta_{s^{**}0}, \beta_{s^{*}1}, \dots \dots \beta_{s^{*}k})$ يمثل متجه المعلمات عندما تكون الاستجابة (s^{*}).

إذا كان هناك (x_i) موجه صفي وهو جزء من مصفوفة المتغيرات التوضيحية (X)ذات الابعاد $(n\times k)$، وان $(E_{s^{**}})$ يمثل مجموعة المتغيرات التوضيحية في المصفوفة (X) عندما يكون متغير الاستجابة $(y_{s^{**}})$، أي ان $(E_{s^{**}})$ يدل على مجموعة المتغيرات التوضيحية عندما تكون الاستجابة s^{**}، وليكن $\beta = (\underline{\beta_1}, \underline{\beta_2}, \dots, \underline{\beta_{g-1}})$ مصفوفة معلمات الاستجابات، أي عدد المعلمات لكل استجابة. وبذلك يكون لدينا مصفوفة أبعادها $v=((p)\times(g-1))$ من المعلمات ويشترط أن يكون عدد المشاهدات أكبر من عدد المعلمات $v<n$، فان دراسة مشكلة فصل بيانات العينة يتم من خلال الفصل التام والفصل شبه التام والتداخل كالاتي. [3]

1.6. الفصل التام

يحدث الفصل التام في مشاهدات العينة أذا كان هناك متجه المعلمات $(\underline{\beta}\epsilon R^{V})$ حيث إن (R^{V}) هو فضاء المعلمات لكل $(\underline{x_i}\epsilon E_j)$ والقيم $(j, t=1,2,\dots,g;\ j\neq t)$ بشرط:

$$\underline{x_i}\left(\underline{\beta_j} - \underline{\beta_t}\right) > 0 \quad (18)$$

ويمكن القول إن في حالة أنموذج الانحدار اللوجستي ثنائي الاستجابة يحدث الفصل التام بين مجموعة من مشاهدات العينة (n) أذا كان لدينا متجه من المعلمات $(\underline{\beta})$ بشرط $(\underline{x_i}\underline{\beta} > 0)$ عندما $(y_i = 1)$ وان $(\underline{x_i}\underline{\beta} < 0)$ عندما $(y_i = 0)$ ويمكن مشاهدة الفصل التام من خلال الجدول (1) التالي: [4]

جدول (1) يوضح الفصل التام للبيانات.

x	Y
-4	0
-3	0
-2	0
-1	0
1	1
2	1
3	1
4	1

2.6. الفصل شبه التام

يمكن ان تكون هناك حالة تسمى الفصل شبه التام بين مشاهدات العينة إذا كان متجه من المعلمات $(\underline{\beta}\epsilon R^{V})$بحيث لكل $(\underline{x_i}\epsilon E_j)$ ولكل قيم $(j, t=1,2,\dots,g;\ j\neq t)$ بشرط.

$$\underline{x_i}\left(\underline{\beta_j} - \underline{\beta_t}\right) \geq 0 \quad (19)$$

يحصل الفصل شبه التام في حالة الأنموذج اللوجستي ثنائي الاستجابة، كلما كانت الدالة من المتغير (x) تولد تنبؤات غير مثالية الى (y) فاذا كان لدينا متجه من المعلمات $(\underline{\beta})$ بحيث $(\underline{x_i}\underline{\beta} \geq 0)$عندما (y=1) و $(\underline{x_i}\underline{\beta} \leq 0)$ عندما (y=0) ويمكن النظر الى حالة الفصل شبه التام من خلال الجدول (2) التالي:

جدول (2) يوضح الفصل شبه التام للبيانات.

X	Y
-4	0
-3	0

-2	0
-1	0
0	0
0	1
1	1
2	1
3	1
4	1

إن في حالة وجود الفصل التام والفصل شبه التام للبيانات يمكن ملاحظة.

1- مقدرات الإمكان الأعظم المرجحة $(\hat{\underline{\beta}})$ تكون غير دقيقة.
2- أن حالة الفصل التام وشبه التام تودي الى حدوث تباعد في عملية التكرار المثالية عند استخراج مقدرات الإمكان الأعظم التكرارية. [4]

3.6. التداخل

التداخل يمثل الحالة الطبيعية لبيانات المتغير التابع ثنائي الاستجابة ففي حالة عدم وجود الفصل التام والفصل شبه التام بين نقاط العينة سوف يكون هناك تداخل للبيانات أي وجود متجه من المعلمات $(\underline{\beta} \epsilon R^V)$ وان $(\underline{x}_i \epsilon E_j)$ للقيم $(j, t=1,2,..., g;\ j\neq t)$ بشرط.

$$\underline{x}_i(\underline{\beta}_j - \underline{\beta}_t) < 0 \tag{20}$$

في حالة التداخل تكون المعلمات المقدرة $(\hat{\underline{\beta}})$دقيقة وان عملية التكرار تتقارب إلى حل نهائي وحيد وتكون تقديرات المعالم ذات قيمة محددة. [3]

7. الكشف عن مشكلة الفصل التام ومشكلة الفصل شبه التام

يتم الكشف عن مشكلة الفصل التام ومشكلة الفصل شبه التام من خلال الاعتماد على مقدرات الإمكان الأعظم التكرارية $(\hat{\underline{\beta}})$ ففي حالة وجود المشكلة عند تقدير المعلمات يستمر التكرار حتى يتم تجاوز التكرار الثابت. لذلك سوف تكون مقدرات المعلمات غير دقيقة، ويكون حد الخطأ كبير للغاية. اقترح الباحثان ألبرت وأندرسون عام 1984 وسيلة تجريبية للكشف عن مشكلة الفصل التي كان تنفيذها في (Protocol Logistic) والذي يحتوي على الخطوات الأتية: [3]

1- إذا تحقق معيار التقارب في غضون ثماني تكرارات نستنتج عدم وجود مشكلة الفصل بصورة عامة اي هناك تداخل بين البيانات.
2- لجميع التكرارات بعدد التكرار الثامن، ان احتمال توقع الاستجابة (دالة الانحدار اللوجستي) لكل مشاهدة يمكن حسابة من قبل الصيغة (21) الأتية:

$$\hat{\pi}_i = \frac{1}{1 + \exp[(2y_i - 1)\underline{x}_i\hat{\underline{\beta}}]} \tag{21}$$

فاذا كان احتمال توقع الاستجابة مساو الى الواحد، ولجميع مشاهدات دالة الانحدار اللوجستي نستنتج أن هناك فصلا تاماً وفي هذه يتوقف التكرار. اما أذا كان احتمال توقع الاستجابة الى مشاهدات دالة الانحدار اللوجستي أكبر من (0.95) لبعض المشاهدات نستنتج الى أن هناك فصل شبه تام ويتوقف التكرار [20]

8. تقدير معلمات أنموذج الانحدار اللوجستي

1.8. مقدرات الإمكان الأعظم الجزائية المزدوجة

في هذه الطريقة يقترح الباحثان (Shen & Gao) إدخال معلمة الحرف على مقدرات الإمكان الأعظم الجزائية، أي تضم هذه الطريقة احتمال (Firth) الجزائي مع معلمة الحرف والتي هي قادرة على التعامل مع كل من مشكلتي الفصل والتعدد الخطي في انموذج الانحدار اللوجستي لتحسين مقدرات المعلمات وتقليل حد الخطأ العشوائي. ومن خلال تعظيم $L(\beta, X)$ تنتج مقدرات الإمكان الأعظم $(\underline{\beta})$ والتي تكون متحيزة عند $(\underline{\beta} = 0)$، ويمكن التعبير عن تحيز مقدرات الإمكان الأعظم (MLE) بالصيغة (22) الاتية: [18]

$$bias\left(\hat{\underline{\beta}}\right) = \left[\frac{b1\left(\underline{\beta}\right)}{n} + \frac{b2\left(\underline{\beta}\right)}{n^2} + \cdots\right] \tag{22}$$

تركز معظم طرائق تصحيح التحيز على إزالة التحيز من الدرجة الأولى، وذلك باستخدام الصيغة (23) التالية:

$$\hat{\underline{\beta}}_{correct} = \hat{\underline{\beta}} - \frac{b_1\left(\hat{\underline{\beta}}\right)}{n} \tag{23}$$

بعد إضافة حد جزائي ثاني الى دالة الإمكان الجزائية (Firth) بواسطة معلمة الحرف التي تفرض على نطاق المعلمات المقيدة فان لوغاريتم دالة الإمكان الجزائية المزدوجة تكون كالآتي:

$$logL(\beta, X)^{**} = logL(\beta, X) + \frac{1}{2} log|l(\beta)| - \lambda ||P\beta||^2 \quad (24)$$

إذا يمثل الرمز (P) مصفوفة من الرتبة (p×p) التي تفرض قيود خطية على المعلمات ($\underline{\beta}$)، في دالة الإمكان الجزائية المزدوجة $L(\beta, X)^{**}$، يكون اختيار معلمة الحرف (λ) يعتمد على درجة التعدد الخطي بين المتغيرات فيتم اختيار معلمة الحرف التي تحقق اقل متوسط مربعات الخطأ [10]. بالتالي فأن المقدرات يمكن الحصول عليها من خلال خوارزمية نيوتن-رافسون التكرارية، كما موضح بالصيغة (25) الاتية:

$$\underline{\hat{\beta}}_{DP}{}^{(t+1)} = \underline{\hat{\beta}}_{DP}{}^{(t)} + l(\beta^*)^{-1(t)} U(\underline{\beta}^{**})^{(t)} \quad (25)$$

إذا ان $(l(\beta^*)^{-1})$ تمثل معكوس مصفوفة المعلومات لدالة الإمكان الجزائية المزدوجة. والتي يمكن التعبير عنها بالصيغة (26) الاتية:

$$l(\beta^*) = \left(\acute{X}WX\right) - 2\lambda I \quad (26)$$

إذا تمثل (I) مصفوفة الوحدة. ويمثل الرمز $(U(\underline{\beta}^{**}))$ المشتقة الأولى إلى لوغاريتم دالة الإمكان الجزائية المزدوجة، ويمكن التعبير عنها بالصيغة (27) الاتية: [18]

$$U\left(\underline{\beta}^{**}\right) = U\left(\underline{\beta}\right) + \frac{\partial}{\partial \underline{\acute{B}}}\left(\frac{1}{2} log|l(\beta)|\right) - 2\lambda\underline{\beta} \quad (27)$$

إذا ان $U\left(\underline{\beta}\right) = \frac{\partial logL(\beta, X)}{\partial \underline{\acute{\beta}}} = \acute{X}(y_i - \pi_i)$

2.8. المقدرات المعدلة

يتم بناء تقديرات هذه الطريقة بتقليص تقديرات الإمكان الأعظم الجزائية بدل من تقديرات الإمكان الأعظم التكرارية من خلال قيمة (C) كما مبين بالصيغة (28) الأتية.

$$\underline{\hat{\beta}}_{Ad} = C\underline{\hat{\beta}}_{pl} \quad (28)$$

تقع قيمة (C) بين الصفر والواحد الصحيح. لذلك سيكون حساب قيمة (C) بالاعتماد على مقدرات الإمكان الأعظم الجزائية كما مبين في الصيغة (29) الأتية:

$$C = \frac{\underline{\hat{\acute{\beta}}}_{pl}\underline{\hat{\beta}}_{pl}}{\{\underline{\hat{\acute{\beta}}}_{pl}\underline{\hat{\beta}}_{pl} + trace(\widehat{\Phi^{**}}^{-1})\}} \quad (29)$$

إذا يمثل الرمز $(\hat{\beta}_{pl})$ مقدرات الإمكان الأعظم الجزائية او مقدرات Firth. والتي يمكن التعبير عنها بالصيغة (30) الاتية [11]

$$\underline{\beta}_{Pl}{}^{(t+1)} = \underline{\beta}_{Pl}{}^{t} - I(\beta)^{-1(t)} U\left(\underline{\beta}^*\right)^{(t)} \quad (30)$$

3.8. اختيار معلمة الحرف

معلمة الحرف (λ) يتم اختيارها لتقليل متوسط مربعات الخطأ لأنموذج الانحدار اللوجستي هناك تعريفات مختلفة عن أخطاء التنبؤ قدمها Efron عام (1986) وعلى سبيل المثال هناك ثلاث تعريفات لقياس أخطاء التنبؤ كما في الصيغ الاتية: [6]

أ- الخطأ التصنيفي (CE)

$$CE\begin{cases} 1 & if\ y = 1\ and\ \hat{\pi} < 1/2\ \ or\ y = 0\ and\ \hat{\pi} > 1/2 \\ 1/2 & if\ \hat{\pi} = 1/2 \\ 0 & otherwiese \end{cases} \quad (31)$$

ب- الخطأ التربيعي (SE)

$$SE = (y_i - \hat{\pi}_i)^2 \quad (32)$$

جـ- خطأ لوغاريتم الإمكان السالب (ml)

$$ml = -\{y_i log\hat{\pi}_i + (1 - y_i) lo\, g(1 - \hat{\pi}_i) \quad (33)$$

اسلوب الخطأ التصنيفي يعبر عن أخطاء تنبؤ الأنموذج في حدود $\pi = 1/2$ بينما اسلوب الخطأ التربيعي واسلوب لوغاريتم الإمكان الناقص يحسب أخطأ التنبؤ في كامل المدى من قيم (π)، ويتم اختيار معلمة الحرف التي تمتلك اقل متوسط مربعات الخطأ المصحح. ان حساب متوسط مربعات الخطأ المصحح (cross validated) $MSEcv^{\lambda}$ يستند على مجموعة البيانات بأكملها عند مقارنة أثار معلمات الحرف المختلفة، ويعرف بالصيغة (34) الاتية:

$$MSEcv^{\lambda} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\pi}^{\lambda}_{(-i)})^2 \quad (34)$$

إذا يدل الرمز $(\hat{\pi}^{\lambda}_{(-i)})$ على تنبؤ $\pi_{(-i)}$ عندما تعطى (λ) استنادا على تقديرات معلمة التصحيح، اذ يتم اختيار (λ) المثالية بالاعتماد على MSE. [5]

9. وصف البيانات

تم الحصول على البيانات من المختبرات الخاصة بأمراض الدم في مستشفى مدينة الطب، وقد تم الاستعانة بمجموعة من الأطباء المتخصصين بأمراض فقر الدم لتصنيف اهم العوامل المؤثر على المرض، اذ تم جمع البيانات الخاصة بالبحث بالعام 2024 وبعينة عشوائية حجمها (29) شخص من خلال اجراء تحاليل الدم لكل شخص واخذ اهم المتغيرات التوضيحية الاتية:

الهيموغلوبين الدم (Hb)، حجم خلايا الدم المرصوصة (PCV)، خلايا الدم البيض(WBC) ، والجنس لكل شخص، وبغية تسهيل مهمة تحليل هذه البيانات فقد تم اعتبار المتغير التابع y هو الإصابة بمرض فقر الدم (y=1) او عدم الإصابة بمرض فقر الدم (y=0) بالاعتماد على المتغيرات التوضيحية الاتية.

10. تحليل وتفسير النتائج

لمعرفة توزيع بيانات متغير الإصابة بفقر الدم (المتغير ثنائي الاستجابة) تم اجراء اختبار حسن المطابقة، اذ اتضح ان البيانات تتبع توزيع برنولي، حيث بلغت قيمة إحصاءه (Anderson Darling) بلغت (11.415)، وكما مبين في الجدول رقم (3) الاتي:

جدول (3) نتائج اختبار حسن المطابقة متغير الإصابة بفقر الدم

# No	Distributions	Anderson Darling	Rank
1	Bernoulli	11.415	1
2	Binomial	11.432	2
3	D. Uniform	21.578	5
4	Geometric	15.509	4
5	Poisson	13.663	3
6	Hypergeometric	No fit	-
7	Logarithmic	No fit	-
8	Neg. Binomial	No fit	-

كما تم استعمال لغة البرمجة الإحصائية (R) للحصول على النتائج اذ تم الكشف عن وجود مشكلة التعدد الخطي بين المتغيرات التوضيحية من خلال حساب مصفوفة الارتباط المقدرة الموزونة $(\hat{\Phi}^{*})$ بين المتغيرات التوضيحية. وكما موضح بالجدول رقم (4) الاتي:

جدول (4) يبين مصفوفة الارتباط المقدرة الموزونة بين المتغيرات التوضيحية

Explanatory Variables	Hb	PCV	WBC	Gander
Hb	1	0.9897	0.9778	0.9398
PCV	0.9897	1	0.9770	0.9418
WBC	0.9778	0.9770	1	0.9576
Gander	0.9398	0.9418	0.9576	1

ونلاحظ من الجدول (4) ان جميع معاملات ارتباط بيرسون ذات قيم كبيرة وطردية الاتجاه لجميع المتغيرات التوضيحية، اذ يرتبط كل متغير مع بقية المتغيرات التوضيحية الأخرى بعلاقات خطية طردية قوية، مما يدل على وجود مشكلة التعدد الخطي الموزون بين المتغيرات التوضيحية. ولبيان تحديد أي المتغيرات التوضيحية يسبب مشكلة التعدد الخطي الموزون تم حساب مصفوفة نسبة التباين الموزون، وكما موضح بالجدول (5) التالي:

جدول (5) يبين الجذور المميزة والعدد الشرطي الموزون ونسبة التباين الموزون للمتغيرات التوضيحية

E-value	K_j	Intercept	Hb	PCV	WBC	Gander
24.8938	1	4.922e-05	2.258e-05	1.709e-05	0.0007	50.001
60.0805	7.7941	e-0412.90	9.530e-04	6.455e-04	0.0004	30.336
10.0243	14.187	e-0424.20	1.169e-03	1.359e-03	0.5236	40.125
70.0010	667.57	e-01308.0	1.125e-01	8.764e-03	0.2145	30.533
40.0002	143.09	e-0131.96	8.853e-01	9.892e-01	0.2608	60.003

تعكس نتائج اعداد الشرط الموزونة (Kj) قيماً كبيره لمتغيرات نموذج الانحدار المتعدد، حيث كانت أكبر تلك القيم هي قيمة المتغير التوضيحي الرابع (الجنس) اذ بلغت (143.09) وهي أكبر من 30 مما يدل على وجود مشكلة التعدد الخطي الموزون بين المتغيرات التوضيحية. كما نلاحظ من الجدول (5) ان قيم نسبة التباين الموزون للمتغير التوضيحي الأول والثاني كبيرتان ويقعان في نفس الجذر المميز الذي يقابل أكبر عدد شرط موزون، ومن ذلك نستنتج إن هناك تعدد خطي يسببه المتغير التوضيحي الأول

والثاني بدرجة كبيرة والمتغير التوضيحي الثالث والرابع بدرجة اقل. وللكشف عن مشكلة الفصل في البيانات تم حساب دالة الانحدار اللوجستي التقديرية (احتمال الاستجابة)، فكانت الاستجابة لبعض المشاهدات أكبر من (0.95) مما يدل على وجود مشكلة الفصل شبه التام كما مبين في الجدول (6) الاتي:

جدول (6) يبين قيم احتمال الاستجابة

i	$\hat{\pi}_i$	i	$\hat{\pi}_i$	i	$\hat{\pi}_i$
1	1	**11**	70.99999	**21**	90.99999
2	1	**12**	1	**22**	1
3	90.99767	**13**	50.99999	**23**	1
4	40.99949	**14**	90.99784	**24**	70.99269
5	20.94475	**15**	00.99959	**25**	80.93166
6	90.98608	**16**	30.99996	**26**	80.99999
7	90.99999	**17**	80.99999	**27**	20.99999
8	70.99999	**18**	40.99999	**28**	10.99994
9	20.98285	**19**	70.99925	**29**	1
10	70.98521	**20**	40.98548		

تم تقدير معلمات انموذج الانحدار اللوجستي باستعمال طرائق التقدير (الإمكان الأعظم الجزائية المزدوجة (DPMLE) وطريقة المقدرات المعدلة (Adjusted))، كذلك تم اجراء اختبار (T) لاختبار معنوية المعلمات المقدرة، وتم الحصول على النتائج الاتية:

جدول (7) تقدير المعلمات انموذج الانحدار اللوجستي واختبار (T test)

Methods / Parameter	DPMLE	Adjusted	T test
$\hat{\beta}_0$	-27.697	-28.273	-9.4008
$\hat{\beta}_1$	0.5943	0.5377	4.0707
$\hat{\beta}_2$	0.4548	0.4645	3.8575
$\hat{\beta}_3$	0.1034	0.2292	3.7723
$\hat{\beta}_4$	2.6157	2.4595	4.4002
MSE	0.0574	0.0344	P-value= 0.002

نلاحظ من الجدول (7) ان طريقة المقدرات المعدلة (Adjusted) أظهرت متوسط مربعات الخطأ (Mse) أصغر من طريقة مقدرات الإمكان الأعظم الجزائية المزدوجة (DPMLE) وبفارق بسيط حيث بلغت قيمته (0.0344)، بالتالي فان طريقة المقدرات المعدلة هي الافضل في معالجة مشكلتي الفصل والتعدد الخطي خاصة عند توفر تلك المشكلتين في البيانات في الوقت ذاته، وذلك لأنها تحقق اقل (Mse). ان تفسير النتائج من خلال اختبار (T) يعطي صوره واضحة عن الإصابة بفقر الدم وما لها من تداعيات، اذ كانت قيمة اختبار (T) المحسوبة معنوية لجميع المعلمات عند مستوى دلاله (p-value =0.01)، اذ وجد ان متغير الجنس هو العامل الأول الذي يؤثر في الإصابة بفقر الدم من خلال قيمة اختبار (T) الخاصة به التي تعطي اعلى قيمة اختبار ثم تأتي بعدها قيمة اختبار (T) الخاصة بالهيموغلوبين (Hb) وبعدها حجم خلايا الدم المرصوصة، واخيرا خلايا الدم البيضاء وهذا ما يتناسب مع واقع الإصابة بفقر الدم.

11. الاستنتاجات

1- اثبتت الطريقة المعدلة (Adjusted) أنها أفضل من طريقة مقدرات الإمكان الأعظم الجزائية المزدوجة (DPMLE) في معالجة البيانات التي تعاني من مشكلة الفصل والتعدد الخطي، وتقدير معلمات انموذج الانحدار اللوجستي من خلال الاعتماد على متوسط مربعات الخطأ (Mse) كمعيار للمقارنة.
2- نستنتج من خلال مصفوفة الارتباطات الموزونة لبيانات الإصابة بفقر الدم، وجود مشكلة التعد الخطي بين المتغيرات التوضيحية، حيث كان المتغير التوضيحي الأول الي يمثل هيموغلوبين الدم يرتبط بعلاقه طردية قوية مع المتغير التوضيحي الثاني الذي يمثل حج خلايا الدم المرصوصة، وهذا يتناسب مع اراء الأطباء من خلال الفحوصات الخاصة للدم التي تجرى في المختبرات الطبية.
3- من خلال نتائج البحث نستنتج ان الهيموغلوبين (Hb) هو العامل الأول والأكثر تأثيراً على الإصابة بمرض فقر الدم، وهذه النتيجة جاءت متوافقة مع اراء الأطباء المختصين في مجال امراض الدم.

12. التوصيات

1- اعتماد طريقة المقدرات المعدلة في تقدير معلمات انموذج الانحدار اللوجستي في حالة وجود مشكلة الفصل والتعدد الخطي.

2- توسيع نطاق البحث من خلال دراسة إمكانية ادخال متغيرات جديدة قد تكون مهمة في عملية تشخيص فقر الدم فهناك متغيرات قد جرى استبعادها لعدم اكتمال البيانات حولها.
3- الاستفادة من دراسة انموذج الانحدار اللوجستي في حالة البيانات متعددة الاستجابة فضلا عن الية تقدير معلمات الانموذج .

References

[1] Abdelwahab, M. M., Abonazel, M. R., Hammad, A. T., & El-Masry, A. M. (2024). Modified Two-Parameter Liu Estimator for Addressing Multicollinearity in the Poisson Regression Model. Axioms, 13(1), 46. https://doi.org/10.3390/axioms13010046
[2] Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71(1), 1-10. https://doi.org/10.1093/biomet/71.1.1
[3] Albert, A., & Lesaffre, E. (1986). Multiple group logistic discrimination. In Statistical Methods of Discrimination and Classification (pp. 209-224). Pergamon. https://doi.org/10.1016/B978-0-08-033621-3.50020-0
[4] Allison, P. D. (2008, March). Convergence failures in logistic regression. In SAS global forum (Vol. 360, No. 1, p. 11). https://doi.org/10.3103/S1066530724700017
[5] Bi, Y., & Jeske, D. R. (2010). The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise. Journal of Multivariate Analysis, 101(7), 1622-1637. https://doi.org/10.1016/j.jmva.2010.03.003
[6] Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association, 70(352), 892-898. https://doi.org/10.1080/01621459.1975.10480305
[7] Guimarães, I. A., & Bubniak, T. S. (2021). Linear programming applied to separation detection in polytomous logistic regression. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, 8(1). . https://doi.org/10.26650/acin.1558583
[8] Lukman, A. F., Adewuyi, E., Månsson, K., & Kibria, B. G. (2021). A new estimator for the multicollinear Poisson regression model: simulation and application. Scientific Reports, 11(1), 3732. https://doi.org/10.1038/s41598-021-82424-x
[9] Marx, B. D., & Smith, E. P. (1990). Weighted multicollinearity in logistic regression: diagnostics and biased estimation techniques with an example from lake acidification. Canadian Journal of Fisheries and Aquatic Sciences, 47(6), 1128-1135. https://doi.org/10.1139/f90-128
[10] McCarthy, W. F. (2007). The Existence of Maximum Likelihood Estimates for the Binary Response Logistic Regression Model. https://scholarscompass.vcu.edu/etd/1545/
[11] Meyer, K. (2016). Simple penalties on maximum-likelihood estimates of genetic parameters to reduce sampling variation. Genetics, 203(4), 1885-1900. https://doi.org/10.1534/genetics.115.185207
[12] Miller, J. M., & Miller, M. D. (2011). Handling quasi-nonconvergence in logistic regression: technical details and an applied example. Interstat, 15(11), 22.
[13] Mohammed, R. F. (2022). Estimating the Parameters of the Poisson Regression Model Under the Multicollinearity Problem. Discovery Summit Americas 2022, 30mp-1148.
[14] Muya, K. B. (2018). Application of Ordinal Logistic Regression in Analyzing Students' Performance at Kenya Certificate of Secondary Education Level in Kiambu County (Doctoral dissertation, University of Nairobi).
[15] Ogoke, U. P., Nduka, E. C., & Nja, M. E. (2013). A new logistic ridge regression estimator using exponentiated response function. Journal of Statistical and Econometric Methods, 2(4), 161-171.
[16] Pearl, R., & Reed, L. J. (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation. Proceedings of the National Academy of Sciences, 6(6), 275-288. . https://doi.org/10.1073/pnas.6.6.275
[17] Shahmandi, M., Farmanesh, F., Gharahbeigi, M. M., & Shahmandi, L. (2013). Data Analyzing by Attention to Weighted Multicollinearity in Logistic Regression Applicable in Industrial Data. British Journal of Applied Science & Technology, 3(4), 748-763. https://doi.org/10.9734/BJAST/2013/3663
[18] Shen, J., & Gao, S. (2008). A solution to separation and multicollinearity in multiple logistic regression. Journal of Data Science: JDS, 6(4), 515. http://jds-online.org/journal/JDS-515/
[19] Urgan, N. N., & Tez, M. (2008). Liu estimator in logistic regression when the data are collinear. In 20th euro mini conference (pp. 323-327).
[20] Won, J. H., Pany, T., & Eissfeller, B. (2012). Iterative maximum likelihood estimators for high-dynamic GNSS signal tracking. IEEE Transactions on Aerospace and Electronic Systems, 48(4), 2875-2893. https://doi.org/10.1109/TAES.2012.6324707

المصادر

[1] عبدالوهاب، م. م.، أبونزل، م. ر.، حماد، أ. ت.، والمسري، أ. م. (2024). مُقدِّر ليو المعدل ذو المعلمتين لمعالجة مشكلة تعدد الخطية في نموذج انحدار بواسون. Axioms، 13(1)، 46. https://doi.org/10.3390/axioms13010046
[2] ألبرت، أ.، وأندرسون، ج. أ. (1984). حول وجود تقديرات الاحتمالية القصوى في نماذج الانحدار اللوجستي. Biometrika، 71(1)، 1-10. https://doi.org/10.1093/biomet/71.1.1

[3] ألبرت، أ.، وليسافر، إ. (1986). التمييز اللوجستي متعدد المجموعات. في Statistical Methods of Discrimination and Classification (ص. 209-224). بيرغامون. https://doi.org/10.1016/B978-0-08-033621-3.50020-0

[4] أليسون، ب. د. (2008، مارس). إخفاقات التقارب في الانحدار اللوجستي. في SAS global forum (المجلد 360، العدد 1، ص. 11). https://doi.org/10.3103/S1066530724700017

[5] بي، ي.، وجيسك، د. ر. (2010). كفاءة الانحدار اللوجستي مقارنة بالتحليل التمييزي الطبيعي في ظل ضجيج التصنيف المشروط بالصنف. Journal of Multivariate Analysis، 101(7)، 1622-1637. https://doi.org/10.1016/j.jmva.2010.03.003

[6] إيفرون، ب. (1975). كفاءة الانحدار اللوجستي مقارنة بالتحليل التمييزي الطبيعي. Journal of the American Statistical Association، 70(352)، 892-898. https://doi.org/10.1080/01621459.1975.10480305

[7] غيماريش، إ. أ.، وبوبنياك، ت. س. (2021). البرمجة الخطية المطبقة على الكشف عن الانفصال في الانحدار اللوجستي متعدد الحدود. . https://doi.org/10.26650/acin.1558583

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics، 8(1).

[8] لقمان، أ. ف.، أديوي، إ.، مانسون، ك.، وكيبريا، ب. ج. (2021). مُقدِّر جديد لنموذج انحدار بواسون متعدد الخطية: محاكاة وتطبيق. Scientific Reports، 11(1)، 3732. https://doi.org/10.1038/s41598-021-82424-x

[9] ماركس، ب. د.، وسميث، إ. ب. (1990). تعدد الخطية الموزون في الانحدار اللوجستي: تشخيصات وتقنيات التقدير المتحيز مع مثال من تحمض البحيرة. Canadian Journal of Fisheries and Aquatic Sciences، 47(6)، 1128-1135. https://doi.org/10.1139/f90-128

[10] مكارثي، و. ف. (2007). وجود تقديرات الاحتمالية القصوى لنموذج الانحدار اللوجستي للاستجابة الثنائية. https://scholarscompass.vcu.edu/etd/1545/

[11] ماير، ك. (2016). عقوبات بسيطة على تقديرات الاحتمالية القصوى للمعلمات الوراثية لتقليل التباين في أخذ العينات. Genetics، 203(4)، 1885-1900. https://doi.org/10.1534/genetics.115.185207

[12] ميلر، ج. م.، وميلر، م. د. (2011). التعامل مع التقارب شبه غير التام في الانحدار اللوجستي: تفاصيل فنية ومثال تطبيقي. Interstat، 15(11)، 22.

[13] محمد، ر. ف. (2022). تقدير معلمات نموذج انحدار بواسون تحت مشكلة تعدد الخطية. Discovery Summit Americas 2022، mp-114830.

[14] مويا، ك. ب. (2018). تطبيق الانحدار اللوجستي الترتيبي في تحليل أداء الطلاب في مستوى شهادة كينيا للتعليم الثانوي في مقاطعة كيامبو (أطروحة دكتوراه، جامعة نيروبي).

[15] أوغوك، و. ب.، ندوكا، إ. س.، ونجا، م. إ. (2013). مقدِّر انحدار التلال اللوجستي الجديد باستخدام دالة الاستجابة الأسية. Journal of Statistical and Econometric Methods، 2(4) ، 161-171.

[16] بيرل، ر.، وريد، ل. ج. (1920). حول معدل نمو سكان الولايات المتحدة منذ عام 1790 وتمثيلهم الرياضي. Proceedings of the National Academy of Sciences، 6(6) ، 275-288. https://doi.org/10.1073/pnas.6.6.275

[17] شاهمنذي، م.، فرمانش، ف.، قره بيجي، م. م.، وشاهمنذي، ل. (2013). تحليل البيانات مع الانتباه إلى تعدد الخطية الموزون في الانحدار اللوجستي القابل للتطبيق في البيانات الصناعية. British Journal of Applied Science & Technology، 3(4)، 748-763 . https://doi.org/10.9734/BJAST/2013/3663

[18] شين، ج.، وقاو، س. (2008). حل لمشكلتي الانفصال وتعدد الخطية في الانحدار اللوجستي المتعدد. Journal of Data Science: JDS، 6(4) ، 515. http://jds-online.org/journal/JDS-515/

[19] أورغان، ن. ن.، وتيز، م. (2008). مُقدِّر ليو في الانحدار اللوجستي عندما تكون البيانات متحدة الخطية. في 20th euro mini conference (ص. 323-327).

[20] وون، ج. هـ.، باني، ت.، وإيسفيلر، ب. (2012). مقدِّرات الاحتمالية القصوى التكرارية لتتبع إشارة GNSS عالية الديناميكية. IEEE Transactions on Aerospace and Electronic Systems، 48(4)، 2875-2893. https://doi.org/10.1109/TAES.2012.6324707

https://doi.org/10.31272/jae.i148.1426
https://admics.uomustansiriyah.edu.iq
P-ISSN: 1813–6729 E-ISSN: 2707-1359

Estimating Logistic Regression Parameters Using the Adjusted Estimator Method

Hasanain Jalil Neamah
Dept. of Business Information Technology, Business Informatics College, University of Information Technology and Communications, Baghdad, Iraq.
Email: hasanien.1975@uoitc.edu.iq , ORCID: https://orcid.org/0009-0009-8128-3068
Firas Monthetr Jassim
Dept. of Statistics, College of Administration & Economics, Mustansiriyah University, Baghdad, Iraq.
Email: firasm@uomustansiriyah.edu.iq, ORCID: https://orcid.org/0000-0002-4948-8583

Article Information

Article History:
Received: 28 / 4 / 2025
Accepted: 26 / 5 / 2025
Available Online: 01 / 06 / 2025
Page no: 88 – 98

Keywords:
Adjusted Estimators, Multi-collinearity, Data Separation, Logistic Function, Weighted variance.

Correspondence:
Researcher name:
Firas Monthetr Jassim

Email:
firasm@uomustansiriyah.edu.iq

Abstract

The problem of separating the binary dependent variable observations that depend on the sample size and the problem of multicollinearity among the explanatory variables are considered two of the most critical issues that arise in the logistic regression model. Treating these problems requires applying effective methods, where the estimation methods of the Double Penalty Maximum Likelihood Estimators (DPMLE) and the Adjusted Estimator Method were adopted in this research. After diagnosing the problem of separation and multicollinearity in the real data, represented by anaemia, which were obtained from the blood disease laboratories at Medicine-City Hospital, several key findings were reached. Chief among them was that the Modified Estimators Method was the best in terms of treating separation and multicollinearity problems. This was reached via dependence on the mean square error (MSE) as a comparison criterion.