



AL KUT JOURNAL OF ECONOMIC AND ADMINISTRATIVE SCIENCES

Publisher: College of Economics and Management - Wasit University



تشخيص ومعالجة مشكلتي التعدد الخطي والفصل في انموذج الانحدار اللوجستي المتعدد للمرضى المصابين بفقر الدم

Diagnosis and Treatment of the Problems Multicollinearity and Separation in the Logistic Regression Model For Patients with Anemic

م.م. احمد رزاق عبد النصير اوي

قسم الإحصاء/ كلية الإدارة والاقتصاد

جامعة واسط

ahmedrazzaq@uowasit.edu.iq

المستخلص

اهم المشاكل التي تظهر في انموذج الانحدار اللوجستي المتعدد هي مشكلة الفصل بين مشاهدات المتغير التابع ثنائي الاستجابة الذي يعتمد على احجام العينات، ومشكلة التعدد الخطي بين المتغيرات التوضيحية.

اذ تم تطبيق بيانات تمثل الإصابة بفقر الدم والتي تم الحصول عليها من مستشفى الزهراء التعليمي قسم الكلى الصناعية من خلال طرائق التقدير وفقاً لطرائق مقدرات الأماكن الأعظم التكرارية ومقدرات الانحدار اللوجستي شتاين التي تخص معالجة مشكلة التعدد الخطي ومقدرات الإمكان الأعظم الجزائية التي تخص معالجة مشكلة الفصل والمقدرات المعدلة التي تخص معالجة مشكلة الفصل ومشكلة التعدد الخطي والتي تمثل أفضل طرائق التقدير لأنها تمتلك اقل متوسط مربعات الخطأ (MSE) لأنموذج الانحدار اللوجستي المتعدد.

Abstract

This research dealt with the subject of study logistic regression model, which is one of nonlinear models Taking character more advanced in the process of statistical analysis, which aims to get the high-level estimates of efficiency.

Of the most important problems that appear in this model is the separation between the observations of the dependent variable binary response and Multicollinearity between the explanatory variables.

As it has been a real data represented injury anemic which have been obtained from kut hospital artificial kidney department through estimation methods according to the modalities Iterative Maximum Likelihood estimators and Stein Logistic Regression Estimators concerning the treatment Multicollinearity problem and penalized maximum likelihood estimators concerning the treatment separation problem and Adjusted Estimators concerning the treatment of separation problem and Multicollinearity problem and that represent the best estimation methods because it has the mean square error (mse) for the logistic regression model.

Keywords: Logistic Regression Model, Problems Multicollinearity, Separation, Iterative Maximum Likelihood Estimators, Penalized Maximum Likelihood Estimators

Introduction

2.1 المقدمة: [5][13]

أن نموذج الانحدار اللوجستي المتعدد من نماذج الانحدار اللاخطية يوضح العلاقة بين المتغير التابع ثنائي الاستجابة والمتغيرات التوضيحية المستقلة, إذ استعمل العديد من الباحثين انموذج الانحدار اللوجستي المتعدد ويعدّ أول من استخدم دالة اللوجستيك (Logistic Function) الباحث (Verhulst) لوصف نمو المجتمع وكانت تسمى هذه الدالة بدالة النمو (Growth Function), ولقد قام الباحثان (Pearl And Reed) في عام 1920 باستخدام الدالة لحساب نمو السكان وأطلق عليها فيما بعد بدالة اللوجستيك, وتبرز العديد من استخدامات هذا الأنموذج في الدراسات المتعلقة بعلوم الحياة وكذلك العلوم الزراعية والطبية وبشكل عام في الدراسات ذات الطابع التجريبي, لكونه من النماذج الملائمة للبيانات الثنائية (Binary Data).^[1] لذلك يعرف الانحدار بشكل عام بأنه التحليل الذي يختص بدراسة اعتماد متغير واحد يعرف بالمتغير التابع (متغير الاستجابة) على متغير واحد أو أكثر تعرف بالمتغيرات التوضيحية (التفسيرية), وذلك لغرض التقدير والتنبؤ بقيمة المتغير التابع باعتماد معلومات المتغيرات التوضيحية (التفسيرية).^[2]

فقد بدأت دراسة بعض المشكلات في انموذج الانحدار اللوجستي المتعدد مثل مشكلة التعدد الخطي (multicollinearity) بين المتغيرات التوضيحية ومشكلة الفصل (Separation) بين مشاهدات المتغير التابع ثنائي الاستجابة، إذ تعتمد مشكلة التعدد الخطي على وجود علاقة خطية بين بعض او كل المتغيرات التوضيحية أي وجود ارتباط بين هذه المتغيرات فيصبح انموذج الانحدار اللوجستي المتعدد غير مستقر.^[3]

اما مشكلة الفصل (separation) تعتمد على شكل انتشار البيانات أي عندما تكون بيانات المتغير التابع ثنائية الاستجابة وعلى حجم العينة أي عندما تكون العينات 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.[1]

إن طرائق التقدير التقليدية لمعاملات انحدار اللوجستي المتعدد عند تحليل البيانات ثنائية الاستجابة (Binary data response) ضعيفة في معالجة المشكلات بين البيانات الامر الذي يستوجب استخدام طرائق لمعالجة تلك المشاكل.

Problem of research

3.1 مشكلة البحث:

من المشاكل التي تظهر في انموذج الانحدار اللوجستي المتعدد هي مشكلة الفصل بين بيانات المتغير التابع ثنائية الاستجابة ومشكلة التعدد الخطي بين المتغيرات التوضيحية، مشكلة الفصل في بيانات المتغير التابع ثنائية الاستجابة تعتمد على حجم العينة أي عندما يكون حجم العينة صغير وعدد مجاميع الاستجابة للمتغير التابع. ان وجود العلاقة القوية بين المتغيرات التوضيحية في انموذج الانحدار اللوجستي المتعدد يؤدي الى ظهور مشكلة التعدد الخطي التي تجعل الانموذج غير مستقر ومعلمات التقدير غير دقيقة،

Object of research

4.1 هدف البحث:

يهدف هذا البحث الى تشخيص وجود مشكلتي الفصل والتعدد الخطي في بيانات التطبيق ثنائية الاستجابة وتطبيق طرائق تقدير معاملات انموذج الانحدار اللوجستي المتعدد في حالة وجود المشكلات، 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. مقارنة طرائق تقدير معاملات انموذج الانحدار اللوجستي المتعدد في حالة وجود مشكلة الفصل والتعدد الخطي من خلال معيار متوسط مربعات الخطأ MSE للأنموذج.

المبحث الثاني: الجانب النظري

Introduction

1.1 المقدمة:

في هذا المبحث سيتم دراسة انموذج الانحدار اللوجستي المتعدد والصيغ العامة والافتراضات الخاصة به، ودراسة المشاكل التي يتعرض لها هذا الانموذج المتمثلة بالتعدد الخطي من خلال نشأتها وتأثيراتها وطرائق تشخيصها وكذلك مشكلة الفصل التي تم تصنيفها الى مشكلة الفصل التام والفصل شبه التام والتداخل، واخيراً سيتم تقدير معاملات انموذج الانحدار اللوجستي المتعدد في ظل وجود مشكلة التعدد الخطي والفصل.

Logistic regression model

4.2 أنموذج الانحدار اللوجستي: [4][3][11][18]

يعرف انموذج الانحدار اللوجستي المتعدد على انه احد نماذج الانحدار الذي تكون فيه العلاقة بين المتغير التابع y والمتغيرات التوضيحية (x_1, x_2, \dots, x_p) غير خطية حيث يكون المتغير التابع y ثنائي الاستجابة او اكثر مفترضاً إحدى القيمتين (1,0) أما النجاح success حدوث الاستجابة باحتمال π_1 أو الفشل failure

عدم حدوث الاستجابة باحتمال $1 - \pi_i$ لذلك يكون المتغير التابع y يتوزع توزيع برنولي وسوف تكون دالة الكثافة الاحتمالية بالصيغة الآتية. [1][2][12]

$$p(Y = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \quad \dots(1)$$

$$y_i = 0,1$$

اذ إن

y_i متغير تابع ثنائي الاستجابة.

π_i احتمال حدوث الاستجابة عندما $y_i = 1$.

لذلك فان توقع المتغير التابع يمثل احتمال حدوث الاستجابة بالصيغة الآتية.

$$E(y_i) = p(Y = 1) = \pi_i \quad \dots(2)$$

أما تباين المتغير التابع حسب توزيع برنولي هو.

$$V(y_i) = \pi_i(1 - \pi_i) \quad \dots(3)$$

ليكن X_1, X_2, \dots, X_k مجموعة من المتغيرات التوضيحية ولتكن n تمثل عدد المشاهدات لهذه المتغيرات التي تكون المصفوفة X .

$$X = (x_{ij})_{n \times k} \quad \dots(4)$$

اذ إن:

$i = 1, 2, \dots, n$ تمثل حجم العينة.

$j = 1, 2, \dots, k$ عدد المتغيرات المستقلة ، $p = (k+1)$ يمثل عدد المعلمات.

فاذا كان $y_i = [y_1, y_2, \dots, y_n]$ عينه عشوائية من المتغير ثنائي الاستجابة وأن $y_i \in \{0,1\}$.

وبالتالي فأن انموذج الانحدار اللوجستي المتعدد يكتب بالصيغة الآتية.

$$y_i = \pi_i + \varepsilon_i \quad \dots(5)$$

اذ إن π_i تمثل دالة الانحدار اللوجستي (احتمال الاستجابة).

$$\pi_i = p(y = 1) = \frac{e^{x_i \theta}}{1 + e^{x_i \theta}} \quad \dots(6)$$

θ : متجه من المعلمات أبعاده $(p \times 1)$.

$\underline{x}_i = \{x_{i1}, x_{i2}, \dots, x_{ip}\}$: متجه صفي من المتغيرات التوضيحية أبعاده $(1 \times p)$.

ε_i : يمثل حد الخطأ العشوائي بمتوسط صفر وتباين $\pi_i(1 - \pi_i)$ ومن الملاحظ 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. [18][19]

The problem of Multicollinearity

3.2 مشكلة التعدد الخطي: [6][8]

تحصل مشكلة التعدد الخطي عندما تكون هناك علاقة خطية بين بعض أو كل المتغيرات التوضيحية وأن الارتباط بين هذه المتغيرات يعرف بالتعدد الخطي، أي تظهر مشكلة التعدد الخطي عندما تكون قيمة احد المتغيرات التوضيحية متساوية لجميع المشاهدات، أو عند اعتماد قيمة احد المتغيرات التوضيحية على قيمة واحدة أو أكثر من المتغيرات التوضيحية في الأنموذج أن احد الشروط الواجب توفرها في انموذج الانحدار عند عدم وجود مشكلة التعدد الخطي بصورة عامة هو شرط الرتبة (rank condition). [6]

$$\text{rank}(X) = k + 1 < n \quad \dots(7)$$

اذ أن X مصفوفة المتغيرات التوضيحية أبعادها $(n \times (k+1))$ وعليه عندما تكون المتغيرات التوضيحية (explanatory variables) مستقلة خطيا يمكن إيجاد معكوس المصفوفة (XX) وبالتالي يمكن إيجاد تقديرات المعلمات , أما إذا كان هناك علاقة خطية بين اثنين أو أكثر من المتغيرات فان ذلك سيؤدي الى انتهاك شرط الرتبة. [6]

$$\text{rank}(X) < k + 1 < n \quad \dots(8)$$

لذا لا يمكن إيجاد معكوس مصفوفة المعلومات (XX) وبالتالي لا توجد مقدرات المعلمات.

اما سبب مشكلة عدم تجانس التباين في انموذج الانحدار اللوجستي المتعددان حد الخطأ العشوائي يكون له متوسط صفر وتباين $\pi_i(1 - \pi_i)$ ومن الملاحظ ان تباين حد الخطأ يعتمد على دالة الانحدار اللوجستي π_i عند كل مشاهدة من i الامر الذي يؤدي الى عدم تجانس تباين حد الخطأ العشوائي عند كل مستوى من المتجه \underline{x}_i . [7]

إن انموذج الانحدار اللوجستي المتعدد يصبح غير مستقر عند وجود الاعتماد القوي بين المتغيرات التوضيحية $(X_1, X_2 \dots X_k)$ لذلك يتم الاعتماد على جميع المتغيرات التوضيحية في الأنموذج لعملية تشخيص مشكلة التعدد الخطي , ومن اجل الحصول على التشخيص المناسب لمشكلة التعدد الخطي الموزون يفضل استعمال مصفوفة المعلومات الموزونة.

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

ويمكن التعبير عن مصفوفة المعلومات الموزونة المقدره بالصيغة الاتية.

$$\widehat{\Phi} = \widehat{X}\widehat{W}\widehat{X} \dots(9)$$

او تكتب بالصيغة الاتية.

$$\widehat{\Phi} = \widehat{S}\widehat{S}$$

$$\widehat{S} = \widehat{W}^{0.5}\widehat{X}$$

وتوجد طريقتان لتشخيص مشكلة التعدد الخطي الموزون في انموذج الانحدار اللوجستي المتعدد وهي كالآتي:

اولاً: قيم اعداد الشرط الموزون تعتمد على مصفوفة الارتباطات الموزونة المقدره $\widehat{\Phi}^*$ في تشخيص مشكلة التعدد الخطي الموزون فان قيم اعداد الشرط التي تكون اكبر من 30 ($k_j > 30$) تشير إلى حالة التشخيص أي وجود مشكلة التعدد الخطي الموزون بين المتغيرات التوضيحية في أنموذج الانحدار اللوجستي.

لتكن $\lambda_0^*, \lambda_1^*, \dots, \dots, \lambda_p^*$ القيم الذاتية المرتبة (الجذور المميزة) الى مصفوفة الارتباطات الموزونة المقدره $\widehat{\Phi}^*$ التي يعبر عنها من خلال الخطوات الاتية:

الوسط الحسابي \bar{s}_j يكون بالصيغة الاتية:

$$\bar{s}_j = \frac{\sum_{i=1}^n \hat{s}_{ij}}{n} \dots(10)$$

\hat{s}_{ij} : يمثل العنصر بالموقع (i,j) من المصفوفة \widehat{S}

\widehat{S}^*_{ij} يمثل العنصر بالموقع (i,j) من المصفوفة \widehat{S}^* والذي يحسب بالصيغة الأتية.

$$\widehat{S}^*_{ij} = \frac{\hat{s}_{ij} - \bar{s}_j}{\sqrt{\sum_{i=1}^n (\hat{s}_{ij} - \bar{s}_j)^2}} \dots(11)$$

مصفوفة الارتباطات الموزونة المقدره $\widehat{\Phi}^*$ يعبر عنها بالصيغة الاتية:

$$\widehat{\Phi}^* = \widehat{S}^*\widehat{S}^* \dots(12)$$

اعداد الشرط تعرف كالآتي.

$$k_j = \left(\frac{\lambda^*_{\max}}{\lambda^*_j}\right)^{0.5} \dots(13)$$

اذ إن λ^*_{\max} هو اكبر الجذور المميزة.

λ^*_j هو الجذر المميزة j في المصفوفة $\hat{\Phi}^*$.

ثانيا: تشخيص مشكلة التعدد الخطي في هذه الحالة يعتمد على قيم نسبة التباين الموزونة والتي تقع بين الصفر والواحد الصحيح، لذا ففي حالة وجود قيمتين من قيم نسبة التباين الموزونة قريبتان من الواحد الصحيح تكون هناك مشكلة التعدد الخطي الموزون بين متغيرات هذه النسب.

لتكن M مصفوفة المتجهات المتعامدة التي يتم الحصول من مصفوفة الارتباطات الموزونة المقدر $\hat{\Phi}^*$, وان Λ^* مصفوفة قطرية من الجذور المميزة لمصفوفة الارتباطات الموزونة المقدر $\hat{\Phi}^*$ التي تحقق الشرط الآتي.

$$M\hat{\Phi}^*M = \Lambda^* \dots(14)$$

لذلك فان أي مشاهدة من نسبة التباين الموزونة وليكن ω_{uj} يمكن التعبير عنها كالآتي.

$$\omega_{uj} = \frac{m_{ju}^2/\lambda_u^*}{c_{jj}} \dots(15)$$

m_{ju} عنصر من مصفوفة المتجهات المميزة M من الرتبة $(j \times u)$.

λ_u^* الجذر المميز المحسوب من مصفوفة المعلومات المقدر $\hat{S}^* \hat{S}^* = \hat{\Phi}^*$.

c_{jj} الجذور المميزة الصغيرة نسبة الى اكبر جذر مميز ويمكن أن يحسب بالصيغة الآتية.

$$c_{jj} = \sum_{u=1}^p \lambda_u^{*-1} m_{ju}^2 \dots(16)$$

اذ إن :

$u=1,2,\dots,p$ عدد الجذور المميز.

4.2 فصل البيانات في الانحدار اللوجستي: [6][9][10][13] Separation of the data in logistic regression

مشكلة الفصل تعتمد على كل من والانموذج ونوع البيانات أي أن مشكلة الفصل تحدث بالمقام الأول مع العينات الصغيرة ويمكن أن تتحقق من خلال زيادة عدد المتغيرات التوضيحية الواردة في الأنموذج وفي انموذج الانحدار اللوجستي المتعدد تظهر مشكلة الفصل بشكل واسع عندما تكون هناك g من مجاميع الاستجابة response categories أي $\{s=1,2,\dots,g\}$. [9]

اظهر ألبرت وأندرسون (1984) أن مجموعة نقاط العينة n يمكن تصنيفها في واحد من ثلاثة تكوينات متنافية للمتغير التابع (الفصل التام، الفصل شبه التام، التداخل) لذا نستعمل مصطلح الفصل لوصف مجموعة من نقاط العينة التي تنتمي الى تكوين الفصل التام أو الفصل شبه التام، عندما يكون هناك فصل تام أو فصل شبه تام بين مشاهدات العينة للبيانات فان مقدرات الإمكان الأعظم تكون غير دقيقة في حدود فضاء المعلمات، وتكون مقدرات الإمكان الأعظم ودقيقة عندما يكون هناك تداخل بين نقاط العينة. [14]

Complete Separation

1.4.2 الفصل التام: [5][6][12]

الفصل التام يحدث بين مجموعة من مشاهدات العينة n أي إذا كان لدينا متجه من المعلمات θ بشرط $\underline{x}_i\theta > 0$ عندما $y_i = 1$ وان $\underline{x}_i\theta < 0$ عندما $y_i = 0$ كما مبين في الجدول رقم (1).

الجدول رقم (1) يستعرض الفصل التام للبيانات [12]

x	Y
-5	0
-4	0
-3	0
-2	0
-1	0
1	1
2	1
3	1
4	1
5	1

quasicomplete separation**2.4.2 الفصل شبه التام:**[8][9][11]

الفصل شبه التام يحدث اذا كان لدينا متجه من المعلمات θ بحيث $\underline{x}_i\theta \geq 0$ عندما $y=1$ و $\underline{x}_i\theta \leq 0$ عندما $y=0$ وبأخذ حالة مساواة واحدة على الأقل في كل فئة من فئات المتغير التابع يحدث الفصل شبه التام كما مبين في الجدول رقم (2).

الجدول رقم (2) يستعرض الفصل شبه التام للبيانات [12]

x	Y
-6	0
-5	0
-4	0
-2	0
-1	0
0	0
0	1
1	1
2	1
3	1
4	1
6	1

أن في حالة وجود الفصل التام والفصل شبه التام للبيانات يمكن ملاحظة ان مقدرات الإمكان الأعظم التكرارية تكون غير دقيقة في حالة الفصل التام والفصل شبه التام. وفي حالة الفصل التام وشبه التام يكون عدد التكرارات كبير في استخراج مقدرات الإمكان الأعظم التكرارية. كذلك يتم تشخيص مشكلة الفصل التام ومشكلة الفصل شبه التام من خلال الاعتماد على مقدرات الإمكان الأعظم التكرارية $\hat{\theta}$ في حالة وجود مشكلة الفصل التام ومشكلة الفصل شبه التام عند تقدير المعلمات يستمر التكرار حتى يتم تجاوز التكرار الثابت لذلك سوف تكون مقدرات المعلمات غير دقيقة وحد الخطأ كبير للغاية , اقترح ألبرت وأندرسون عام 1984 وسيلة تجريبية للكشف عن مشكلة الفصل التي كان تنفيذها في PROTCL LOGISTIC والذي يحتوي على الخطوات الآتية. [8] [11]

- 1- إذا كان معيار التقارب يتحقق في ثمانية تكرارات نستنتج إلى أن هناك لا توجد مشكلة الفصل بصورة عامة اي هناك تداخل بين البيانات.
- 2- لجميع التكرارات بعدد التكرار الثامن احتمال توقع الاستجابة (دالة الانحدار اللوجستي) لكل مشاهدة يمكن حسابة من قبل الصيغة الآتية.

$$\hat{\pi}_i = \frac{1}{1+\exp [(2y_i-1)\underline{x}_i\hat{\theta}]} \dots(17)$$

* إذا كان احتمال حصول الاستجابة هو واحد لجميع مشاهدات دالة الانحدار اللوجستي نستنتج أن هناك فصلا تاماً وفي هذه يتوقف التكرار.

* إذا كان احتمال حصول الاستجابة الى مشاهدات دالة الانحدار اللوجستي أكبر من (0.95) ، لبعض المشاهدات وليس لكلها نستنتج الى أن هناك فصل شبه تام ووقف التكرار.

5.2 طرائق تقدير معلمات أنموذج الانحدار اللوجست [3][9][10][12][13][14][16][17][18]

Methods estimation parameters logistic regression model

1.5.2 مقدرات الإمكان الأعظم التكرارية: [1] [10] [11] [12] Iterative Maximum Likelihood Estimators

مقدرات الإمكان الأعظم هي اي المعلمات التي تجعل دالة الإمكان أعظم ما يمكن، دالة الإمكان الأعظم لأنموذج الانحدار اللوجستي المتعدد الذي يتبع توزيع برنولي تكون بالصيغة الآتية.

$$L(\theta, X) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \dots(18)$$

وحسب خاصية التحويل لدالة اللوجت

$$\log \frac{\pi_i}{1-\pi_i} = \underline{x}_i \theta \dots(19)$$

$$L(\theta, X) = \prod_{i=1}^n (1 - \pi_i) \exp [\sum_{i=1}^n y_i (\underline{x}_i \theta)]$$

وبأخذ اللوغاريتم إلى دالة الإمكان.

$$\log L(\theta, X) = [\sum_{i=1}^n y_i (\underline{x}_i \theta)] - \sum_{i=1}^n \log (1 + e^{\underline{x}_i \theta})$$

أن المعادلات الناتجة من المشتقة الأولى ليس لها حل، في مثل هذه الحالات لابد من حل المعادلات عن طريق الطرائق العددية والتي منها الطريقة العددية الأكثر شيوعا هي خوارزمية نيوتن رافسون التي تعطى بالصيغة الآتية. [12]

$$\underline{\theta}_{imle}^{(t+1)} = \underline{\theta}_{imle}^t - I(\theta)^{-1(t)}U(\theta)^{(t)} \dots(20)$$

عندما $U(\theta)$ يمثل المشتقة الأولى الى لوغاريتم دالة الإمكان الأعظم.

$$U(\theta) = \frac{\partial \log L(\theta, X)}{\partial \underline{\theta}} = \sum_{i=1}^n \underline{x}_i (y_i - \pi_i) \dots(21)$$

ويمكن كتابة $U(\theta)$ بشكل متجه عمودي كالآتي .

$$U(\theta) = \underline{X}(y_i - \pi_i)$$

وان $I(\theta)^{-1}$ معكوس مصفوفة تقييم المعلومات إلى θ يمكن الحصول على $I(\theta)$ من المشتقة الثانية إلى لوغاريتم دالة الإمكان الأعظم لأنموذج الانحدار اللوجستي.

$$I(\theta) = \frac{\partial^2 \log L(\theta, X)}{\partial \underline{\theta} \partial \underline{\theta}} = - \sum_{i=1}^n \underline{x}_i \underline{x}_i' \pi_i (1 - \pi_i) \dots(22)$$

وبذلك تكون مصفوفة المعلومات بالصيغة الآتية.

$$I(\theta) = -\underline{X} \text{diag} \pi_i (1 - \pi_i) \underline{X}' = -\underline{X} W \underline{X}'$$

t: يمثل عدد التكرارات.

$\underline{\theta}_{imle}^t$: متجه عمودي من المعلمات المقدرة أبعاده $(p \times 1)$ في التكرار s حيث تكون في بدايت التكرار تساوي صفر.

X: مصفوفة المتغيرات التوضيحية أبعاده $(n \times p)$.

2.5.2 مقدرات الانحدار اللوجستي شتاين: [15][17][18] Stein Logistic Regression Estimators

شتاين 1960 اقترح مقدرات يتم من خلالها معالجة مشكلة التعدد الخطي في الانحدار اللوجستي حيث يتم تقليص مقدرات الإمكان الأعظم التكرارية على النحو الآتي.

$$\hat{\underline{\theta}}_{slre} = C \hat{\underline{\theta}}_{imle} \dots(23)$$

حيث تكون قيمة C تقع بين الصفر والواحد الصحيح $0 < C < 1$ والغرض من اسلوب شتاين هو تقليص كلا من متجه المعلمات المقدرة والأخطاء المعيارية بواسطة تقنية القياس البسيط C التي يتم اختيارها مما يقلل من دالة الخسارة التي تحسب بالصيغة الآتية:

$$E(L^2) = (C \hat{\underline{\theta}}_s - \underline{\theta})' (C \hat{\underline{\theta}}_s - \underline{\theta}) \dots(24)$$

لذلك سوف يكون المعيار المتعلق ب C كالآتي:

$$C = \frac{\hat{\theta}_{imle}\hat{\theta}_{imle}}{\{\hat{\theta}_{imle}\hat{\theta}_{imle} + \text{trace}(\hat{\Phi}^{*-1})\}} \dots(25)$$

عندما $\hat{\theta}_{imle}$ مقدرات الإمكان الأعظم التكرارية.

$\text{trace}(\hat{\Phi}^{*-1})$ مجموع عناصر القطر الرئيسي لمعكوس مصفوفة المعلومات $\hat{\Phi}^{**}$.

الخطأ المعياري لمقدرات شتاين يمثل C من التكرارات للخطأ المعياري لمقدرات الإمكان الأعظم التكرارية وان خواص طريقة شتاين مشابه إلى طريقة الحرف.

Penalized Estimators

3.5.2 المقدرات الجزائية: [18][14][12]

اقترحت مقدرات الجزائية من قبل (Firth 1993) [15] إجراء (Firth) وضع أصلا للحد من التحيز في مقدرات الإمكان الأعظم التكرارية لتوفير الحل المثالي الى مشكلة الفصل، تستخدم للسيطرة على مشكلة الفصل التام وشبه التام.

$$L(\theta, X)^* = L(\theta, X)|I(\theta)|^{1/2} \dots(26)$$

من اجل الحد من تحيز العينات في تقديرات (Firth) اقترح إسناد التقديرات على معادلة النتيجة المعدلة التي تمثل المشتقة الأولى الى لوغار يتم دالة الإمكان الجزائية والتي تكون بالصيغة الآتية.

$$U(\theta^*) = U(\theta) + \frac{1}{2} \text{trace} \left[I(\theta)^{-1} \left\{ \frac{\partial I(\theta)}{\partial \theta} \right\} \right] \dots(27)$$

Firth اظهر أن تحيز مقدرات الإمكان الأعظم تم إزالته , أن غرض صيغة الانحدار اللوجستي الجزائية (المقطعية) الحد من التحيز مشكلة الفصل , أن الفكر العامة إلى firth هو استبدال معادلة النتيجة = $U(\theta) \sum_{i=1}^n \underline{x}_i (y_i - \pi_i)$ بمعادلة النتيجة المعدلة التي تكتب بالصيغة الآتية:

$$U(\theta^*) = \sum_{i=1}^n [(y_i - \pi_i) + h_i(\frac{1}{2} - \pi_i)] \underline{x}_i \dots(28)$$

عندما h_i عناصر القطر الرئيسي للمصفوفة H

$$H = W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}} \dots(29)$$

حيث أن w تمثل المصفوفة $w = \text{diag}\{\pi_i(1 - \pi_i)\}$ وبصيغة المصفوفات يمكن ان تكتب معادلة النتيجة المعدلة بالصيغة لتالية .

$$U(\theta^*) = \hat{X}(y_i - \pi_i) + \hat{X}B\left(\frac{1}{2} - \pi_i\right) \dots(30)$$

B تمثل المصفوفة القطرية تكتب بالشكل التالي $B = \text{diag}(B)$

مقدرات $\text{firt}B$ يمكن الحصول عليها تكراريا من خلال طريقة نيوتن رافسون.

$$\underline{\theta}_{\text{pmle}}^{(t+1)} = \underline{\theta}_{\text{pmle}}^t - I(\theta)^{-1(t)}U(\theta^*)^{(t)} \dots(31)$$

أن t عدد التكرارات ويتم التوقف عن عدد التكرارات حتى يكون الفرق بين عمليتين تكراريتين هو صفر أو قريب من الصفر وبذلك تكون طريقة $\text{Firt}B$ قد عالجت مشكلة الفصل.

Adjusted Estimators

4.5.2 المقدرات المعدلة:

في هذه الطريقة يتم الجمع بين خواص تقديرات $\text{Firt}B$ في الانحدار اللوجستي التي يتم من خلالها معالجة مشكلة الفصل بين بيانات العينة المبينة في الصيغة (31) وطريقة مقدرات الانحدار اللوجستي شتاين المبينة في الصيغة (24) التي تتعامل مع مشكلة التعدد الخطي بين المتغيرات المستقلة.

يتم بناء تقديرات هذه الطريقة بتقليص تقديرات الإمكان الأعظم الجزائية بدل من تقديرات الإمكان الأعظم التكرارية من خلال قيمة C كما مبين بالصيغة الآتية.

$$\hat{\theta}_{ae} = C\hat{\theta}_{\text{pmle}} \dots(32)$$

وكما معروف ان قيمة C تقع بين الصفر

والواحد الصحيح $0 < C < 1$ لذا سوف يكون حساب قيمة C يعتمد على مقدرات الإمكان الأعظم الجزائية كما مبين في الصيغة الآتية.

$$C = \frac{\hat{\theta}_{\text{pmle}}\hat{\theta}_{\text{pmle}}}{\{\hat{\theta}_{\text{pmle}}\hat{\theta}_{\text{pmle}} + \text{trace}(\widehat{\Phi}^{*-1})\}}$$

$\hat{\theta}_{\text{pmle}}$ مقدرات الإمكان الأعظم الجزائية.

المبحث الثالث: الجانب التطبيقي

Classification of data

1.3 تصنيف البيانات:

جمعت البيانات الخاصة بالبحث من مستشفى الكوت قسم الكلى الصناعية من خلال المختبرات الخاصة بأمراض الدم، وقد تم الاستعانة بمجموعة من الأطباء المتخصصين بأمراض فقر الدم لتصنيف اهم العوامل المؤثرة على المرض، اذ تم جمع البيانات الخاصة بسنة 2015 وبعينة حجمها (40) شخص من خلال اجراء تحاليل الدم لكل شخص واخذ اهم المتغيرات الاتية.

الهيموغلوبين الدم (Bb) , حجم خلايا الدم المرصوفة (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 .

المتغير (x_1) يمثل هيموغلوبين الدم ويقاس بوحد (غم/100مل من الدم) او (غم/ديسي لتر من الدم) ، المتغير (x_2) يمثل حجم خلايا الدم المرصوفة (pcv) ويقاس بالنسبة المئوية، المتغير (x_3) يمثل خلايا الدم البيضاء (wbc) ويقاس بوحد قياس (خلية لكل الف مايكرو لتر من الدم)، المتغير (x_4) ويمثل الجنس ويضم فئتين الذكور ($x_4 = 1$) والاناث ($x_4 = 2$) .

The problem of multicollinearity

2.3 مشكلة التعدد الخطي:

لفحص عن وجود تعدد خطي بين اثنين او اكثر من المتغيرات التوضيحية قيد التطبيق استخدم البرنامج الخاص بلغة (R) في حساب مصفوفة الارتباط الموزونة $\hat{\Phi}^*$ بين المتغيرات التوضيحية و الجذور المميزة والعدد الشرطي الموزون ونسبة التباين الموزونة.

من مصفوفة الارتباط الموزونة الموضحة بالجدول (4) وجود معاملات الارتباط ذات قيم كبيرة وطرديّة الاتجاه لجميع المتغيرات التوضيحية , اذ يرتبط كل واحد منهم مع كافة المتغيرات التوضيحية الأخرى بعلاقات خطية طردية قوية كما مبين في الجدول (4) , اعداد الشرط الموزونة المبينة من المعادلة (12) تعكس نتائجها قيماً كبيره للمتغيرات الانموذج اذ كان اكبر تلك القيم الخاصة بالمتغير الأخير الذي يمثل المتغير التوضيحي الرابع حيث كانت قيمة عدد الشرط الموزون الخاص به (143.094711) وهي اكبر من القيمة 30 مما يدل على وجود مشكلة التعدد الخطي الموزون بين المتغيرات التوضيحية , ولبيان تحديد أي المتغيرات التوضيحية يسبب مشكلة التعدد الخطي الموزون يتم من خلال حساب قيم مصفوفة نسبة التباين الموزون المبينة في المعادلة (14) اذ وجد ان قيم نسبة التباين الموزون للمتغير التوضيحي الأول والثاني كبيرتان وقريبتان من الواحد ويقعان في نفس الجذر المميز الذي يقابل اكبر عدد شرط موزون كما في الجدول (5) , ومن ذلك نستنتج بأن هناك تلازم وتعدد خطي بين المتغيرات التوضيحية يسببه المتغير التوضيحي الأول والثاني بدرجة كبيرة جداً والمتغير التوضيحي الثالث والرابع بدرجة اقل.

جدول (4) يبين مصفوفة الارتباط الموزونة بين المتغيرات التوضيحية

	Intercept	X ₁	X ₂	X ₃	X ₄
Intercept	1	0.99	0.98	0.97	0.956
X ₁	0.98	1	0.97	0.97	0.94
X ₂	0.98	0.96	1	0.95	0.94
X ₃	0.97	0.96	0.95	1	0.93
X ₄	0.92	0.91	0.91	0.91	1

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

E-value	K _j	WeigBted variance proportion				
		intercept	X ₁	X ₂	X ₃	X ₄

3.795	1	7.967564e-05	3.258413e-05	1.8653322e-05	0.0004765330
0.180	8.84				0.00175785
0.0544	16.175	1.9000927e-04	6.530405e-04	6.454593e-04	0.0004302261
0.0021	47.5336				0.46334222
0.0002	123.06	4.213457e-04	1.169124e-03	1.359624e-03	0.5236032778
					0.125357337
		9.086544e-01	1.125518e-01	7.763453e-03	0.2145336251
					0.533322884
		0.985322e-01	7.853035e-01	8.892144e-01	0.2607640230
					0.004321564

The problem of separation

3.3 مشكلة الفصل:

الحصول على مشكلة الفصل تام او شبه تامة الفصل بين المشاهدات تقل كلما ازداد عدد المشاهدات لذلك فقد تم اخذ عينة من 40 مشاهدة من اجل الحصول على مشكلة الفصل، فان عملية الكشف عن مشكلة الفصل تتم من خلال حساب دالة الانحدار اللوجستي التقديرية (احتمال الاستجابة) المبينة في المعادلة (16)، فكانت الاستجابة لبعض المشاهدات أكبر من 0.95 مما يدل على وجود مشكلة الفصل شبه التام كما مبين في الجدول (6).

جدول (6) يمثل قيم $\hat{\pi}_i$ (احتمال الاستجابة)

i	$\hat{\pi}_i$	i	$\hat{\pi}_i$	i	$\hat{\pi}_i$	i	$\hat{\pi}_i$	i	$\hat{\pi}_i$
1	1	8	0.9999997	15	0.9995903	22	1	30	1
2	1	9	0.9828515	16	0.9999632	23	1	31	1
3	0.9976788	10	0.9852172	17	0.9999989	24	0.9926966	32	1
4	0.9994942	11	0.9999997	18	0.9999994	25	0.9316680	34	0.944553
5	0.9447515	12	1	19	0.9992571	26	0.9999986	35	0.936868
6	0.9860895	13	0.9999948	20	0.9854837	27	0.9999992	36	0.925778
7	0.9999990	14	0.9978494	21	0.9999999	28	0.9999414	37	0.913468
						29	1	38	9.034469
								39	1
								40	0.993546

4.3 تقدير معاملات انحدار اللوجستي: Estimate parameter logistic regression models

يتم تقدير معاملات انحدار اللوجستي المتعدد من خلال الاعتماد على عدة طرائق التقدير لمعالجة مشكلة الفصل والتعدد الخطي في عينة بيانات التطبيق , تمت المقارنة بين طرائق التقدير من خلال متوسط مربعات الخطأ (MSE) للأنموذج , ومن الجدير بالذكر استخدم برنامج مكتوب بلغة البرمجة (R) لطرائق تقدير معاملات انموذج

الانحدار اللوجستي المتعدد التي يتم ذكرها في صيغ المعادلات , مقدرات الإمكان الأعظم التكرارية (IMLE) المتمثلة بالصيغة (19) , ومقدرات الانحدار اللوجستي شتاين (SLRE) المتمثلة بالصيغة (24) مقدرات الإمكان الأعظم الجزائية (PMLE) المتمثلة بالصيغة (32) , المقدرات المعدلة (AE) المتمثلة بالصيغة (33) . تم الحصول النتائج الأتية.

جدول (7) يبين تقدير المعلمات ومتوسط مربعات الخطأ لجميع طرائق التقدير

MetBods parameter	IMLE	SLRE	PMLE	Adj
$\hat{\theta}_0$	-86.5645635	-85.555452	-45.97354	18.64536331-6
$\hat{\theta}_1$	9.0534710	8.0532364	0.76453198	0.4476568
$\hat{\theta}_2$	-0.75646301	-0.688701	0.5894145	0.65434
$\hat{\theta}_3$	0.4576709	0.31445617	0.2907748	0.869943
$\hat{\theta}_4$	9.5645146	5.9649988	4.84465643	1.5666413
MSE	0.54540163	0.6763249	0.03565295	0.02544303

لوحظ من الجدول (7) ان نتائج التطبيق للبيانات الحقيقية جاءت متوافقة مع الواقع الحقيقي للبيانات حيث أظهرت مقدرات الانحدار اللوجستي شتاين اقل متوسط مربعات الخطأ للأنموذج من مقدرات الإمكان الأعظم التكرارية في معالجة مشكلة التعدد الخطي ومقدرات الإمكان الأعظم الجزائية اقل متوسط مربعات الخطأ للأنموذج من مقدرات الانحدار اللوجستي شتاين ومقدرات الإمكان الأعظم التكرارية (IMLE) في معالجة مشكلة الفصل حين أظهرت المعدلة Adjusted اقل متوسط مربعات الخطأ من طريقة مقدرات الإمكان الأعظم الجزائية وذلك

لأنها تعالج مشكلة الفصل ومشكلة التعدد الخطي وبالتالي تعد طريقة المقدرات المعدلة Adjusted الطريقة المثالية في معالجة مشكلة الفصل والتعدد الخطي لأنها تحقق اقل MSE من جميع طرائق التقدير الأخرى.

ان تفسير هذه النتائج يعطي صورته واضحة عن الإصابة بفقر الدم وما لها من تداعيات حيث وجد ان الهيموغلوبين (Bb) العامل الأول الذي يؤثر في الإصابة بفقر الدم، من خلال تقدير المعلمة الخاصة به التي تعطي اعلى تقدير من بين المعلمات في الإصابة بفقر الدم ثم تأتي بعدها معلمة تأثير نوع الجنس وبعدها معلمة حجم خلايا الدم المرصوفة واخيرا معلمة خلايا الدم البيضاء وهذا ما يتناسب مع واقع الإصابة بفقر الدم الذي تم التطرق اليه في مقدمة هذا الفصل.

المبحث الرابع: الاستنتاجات والتوصيات

Conclusions

1.4 الاستنتاجات:

- 1- مقدرات الإمكان الاعظم تؤدي الى حل غير دقيق في حدود فضاء المعلمات لأنها تمتلك أكبر متوسط مربعات الخطأ من بين طرائق التقدير الأخرى لأنموذج الانحدار اللوجستي.
- 2- يحدث تباعد في عملية التكرار المثالية لمقدرات الإمكان الأعظم التكرارية في حالة وجود مشكلة الفصل التام والفصل شبه التام لأنموذج الانحدار اللوجستي.
- 3- اثبتت طريقة مقدرات الانحدار اللوجستي اثنتان أفضل من مقدرات الإمكان الأعظم التكرارية في معالجة مشكلة التعدد الخطي من خلال (MSE) لأنموذج
- 4- تكون مشكلة الفصل أكبر تأثير من مشكلة التعدد الخطي والذي يتم مشاهدته من خلال مقدرات الإمكان الأعظم الجزائي ومقدرات الانحدار اللوجستي اثنتان ومتوسط مربعات الخطأ للمقدرات.
- 5- اثبتت طريقة المقدرات المعدلة أفضل من طريقة مقدرات الإمكان الأعظم الجزائية ومقدرات الانحدار اللوجستي اثنتان ومقدرات الإمكان الأعظم التكرارية في معالجة مشكلة الفصل والتعدد الخطي في تقدير المعلمات من خلال مقياس (MSE) لأنموذج.
- 6- من خلال جدول تحليل البيانات في الجانب التطبيقي دراسة العوامل المؤثرة على مرض فقر الدم تم التوصل ان الهيموغلوبين (Bb) العامل الأول الذي يؤثر على الإصابة وهذا جاء متوافق مع رأي الأطباء.
- 7- من خلال مصفوفة الارتباطات الموزونة لبيانات الإصابة بفقر الدم تم التوصل الى وجود مشكلة التعدد الخطي بين المتغيرات التوضيحية، فكان المتغير التوضيحي الأول الي يمثل هيموغلوبين الدم له علاقة خطية طردية مع المتغير التوضيحي الثاني الذي يمثل حجم خلايا الدم المرصوفة وهذا يتناسب مع اراء الأطباء من خلال الفحوصات الخاصة للدم التي تجرى في المختبرات الطبية.

TBe Recommendations

2.4 التوصيات:

- 1- اعتماد طريقة المقدرات المعدلة في تقدير معلمات انموذج الانحدار اللوجستي المتعدد في حالة وجود مشكلة الفصل والتعدد الخطي.
- 2- استعمال اساليب تكرارية أخرى في تقدير معلمات انموذج الانحدار اللوجستي المتعدد في حالة وجود مشكلة الفصل والتعدد الخطي.
- 3- اما في الجانب التطبيقي يجب دراسة إمكانية ادخال متغيرات جديدة قد تكون مهمة في عملية تشخيص فقر الدم فهناك متغيرات قد جرى استبعادها لعدم اكتمال البيانات حولها وكذلك يجب استخدام البرامج الحديثة في تبويب وارشفة البيانات في المؤسسات الحكومية خصوصا المؤسسات الطبية.
- 4- دراسة انموذج الانحدار اللوجستي المتعدد في حالة وجود مشكلة الفصل والتعدد الخطي على بيانات طبيه أخرى مثل مرض التدرن او في المجالات الاجتماعية الأخرى.
- 5- على كل شخص اجراء الفحوصات الخاصة بتشخيص فقر الدم.

المصادر

1. الربيعي, عباس حسين, وجبار, علي مقيم (2010), "دراسة لبعض التغيرات الدموية والكيموحيوية في الأطفال المصابين بمرض التلاسيميا في محافظة بابل", مجلة كلية التربية الاساسية-جامعة بابل, العدد 2, الصفحة 317-310.
2. الراوي, خاشع محمود (1978), "مدخل الى تحليل الانحدار", جامعة الموصل.
3. بيثون, نغم نافع (1992), "خواص قوة الاختبار وحدود الثقة لمعاملات الانموذج اللوجستي الخطي دراسة مقارنة", رسالة ماجستير في الإحصاء, كلية الإدارة والاقتصاد, جامعة بغداد.
4. علي, مهدي كاظم, ورشيد, وسن سعيد (2013), "استجابات عدد من مكونات الدم والإدرار لعدو ١٠٠٠٠ متر وفي فترة استعادة الشفاء لدى عدائي المسافات الطويلة", مجلة كلية التربية الرياضية-جامعة بغداد, المجلد 25, العدد 2, الصفحة 31-52.
5. كاظم, اموري هادي, ومسلم, باسم شيلبية (2002), "القياس الاقتصادي المتقدم النظرية والتطبيق", مطبعة دنيا الأمل, العراق, بغداد.
6. يحيى, مزاحم محمد, وعبدالله, محمود حمدان (2007), "تشخيص التعدد الخطي واستخدام انحدار الحرف في اختيار دالة الاستثمار الزراعي في العراق للفترة (1980-2000)", مجلة تكريت للعلوم الإدارية والاقتصادية, المجلد 3, العدد 8.
7. Allison, P. D (2008), "Convergence Failures in Logistic Regression", University of Pennsylvania, Statistics and Data Analysis, Paper 360.
8. Allison, P. D (1999), "Logistic Regression Using tBe SAS® System: TBeory and Application", CopyrigBt © 1999 by SAS Institute Inc., Cary, NC, USA.
9. Albert, A., and Anderson, J. A. (1984), "On tBe Existence of Maximum LikeliBood Estimates in Logistic Regression Models", Biometrika , Vol. 71, No. 1 (Apr., 1984), pp. 1-10.
10. Albert, A. and Lesaffre, E. (1986), "Multiple Group Logistic Discrimination", Comp & MtBs, Vol. 12A, No. 20, PP. 209-224, Printed in Great Britain.
11. Beinze, G. and ScBemper, M. (2002), "A solution to tBe problem of separation in logistic regression" Statist. Med. , 21:2409–2419 (DOI: 10.1002/sim.1047).

12. GBosB , Joyee , & Liy , Yingbo & Robin , Mitra (2017) , “ On tBe Use of CaucBy Prior Distributions for Bayesian Logistic Regression “ , arXiv:1507.07170v2 [stat.ME] 9 Feb 2017 .
13. Konis, K. (2007), “Linear Programming AlgoritBms for Detecting Separated Data in Binary Logistic Regression Models”, A tBesis submitted for tBe degree of Doctor of PBilosopBy in Statistics, Worcester College, University of Oxford.
14. Lesaffre, E. and Marx, B. D. (1993), “Collinearity in Generalized Linear Regression”, COMMUN. STATIST.-TBEORY METB., Vol. 22, No. 7, pp. 1933-1952.
15. Marx, B. D., and SmitB, E. P. (1990), “WeigBted multicollinearity in logistic regression : diagnostics and biased estimation tecBniques witB an example from lake acidification”, journal Canadian des sciences aquatiques, volume 47, No. 6, pp. 1128-1135.
16. SBen, J., and Gao, S. (2008), “A Solution to Separation and Multicollinearity in Multiple Logistic Regression”, Indiana University ScBool of Medicine, Journal of Data Science 6, PP. 515-531.
17. SBaefer, R. L. (1979), “Multicollinearty and logistic regression”, pB.D. dissertation, tBe university of MicBigan, USA.
18. SBaBmandi, M., and FarmanesB, F., and GBaraBbeigi M., (2013), “Data Analyzing by Attention to WeigBted Multicollinearity in Logistic Regression Applicable in Industrial Data”, BritisB Journal of Applied Science & TecBnology3(4), PP. 748-763.

