استعمال الطرائق الجزائية في تقدير معلمات أنموذج الانحدار اللوجستي في حالة وجود مشكلة الفصل والتعدد الخطي مع تطبيق عملي

أ.م.د.عماد حازم عبودي   طالب الماجستير. اياد حبيب شمال (*)
جامعة بغداد / كلية الادارة والاقتصاد

المبحث الأول: منهجية البحث

1.1 المستخلص:

تناول هذا البحث دراسة موضوع انموذج الانحدار اللوجستي الذي يعد من النماذج اللاخطية اذ يأخذ طابعاً اكثر تقدماً في عملية التحليل الاحصائي الذي يهدف للحصول على تقديرات ذات مستوى عالٍ من الكفاءة.

ان من اهم المشاكل التي تظهر في هذا الانموذج هو الفصل بين مشاهدات المتغير التابع ثنائي الاستجابة الذي يعتمد على احجام العينات ، والتعدد الخطي بين المتغيرات التوضيحية .

اذ تم تطبيق بيانات حقيقية متمثلة بالإصابة بفقر الدم والتي تم الحصول عليها من مستشفى الكوت قسم الكلى الصناعية من خلال طرائق التقدير وفقاً لطرائق مقدرات الأماكن الأعظم الجزائية ومقدرات الإمكان الأعظم الجزائية المزدوجة فكانت مقدرات الإمكان الأعظم الجزائية المزدوجة هي الأفضل على مقدرات الإمكان الأعظم الجزائية في معالجة مشكلة الفصل والتعدد الخطي.

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.

It was has been represented by a real anemia and data that have been true of incidence of, which was obtained from Kut hospital of Department artificial during estimate methods according to the penalized maximum likelihood estimators and double penalized maximum likelihood estimators parameters were double penalized maximum likelihood estimators is the best on penalized maximum likelihood estimators in the treatment of separation problem and Multicollinearity.

(*) جزء مستل من رسالة ماجستير للباحث الثاني.

2.1 المقدمة : Introduction

يُعرف 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) وبناء على ذلك فان أسلوب الانحدار يستعمل للتوصل إلى نموذج رياضي يوضح العلاقة الكمية بين المتغير التابع المراد التنبؤ بقيمته والمتغيرات التوضيحية.[2]

استعمل العديد من الباحثين أنموذج الانحدار اللوجستي ويعتبر 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 في ما 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).[4]

فقد بدأت تظهر بعض المشاكل في البيانات الثنائية لأنموذج الانحدار اللوجستي مثل مشكلة التعدد الخطي (multicollinearty) بين المتغيرات التوضيحية ومشكلة الفصل (separation) مشكلة التعدد الخطي تعتمد على وجود علاقة خطية بين بعض او كل المتغيرات التوضيحية أي وجود الارتباط بين هذه المتغيرات فيصبح انموذج الانحدار اللوجستي غير مستقر.

اما مشكلة الفصل (separation) فأنها تعتمد على شكل البيانات أي عندما تكون بيانات المتغير التابع ثنائية الاستجابة وعلى حجم العينة أي مشكلة الفصل تحدث عندما تكون العينات صغيرة وكلما ازداد عدد المشاهدات تقل فرصة الحصول على مشكلة الفصل ان مكانية حصول مشكلة الفصل شائع اكثر مما هو متوقع في مجال التطبيقات الطبية.[1]

ان طرائق التقدير الكلاسيكية للمعلمات أنموذج الانحدار اللوجستي عند تحليل البيانات ثنائية الاستجابة(Binary data response) ضعيفة في معالجة مشكلة التعدد الخطي والفصل بين البيانات الامر الذي يستوجب استخدام طرائق لمعالجة تلك المشاكل.

3.1 مشكلة البحث : Problem of research

عند استعمال أنموذج الانحدار اللوجستي مع البيانات ثنائية الاستجابة فأن من اهم المشاكل التي تظهر في هذا الانموذج هو الفصل بين بيانات المتغير التابع ثنائية الاستجابة والتعدد الخطي بين المتغيرات التوضيحية ، ان وجود العلاقة القوية بين المتغيرات التوضيحية في انموذج الانحدار اللوجستي المتعدد يؤدي الى ظهور مشكلة التعدد الخطي الذي يجعل الانموذج غير مستقر ومعلمات التقدير غير دقيقة ، لذلك فان تفسير العلاقة بين متغير الاستجابة وكل المتغيرات التوضيحية من حيث نسبة الارجحية قد تكون خاطئة ، اما الفصل في بيانات المتغير التابع ثنائية الاستجابة يعتمد على حجم العينة أي عندما يكون حجم العينة صغير.

4.1 هدف البحث : Object of research

من خلال الإجابة على التساؤلات في هذا البحث سيحقق الأهداف الأتية.

1- معالجة مشكلة الفصل والتعدد الخطي في انموذج الانحدار اللوجستي المتعدد من خلال طرائق تقدير معلمات هذا الانموذج، اذ تم ذلك من خلال تطبيق طرائق التقدير على بيانات فقر الدم التي تعاني مشكلة الفصل والتعدد الخطي بغية التوصل الى اهم العوامل المؤثر على الإصابة بفقر الدم.

2- مقارنة طرائق تقدير معلمات انموذج الانحدار اللوجستي في ظل وجود مشكلة الفصل والتعدد الخطي من خلال معيار متوسط مربعات الخطأ MSE لتحديد الأفضل منها .

المبحث الثاني: الجانب النظري

1.2 المقدمة: Introduction

في هذا المبحث سيتم التطرق الى انموذج الانحدار اللوجستي والصيغ العامة والافتراضات الخاصة به ، مروراً بالتطرق الى 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 استعراض تقدير معلمات انموذج الانحدار اللوجستي في ظل وجود مشكلة التعدد الخطي والفصل.

2.2 انموذج الانحدار اللوجستي:[1][2][18][19] Logistic regression model

يعرف أنموذج الانحدار اللوجستي على انه احد نماذج الانحدار الذي تكون فيه العلاقة بين المتغير التابع y والمتغيرات التوضيحية $(x_0, x_1, x_2 \dots x_p)$ غير خطية حيث يكون المتغير التابع y ثنائي الاستجابة مفترضاً إحدى القيمتين (1,0) أما النجاح success حدوث الاستجابة باحتمال π_i أو الفشل failure عدم حدوث الاستجابة باحتمال $1-\pi_i$ لهذا يكون المتغير التابع y يتبع توزيع برنولي وسوف تكون دالة الكثافة الاحتمالية بالصيغة الأتية.[1][2]

$$p(Y = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad . \; . \; . \; (1)$$
$$y_i = 0,1$$

اذ أن

y_i متغير تابع ثنائي الاستجابة.

π_i احتمال حدوث الاستجابة عندما $y_i = 1$.

لذلك فان توقع المتغير التابع يمثل احتمال حدوث الاستجابة .

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

أما تباين المتغير التابع حسب توزيع برنولي هو .

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

ليكن $x_0, x_1, x_2, \dots \dots x_p$ مجموعة من المتغيرات التوضيحية ولتكن n تمثل عدد المشاهدات لهذه المتغيرات التي تكون المصفوفة X

$$X = (x_{ij})_{n\times p} \quad . \; . \; . \; (4)$$

اذ ان :

$i= 1,2\dots\dots n$ (n) تمثل حجم العينة.

$j= 0,1,2\dots p$ (p) عدد المعلمات.

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

هذا يؤدي الى أن أنموذج الانحدار اللوجستي يعطى بالصيغة الأتية.

$$y_i = \pi_i + \varepsilon_i \quad . \; . \; . \; (5)$$

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

$$\pi_i = p(y=1) = \frac{e^{\underline{x}_i\underline{\beta}}}{1+e^{\underline{x}_i\underline{\beta}}} \quad . \; . \; . \; (6)$$

$\underline{\beta}$:متجه من المعلمات أبعاده $(p\times1)$.

$\underline{x}_i = \{x_{i0}, x_{i1}, \dots x_{ip}\}$:موجه صفي من المتغيرات التوضيحية أبعاده $(1\times p)$.

ε_i :حد الخطأ العشوائي بمتوسط صفر كما مبين في الصيغة الاتية .

$$\varepsilon_i = y_i - \pi_i$$

$$E(\varepsilon_i) = E(y_i) - E(\pi_i) = \pi_i - \pi_i = 0 \quad . . . (2-7)$$

أما تباين حد الخطأ العشوائي فانه يساوي تباين المتغير المعتمد ثنائي الاستجابة .

$$V(\varepsilon_i) = V(y_i) = \pi_i(1-\pi_i) \quad . . . (2\text{-}8)$$

لذا فان حد الخطأ العشوائي يتبع توزيع برنولي 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 فيكون تشخيص التعدد الخطي بوجود الاوزان.[19][18]

3.2 التحويل الخطي إلى دالة الانحدار اللوجستي (دالة اللوجت):[16][4]

يميل الكثير من الإحصائيين إلى أزاله انحناءات معلمات دالة الانحدار اللوجستي من خلال استعمال تحويل دالة اللوجت ، للتأثير السلبي لهذه الانحناءات في حالة وجودها على خصائص مقدرات المعلمات ومن ثم قيم حدوث الاستجابة التي يتم التنبؤ بها .

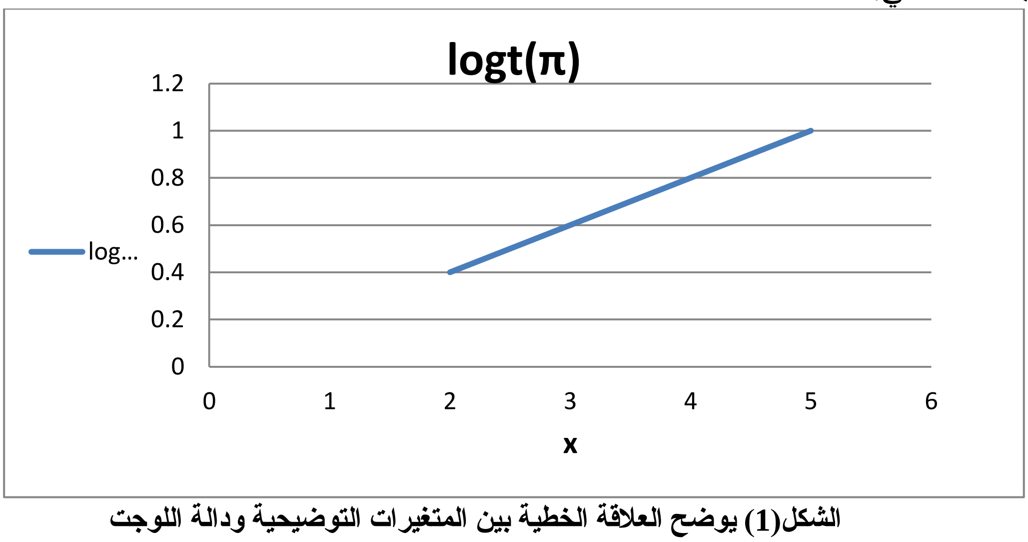
تمكن الباحث Berkson عام 1944 من أيجاد علاقة لوغاريتمية لتحويل العلاقة بين المتغيرات التوضيحية واحتمال حدوث الاستجابة π_i إلى علاقة خطية وذلك برسم $\text{logit}(\pi_i)$ والتي تكون بالشكل الاتي.

$$\text{logit}\,(\pi_i) = \ln \frac{\pi_i}{1-\pi_i} \quad . . . (9)$$

$$\text{logit}\,(\pi_i) = \ln \frac{\frac{e^{\underline{x}_i\underline{\beta}}}{1+e^{\underline{x}_i\underline{\beta}}}}{1-\frac{e^{\underline{x}_i\underline{\beta}}}{1+e^{\underline{x}_i\underline{\beta}}}}$$

$$\text{logit}\,(\pi_i) = \ln\left(e^{\underline{x}_i\underline{\beta}}\right) = \underline{x}_i\underline{\beta} \quad . . . (10)$$

أن المعادلة رقم (9) تمثل العلاقة الخطية بين احتمال حدوث الاستجابة π_i ودالة اللوجت $\text{logit}\,(\pi_i)$ أي تمثل العلاقة بين المتغيرات التوضيحية ودالة اللوجت ويمكن أن تمثل بالرسم بالشكل الاتي.



الشكل(1) يوضح العلاقة الخطية بين المتغيرات التوضيحية ودالة اللوجت

4.2 تشخيص مشكلة التعدد الخطي الموزون في الانحدار اللوجستي:[18][17][6]

أنموذج الانحدار اللوجستي يصبح غير مستقر عند وجود الاعتماد القوي بين المتغيرات التوضيحية $(x_1, x_2 \ldots x_k)$ لذلك يبدو انه لا يوجد متغير مهم عن كل المتغيرات الأخرى التوضيحية في الأنموذج لعملية تشخيص مشكلة التعدد الخطي ، ومن اجل الحصول على التشخيص المناسب لمشكلة التعدد الخطي الموزون يفضل استعمال مصفوفة المعلومات الموزونة.

لتكن $[y_1,y_2,\ldots..y_n]$ تمثل (n) من المتغيرات العشوائية ثنائية الاستجابة المستقلة في ما بينها والتي تتوزع حسب توزيع برنولي فان المصفوفة القطرية $\widehat{W}$ تحتوي على التباين المقدر لقيم y.

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

ويمكن التعبير عن مصفوفة المعلومات الموزونة المقدرة بالصيغة الاتية .

$$\hat{\phi} = \acute{X}\widehat{W}X \quad . \ . \ . \ (11)$$

او تكتب بالصيغة الاتية .

$$\hat{\phi} = \acute{\hat{S}}\hat{S}$$

اذ ان $\hat{S} = \widehat{W}^{0.5}X$ لذلك سوف تكون مصفوفة التباين والتباين المشترك لمعلمات مقدرات الإمكان الأعظم التكرارية بالصيغة الأتية .

$$\hat{v}ar(\hat{B}) = \hat{\phi}^{-1} \quad . \ . \ . \ (12)$$

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

1.4.2 العدد الشرطي الموزون:[18][17] weighted condition number

قيم اعداد الشرط الموزون تعتمد على مصفوفة الارتباطات الموزونة 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 بين المتغيرات التوضيحية في أنموذج الانحدار اللوجستي.

لتكن $\lambda_0^*, \lambda_1^*, \ldots \ldots \ldots, \lambda_p^*$ القيم الذاتية المرتبة (الجذور المميزة) الى مصفوفة الارتباطات الموزونة المقدرة.

$$\hat{\phi}^* = \acute{\hat{S}}^*\hat{S}^* \quad . \ . \ . \ (13)$$

عندما ${\hat{S}^*}_{ij}$ تمثل بالصيغة الأتية.

$${\hat{S}^*}_{ij} = \frac{\hat{s}_{ij}-\bar{s}_j}{\sqrt{\sum_{i=1}^{n}(\hat{s}_{ij}-\bar{s}_j)^2}} \quad . \ . \ . \ (14)$$

وان الوسط الحسابي $\bar{s}_j$ يكون.

$$\bar{s}_j = \frac{\sum_{i=1}^{n}\hat{s}_{ij}}{n} \quad . \ . \ . \ (15)$$

اذ ان اعداد الشرط تعرف كالاتي:

$$k_j = \left(\frac{\lambda^* \max}{\lambda_j^*}\right)^{0.5} \quad . \ . \ . \ (16)$$

اذ أن:

$\lambda^*\max$ هو اكبر الجذور المميزة لمصفوفة الارتباطات الموزونة.

λ_j^* هو الجذور المميزة j في المصفوفة $\hat{\phi}^*$.

2.4.2 نسبة التباين الموزون:[18][17][6] weighted variance proportion

الكشف عن مشكلة التعدد الخطي في هذه الحالة يعتمد على قيم نسبة التباين الموزونة حيث تشكل القيم مصفوفة إبعادها (u×j) إي أنها تكون u من الجذور المميزة لذ ففي حالة وجود قيمتين

من قيم نسبة التباين الموزونة كبيرتان وبالقرب من بعضهما تحدث مشكلة التعدد الخطي الموزون التي تضر بالخصائص المرغوب فيها من الانحدار اللوجستي.

ليكن m_{ju} عنصر في مصفوفة المتجهات المميزة التي يتم الحصول عليها من مصفوفة المعلومات المقدرة $\hat{\phi}^*$ وان Λ^* مصفوفة قطرية من الجذور المميزة لمصفوفة المعلومات المقدرة $\hat{\phi}^*$ التي تحقق الشرط الاتي .

$$\acute{M}\hat{\phi}^* M = \Lambda^* \quad . . . \ (17)$$

اذ ان M مصفوفة المتجهات المتعامدة المميزة لمصفوفة المعلومات المقدرة $\hat{\phi}^*$.

فان نسبة التباين ω_{uj} الموزونة يمكن التعبير عنها كالاتي .

$$\omega_{uj} = \frac{{m_{ju}}^2/\lambda_u^*}{c_{jj}} \quad . . . \ (18)$$

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

λ_u^* الجذر المميز المحسوب من مصفوفة المعلومات المقدرة $\hat{\phi}^* = \acute{\hat{S}}^*\hat{S}^*$

c_{jj} الجذور المميزة الصغيرة (القيم الذاتية الصغيرة نسبة الى اكبر جذر مميز) ويمكن أن يحسب بالصيغة الأتية.

$$c_{jj} = \sum_{u=0}^{p} {\lambda_u^*}^{-1} {m^2}_{ju} \quad . . . \ (19)$$

اذ ان :

$u=0,1,2\ldots\ldots\ldots\ldots\ldots\ldots p$ (u) عدد الجذور المميز

وعلى سبيل المثال إذا كان لدينا قيمة نسبة التباين الموزونة ω_{32} والقيمة ω_{34} كبيرتان وبالقرب من بعضهما إي إن كلا القيمتان تقع في نفس المتجه المميز رقم (3) سوف تكون لهما علاقة بمشكلة التعدد الخطي ويكون التباين بين $\hat{\beta}_2$ و $\hat{\beta}_4$ كبير جدا.

5.2 فصل البيانات في الانحدار اللوجستي:[7][8][10][13] separation of the data in logistic regression

اظهر ألبرت وأندرسون (1984) أن مجموعة نقاط العينة n يمكن تصنيفها في واحد من ثلاث تكوينات متنافية للمتغير التابع (الفصل التام ،الفصل شبه التام ، التداخل)، نحن 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 نقاط العينة مقدرات الإمكان الأعظم تؤدي إلى حل غير وحيد في حدود فضاء المعلمات ، وتكون مقدرات الإمكان الأعظم محددة عندما يكون هناك تداخل بين نقاط العينة.[13]

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 تظهر مشكلة الفصل بشكل واسع عندما تكون هناك g من مجاميع الاستجابة response categories أي $\{s=1,2,\ldots\ldots\ldots,g\}$ وبالتالي وجود الفصل قد يشير فقط الى أن هذا الانموذج هو أكثر تركيب للبيانات لذلك عند اختيار انموذج مختلف على سبيل المثال تحليل التمايز يمكن تجنب المشاكل الناجمة عن الفصل تماما.[8]

1.5.2 الفصل التام: [7][8][10] ▯▯▯▯▯▯▯▯ ▯▯▯▯▯▯▯▯▯▯

يحدث الفصل التام في نقاط العينة اذا كان هناك موجه من المعلمات يخصص كل المشاهدات الى مجموعاتهم بشكل صحيح فاذا كان هناك مجموعتين من الاستجابة $(E_1 ، E_2)$ في مشاهدات العينة اي عندما يكون $\underline{x}_i\beta > 0$ لكل $x_i \epsilon E_1$ وان $\underline{x}_i\beta < 0$ لكل $x_i \epsilon E_2$ ، وفي حالة أنموذج الانحدار اللوجستي ثنائي الاستجابة الفصل التام يحدث بين مجموعة من مشاهدات العينة n أي

أذا كان لدينا متجه من المعلمات $\underline{\beta}$ بشرط $\underline{x_i\beta} > 0$ عندما $y_i = 1$ و ان $\underline{x_i\beta} < 0$ عندما $y_i = 0$ كما مبين في الجدول رقم (1).

الجدول رقم (1) يستعرض الفصل التام للبيانات

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

2.5.2 الفصل شبه التام:[10][8][7] □□□□□□□□□□□□□ □□□□□□□□□n

الفصل شبه التام يحصل عندما تكون الدالة من x تولد تنبؤات غير مثالية الى y فاذا كان لدينا موجه من المعلمات β بحيث $\underline{x_i\beta} \geq 0$ عندما $y=1$ و $\underline{x_i\beta} \leq 0$ 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 شبة التام للبيانات

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

أن في حالة وجود الفصل التام و الفصل شبه التام للبيانات يمكن ملاحظة .

1- مقدرات الإمكان الأعظم المرجحة $\hat{\beta}$ تكون غير محددة في حالة الفصل التام و الفصل شبه التام.

2- أن حالة الفصل التام وشبه التام تودي الى حدوث تباعد في عملية التكرار المثالية عند استخراج مقدرات الإمكان الأعظم التكرارية.

3.5.2 التداخل:[7][8][10]

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 ، وفي حالة التداخل تكون المعلمات المقدرة $\hat{\beta}$ موجودة وان عملية التعظيم التكراري تتقارب إلى حل نهائي وحيد .

6.2 الكشف عن الفصل التام والفصل شبه التام:[7][10]

الكشف عن الفصل التام والفصل شبه التام يتم ذلك من خلال الحصول على مقدرات الإمكان الأعظم التكرارية $\hat{\beta}$ التي تعطينا تشخيصاً صحيحا لجميع مواقع المشاهدات ففي حالة وجود الفصل التام والفصل شبه التام يستمر التكرار حتى يتم تجاوز التكرار الثابت لذلك سوف يكون تقديرات المعلمات كبيرة وحد الخطأ كبير للغاية اقترح ألبرت وأندرسون عام 1984 وسيلة تجريبية للكشف عن الفصل التي كان تنفيذها في PROTCLO LOGISTIC انه يحتوي على الخطوات الأتية.

1- إذا معيار التقارب تحقق في غضون ثمانية تكرارات نستنتج إلى أن هناك لا توجد مشكلة الفصل التام والفصل شبه التام بصورة عامة أي هناك تداخل بين البيانات.
2- لجميع التكرارات بعدد التكرار الثامن احتمال توقع الاستجابة لكل مشاهدة تعطى من قبل الصيغة الأتية.

$$\hat{\pi}_i = \frac{1}{1+\exp[(2y_i-1)\underline{x}_i\underline{\hat{\beta}}]} \quad . \; . \; . \quad (20)$$

اذا كان احتمال التوقع هو واحد لجميع المشاهدات نستنتج أن هناك فصل تام وفي هذه يتوقف التكرار.

أذا كان احتمال استجابة المشاهدات اكبر من (0.95) لبعض المشاهدات وليس لكلها نستنتج أن هناك فصل شبه تام ووقف التكرار.

7.2 طرائق تقدير معلمات نموذج الانحدار اللوجستي:[1][9][10][11][12][14][15][19]

1.7.2 مقدرات الإمكان الأعظم التكرارية:[1][9][10] Iterative Maximum Likelihood Estimators

هدف مقدرات الإمكان الأعظم هو أيجاد مجموعة من القيم المقدرة الى $\underline{\beta}$(تقدير المعلمات التي تجعل دالة الإمكان أعظم ما يمكن) ، دالة الإمكان الأعظم لنموذج الانحدار اللوجستي الذي يتبع توزيع برنولي تكون بالصيغة الأتية.

$$L(\beta, X) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \; . \, . \, . \, (21)$$

$$L(\beta, X) = \prod_{i}^{n}(1-\pi_i)\exp[\textstyle\sum_{i=1}^{n} y_i \log(\frac{\pi_i}{1-\pi_i})]$$

وحسب خاصية التحويل لدالة اللوجت

$$\log\frac{\pi_i}{1-\pi_i} = \underline{x}_i\underline{\beta}$$

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

وبأخذ اللوغاريتم إلى دالة الإمكان.

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

أن النهج المعروف لدالة الإمكان الأعظم فيما يتعلق في β من خلال مساواة المشتقة الأولى الى لوغاريتم دالة الإمكان الأعظم للصفر ثم حل مجموعة من المعادلات الناتجة من المشتقة :

$$\frac{\partial \log L(\beta,X)}{\partial \underline{\acute{\beta}}} = 0 \quad \ldots (22)$$

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 الأتية.

$$\underline{\beta}^{(s+1)} = \underline{\beta}^{S} - I(\beta)^{-1(s)} U(\beta)^{(s)} \quad \ldots (23)$$

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

$$U(\beta) = \frac{\partial \log L(\beta,X)}{\partial \underline{\acute{\beta}}} = \sum_{i=1}^{n} \underline{\acute{x}}_i (y_i - \pi_i) \ldots (24)$$

ويمكن كتابة $U(\beta)$ بشكل متجه عمودي كالاتي .

$$U(\beta) = \acute{X}(y_i - \pi_i)$$

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

$$I(\beta) = \frac{\partial^2 \log L(\beta,X)}{\partial \acute{\beta}\, \partial \beta} = -\sum_{i=1}^{n} \underline{\acute{x}}_i \underline{x}_i \pi_i (1 - \pi_i) \ldots (25)$$

وبذلك تكون مصفوفة المعلومات بالصيغة الأتية.

$$I(\beta) = -\acute{X} \operatorname{diag} \pi_i (1 - \pi_i) X = -\acute{X} W X$$

s: يمثل عدد التكرارات.

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

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

2.7.2 مقدرات الإمكان الأعظم الجزائية: [19][15][12] Penalized Maximum Likelihood Estimators

اقترحت مقدرات الإمكان الأعظم الجزائية من قبل (Firth 1993)[15] إجراء (Firth) وضع أصلا للحد من التحيز في مقدرات الإمكان الأعظم التكرارية لتوفير الحل المثالي الى الفصل ، وينتج تقديرات معلمة محددة بواسطة مقدرات الإمكان الأعظم الجزائية في حالة وجود الفصل التام والفصل شبه التام الذي يحدث على بيانات العينة ، فان دالة الإمكان الاعظم الجزائية لـFirth هي.

$$L(\beta, X)^* = L(\beta, X) |I(\beta)|^{1/2} \quad \ldots (26)$$

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

$$U(\beta^*) = U(\beta) + \frac{1}{2}\ \operatorname{trace}\left[I(\beta)^{-1} \left\{ \frac{\partial I(\beta)}{\partial \acute{\beta}} \right\} \right] \ldots (27)$$

Firth اظهر أن تحيز مقدرات الإمكان الأعظم تم إزالته ، أن غرض صيغة الانحدار اللوجستي الجزائية (المقطعية) تخفيض التحيز وحل مشكلة الفصل ، أن الفكر العامة إلىfirth هو استبدال معادلة النتيجة$U(\beta) = \sum_{i=1}^{n} \underline{x}_i (y_i - \pi_i)$ بمعادلة النتيجة المعدلة التي تكتب بالصيغة.

$$U(\beta^*) = \sum_{i=1}^{n} [(y_i - \pi_i) + h_i(\tfrac{1}{2} - \pi_i)] \underline{\acute{x}}_i \ldots (28)$$

عندما h_i عناصر القطر الرئيسي للمصفوفة H

$$H = W^{\frac{1}{2}}X(\acute{X}WX)^{-1}\acute{X}W^{\frac{1}{2}} \quad \ldots (28)$$

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

$$U(\beta^*) = \acute{X}(y_i - \pi_i) + \acute{X}h(\frac{1}{2} - \pi_i) \quad \ldots (29)$$

h تمثل المصفوفة القطرية تكتب بالشكل التالي $h = diag(H)$

مقدرات firth يمكن الحصول عليها تكراريا من خلال طريقة نيوتن رافسون.

$$\beta^{(s+1)} = \beta^{S} - I(\beta)^{-1(s)}U(\beta^*)^{(s)} \quad \ldots (30)$$

أن s عدد التكرارات ويتم التوقف عن عدد التكرارات حتى يكون الفرق بين عمليتين تكراريتين هو صفر أو قريب من الصفر وبذلك تكون طريقة Firth قد عالجت مشكلة الفصل.

3.7.2 مقدرات الإمكان الأعظم الجزائية المزدوجة:[19][14][11] Double Penalized Maximum Likelihood Estimators

في هذه الطريقة يقترح إدخال معلمة الحرف على مقدرات الإمكان الأعظم الجزائية، أي تضم هذه الطريقة احتمال Firth الجزائي مع مصطلح معلمة الحرف والتي هي قادرة على التعامل مع كل من الفصل والتعدد الخطي لتحسين مقدرات المعلمات وتقليل أخطاء التنبؤ.

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.

$$\log L(\beta, X)^{**} = \log L(\beta, X) + \frac{1}{2}\log|I(\beta)| - \lambda\|P\beta\|^2 \quad \ldots (31)$$

عندما P مصفوفة من الرتبة $(p\times p)$ التي تفرض قيود خطية على المعلمات β .

خوارزمية نيوتن رافسون التكرارية إلى دالة الإمكان الجزائية المزدوجة تعرف كالاتي .

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} + I(\beta^*)^{-1(s)}U(\beta^{**})^{(s)} \quad \ldots (32)$$

$I(\beta^*)^{-1}$ معكوس مصفوفة تقيم المعلومات لدالة الإمكان الجزائية المزدوجة.

$U(\beta^{**})$ المشتقة الأولى إلى لوغاريتم دالة الإمكان الجزائية المزدوجة (معادلة النتيجة المعدلة) والتي تكون بالصيغة الأتية .

$$U(\beta^{**}) = U(\beta) + \frac{\partial}{\partial\acute{B}}\left(\frac{1}{2}\log|I(\beta)|\right) - 2\lambda\beta \quad \ldots (33)$$

أما الصيغة الملائمة حسابيا في شكل المصفوفة $\frac{\partial}{\partial B}\left(\frac{1}{2}\log|I(B)|\right)$ قدمها بيل وآخرون 2002 .

$$\frac{\partial}{\partial\acute{B}}\left(\frac{1}{2}\log|I(\beta)|\right) = -\frac{1}{2}\{\acute{X}Q(X \otimes X)vec(I(\beta)^{-1})\} \quad \ldots (34)$$

اذ ان V مصفوفة: $Q = V \otimes ei \quad \ldots (35)$

قطرية من الرتبة $(n\times n)$ عناصر القطر الرئيسي تساوي $\pi_i(1-\pi_i)(1-2\pi_i)$

$\otimes$: يمثل الظرب الداخلي للمصفوفات

ei:موجه من الرتبة $(1\times n)$ يساوي واحد في العمود i وصفر ما عدا ذلك في الظرب الداخلي للمصفوفات.

$vec(I(\beta)^{-1})$: جعل أعمدة المصفوفة $I(\beta)^{-1}$ في متجه عمودي واحد.

مصفوفة المعلومات إلى دالة الإمكان الجزئية المزدوجة التي تمثل المشتقة الثانية الى لوغاريتم دالة الإمكان الجزائية المزدوجة تكون كالآتي .

$$I(\beta^*) = -I(\beta) + \frac{1}{2}\frac{\partial^2\log|I(\beta)|}{\partial\beta^2} - 2\lambda I \quad \ldots (36)$$

وبما أن من الصعب أيجاد المشتقة الثانية اقترح (Shen & Gao)[19] استخدام.

$$I(\beta^*) = -I(\beta) - 2\lambda I \quad \ldots (37)$$

λ :معلمة الحرف.

I : تمثل مصفوفة الوحدة.

المبحث الثالث: الجانب التطبيقي

1.3 وصف البيانات:[5][3] Description of data

فقر الدم هو نقص في عدد كريات الدم الحمراء السليمة المسؤولة عن نقل الاوكسجين لكافة انحاء الجسم وتعددت أنماط فقر الدم اعتماداً على العامل المسبب له، وهو حالة تحدث بسبب انخفاض تركيز الهيموغلوبين عن المستوى الطبيعي عند الاناث البالغات اقل من 11 غم/ ديسي لتر والذكور البالغين اقل من 13 غم/ ديسي لتر، وبسبب الهبوط في مستوى الهيموغلوبين تعاني الأجهزة من عدم الحصول على ما يكفي من الاوكسجين، فيما يتعلق بدراستنا التطبيقية فسوف يتم دراسة فقر الدم من خلال المكونات الأساسية للدم وهي:

1- **الهيموغلوبين (Hb):**هو عبارة عن صبغة حمراء ناقلة للأوكسجين ويتكون الجزء الواحد من الهيموغلوبين من جزء بروتين يسمى الكلوبين الذي يرتبط بأربعة جزيئات من الهيم المحتوي على الحديد ، وتتراوح الكمية الطبيعية للهيموغلوبين لدى الذكور (13-17)غم/ 100 مل من الدم بينما تتراوح هذه النسبة لدى الاناث (11-15) غم/ 100 مل من الدم.
2- **حجم خلايا الدم المرصوصة (pcv) :**يعرف حجم الخلايا المرصوصة (pcv) بانة نسبة حجم الدم المشغول بخلايا الدم الحمراء الى حجم الدم الكامل ،ويبلغ الحجم للخلايا المرصوصة عادة لدى الرجال حوالي 48% والدى النساء حوالي 38% وتعد جزء مكمل لدم الشخص الكامل ويستعمل حجم خلايا الدم المرصوصة لتحديد كتلة خلايا الدم الحمراء بعد اجراء عملية الطرد المركزي على الدم وترسب خلايا الدم الحمراء في أنبوب الهيماتوكرت.
3- **خلايا الدم البيضاء (WBC) :**عبارة عن كرات صغيرة عديمة اللون توجد في الدم والسائل اللمفاوي وتتحرك حركة اميبية تعمل على التهاب الميكروبات وهي عبارة عن خلايا كبيرة مقارنة بخلايا الدم الحمراء وتحتوي على النواة ولا تحتوي على صبغة الهيموغلوبين لذلك تبدو شفافة وبيضاء اللون وهي اقل وفرة من خلايا الدم الحمراء ويتراوح عددها ما بين (5000-11000) خلية لكل مايكرو لتر من الدم توجد في اللمف وسائل النسيج وتكون فترة حياة عدد من خلايا الدم البيضاء أياما قليلة ، في حين تعيش الأخرى لأشهر وحتى لسنين وتقوم بوظيفة الدفاع عن الجسم من خلال عملية البلعمة Phagocytosis للأجسام الغريبة ومن خلال استجابات مناعية متخصصة.

2.3 تصنيف البيانات: Classification of data

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 2014 وبعينة حجمها (29) شخص من خلال اجراء تحاليل الدم لكل شخص واخذ اهم المتغيرات 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=0) او عدم الإصابة بمرض فقر الدم (y=1) بالاعتماد على المتغيرات التوضيحية الاتية .

1- المتغير(x_1) يمثل هيموغلوبين الدم ويقاس بوحدة (غم/100مل من الدم) او (غم/ديسي لتر من الدم).
2- المتغير(x_2) يمثل حجم خلايا الدم المرصوصة (pcv) ويقاس بالنسبة المئوية.

3- المتغير (x_3) يمثل خلايا الدم البيضاء (wbc) ويقاس بوحدة قياس (خلية لكل الف مايكرو لترمن الدم).

4- المتغير (x_4) ويمثل الجنس ويضم فئتين الذكور $(x_4 = 1)$ والاناث $(x_4 = 2)$.

3.3 اختبار بيانات المتغير التابع:

لمعرفة طبيعة التوزيعات الاحتمالية التي من الممكن ان يتوزع المتغير التابع تبعاً لها تم استعمال البرنامج الجاهز (Easy Fit) وقد ظهر بان المتغير التابع y يتبع توزيع برنولي بمعلمة توزيع قدرها $(\pi=0.62069)$.

ومن اجل التحقق من ملائمة توزيع برنولي للمتغير التابع في الجانب التطبيقي اظهر الجدول (3) توافق نتائج اختبار حسن المطابقة لمتغير الإصابة بفقر الدم من خلال اختبار (اندرسون-دارلينغ) والذي من خلاله يتم التأكد على ان توزيع برنولي من ابرز التوزيعات التي من الممكن ان يتبع لها المتغير التابع.

الجدول (3) نتائج اختبار حسن المطابقة لبيانات المتغير المعتمد في التطبيق

#	Distribution	Anderson Darling	
		Statistic	Rank
1	Bernoulli	11.415	1
2	Binomial	11.415	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 (data min < 1)	
8	Neg. Binomial	No fit	

4.3 مشكلة التعدد الخطي الموزون في بيانات التطبيق:

لأجل الكشف عن وجود ازدواج وتعدد 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 والجذور المميزة والعدد الشرطي الموزون ونسبة التباين الموزونة.

يلاحظ عبر مصفوفة الارتباط الموزونة الموضحة بالجدول (4) وجود معاملات الارتباط ذات قيم كبيرة وطردية 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 (16) 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 (18) 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.

الجدول (4) يبين مصفوفة الارتباط الموزونة بين المتغيرات التوضيحية

	Intercept	X_1	X_2	X_3	X_4
Intercept	1.0000000	0.9966068	0.9970661	0.9864695	0.9595706
X_1	0.9966068	1.0000000	0.9996664	0.9777572	0.9397705
X_2	0.9970661	0.9996664	1.0000000	0.9769747	0.9418057
X_3	0.9864695	0.9777572	0.9769747	1.0000000	0.9576042
X_4	0.9595706	0.9397705	0.9418057	0.9576042	1.0000000

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

E-value	K_j	Weighted variance proportion				
		intercept	X_1	X_2	X_3	X_4
4.893815	1.00000	4.922286e-05	2.258413e-05	1.709794e-05	0.0006688480	0.001493101
0.080559	7.79410	2.900033e-04	9.530405e-04	6.454593e-04	0.0004302261	0.336256395
0.024314	14.1870	4.201627e-04	1.169124e-03	1.359624e-03	0.5236032778	0.125357337
0.001071	67.5758	8.029984e-01	1.125518e-01	8.763453e-03	0.2145336251	0.533322884
0.000239	143.094	1.962422e-01	8.853035e-01	9.892144e-01	0.2607640230	0.003570284

5.3 مشكلة الفصل في بيانات التطبيق:

ان فرصة الحصول على تشكيلة من نقاط البيانات تامة الفصل او شبه تامة الفصل تقل كلما ازداد عدد المشاهدات لذلك فقد تم اخذ عينة من 29 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 (20) ، فكانت الاستجابة لبعض المشاهدات اكبر 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	$\hat{\pi}_i$
1	1.0000000	8	0.9999997	15	0.9995903	22	1.0000000
2	1.0000000	9	0.9828515	16	0.9999632	23	1.0000000
3	0.9976788	10	0.9852172	17	0.9999989	24	0.9926966
4	0.9994942	11	0.9999997	18	0.9999994	25	0.9316680
5	0.9447515	12	1.0000000	19	0.9992571	26	0.9999986
6	0.9860895	13	0.9999948	20	0.9854837	27	0.9999992
7	0.9999990	14	0.9978494	21	0.9999999	28	0.9999414
						29	1.0000000

6.3 تقدير معلمات انموذج الانحدار اللوجستي:

تم تقدير معلمات الانموذج (5) باستعمال عدة طرائق التقدير لمعالجة مشكلة الفصل والتعدد الخطي في عينة بيانات التطبيق ، المتمثلة بالإصابة بفقر الدم بعد الكشف عن وجود مشكلة الفصل شبه التام والتعدد الخطي في هذه البيانات، وتمت المقارنة بين طرائق التقدير من خلال متوسط مربعات الخطأ (MSE) للأنموذج ، ومن الجدير بالذكر استخدم برنامج مكتوب بلغة البرمجة (R) لطرائق تقدير معلمات انموذج الانحدار اللوجستي التي يتم ذكرها في صيغ المعادلات الأتية .

مقدرات الإمكان الأعظم التكرارية (IMLE) المتمثلة بالصيغة (23).

مقدرات الإمكان الأعظم الجزائية (PMLE) المتمثلة بالصيغة (30).

مقدرات الإمكان الأعظم الجزائية المزدوجة (DPMLE) المتمثلة بالصيغة (32).

بعد تطبيق صيغ المعادلات أعلاه على البيانات الحقيقية تم الحصول النتائج الأتية .

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

Methods parameter	IMLE	PMLE	DPMLE
$\hat{\beta}_0$	-82.952020	-35.875354	-27.697324
$\hat{\beta}_1$	8.0534710	0.6822198	0.5942665
$\hat{\beta}_2$	-0.6886301	0.5894145	0.4547526
$\hat{\beta}_3$	0.3142809	0.2907748	0.1033230
$\hat{\beta}_4$	7.4082146	3.1207516	2.6157121
MSE	0.5250163	0.07429295	0.0573744

لوحظ من الجدول (7) ان نتائج التطبيق للبيانات الحقيقية جاءت متوافقة مع الواقع الحقيقي للبيانات حيث اظهرت مقدرات الإمكان الأعظم الجزائية اقل متوسط مربعات الخطأ من مقدرات

(IMLE) هذا يعني ان تأثير مشكلة الفصل شبة التام والتعدد الخطي اكثر تفوق على البيانات في حين أظهرت مقدرات الإمكان الأعظم الجزائية المزدوجة DPMLE اقل متوسط مربعات الخطأ من طريقة مقدرات الإمكان الأعظم الجزائية وذلك لأنها تعالج مشكلة الفصل والتعدد الخطي وبفارق بسيط وبالتالي تعد طريقة المقدرات DPMLE الطريقة المثالية في معالجة مشكلة الفصل والتعدد الخطي لأنها تحقق اقل MSE من جميع طرائق التقدير الأخرى.

ان تفسير هذه النتائج يعطي صوره واضحة عن الإصابة بفقر الدم وما لها من تداعيات حيث وجد ان الهيموغلوبين (Hb) 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 المرصوصة واخيرا معلمة خلايا الدم البيضاء وهذا ما يتناسب مع واقع الإصابة بفقر الدم الذي تم التطرائق اليه في مقدمة هذا الفصل.

المبحث الرابع: الاستنتاجات والتوصيات

1. الاستنتاجات:

1- مقدرات الإمكان الاعظم تؤدي الى حل غير وحيد في حدود فضاء المعلمات في حالة وجود مشكلة الفصل في انموذج 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 الإمكان الأعظم الجزائية المزدوجة افضل من طريقة مقدرات الإمكان الأعظم الجزائية ومقدرات الإمكان الأعظم التكرارية في معالجة مشكلة الفصل والتعدد الخطي في تقدير معلمات انموذج الانحدار اللوجستي من خلال مقياس (MSE) للأنموذج.
4- مقدرات الامكان الأعظم التكرارية تمتلك اكبر قيم متوسط مربعات الخطأ (MSE) للأنموذج من طرائق التقدير الأخرى.
5- من خلال جدول تحليل البيانات في الجانب التطبيقي دراسة العوامل المؤثرة على مرض فقر الدم تم التوصل ان الهيموغلوبين (Hb) العامل الأول الذي يؤثر على الإصابة وهذا جاءَ متوافق مع رأي الأطباء.

2. التوصيات:

1- اعتماد طريقة مقدرات الإمكان الأعظم الجزائية المزدوجة في تقدير معلمات انموذج الانحدار اللوجستي في حالة وجود مشكلة الفصل والتعدد الخطي.
2- استعمال اساليب تكرارية أخرى في تقدير معلمات انموذج الانحدار اللوجستي 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 لذلك يجب استخدام البرامج الحديثة في تبويب وارشفة البيانات في المؤسسات الحكومية خصوصا المؤسسات الطبية.
4- على كل شخص اجراء الفحوصات الخاصة بتشخيص فقر الدم .

المصادر

1. البلداوي، تنسيم حسن (1996) م،"مقارنة تحليلية بين انموذج الانحدار اللوجستي ونماذج الدوال التميزية" أطروحة دكتورا في الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد.
2. الراوي، خاشع محمود (1978) م،"مدخل الى تحليل الانحدار"، جامعة الموصل.

3. الربيعي، عباس حسين، وجبار، علي مقيم (2010) م،"دراسة لبعض التغيرات الدموية والكيموحيوية في الأطفال المصابين بمرض الثالاسيميا في محافظة بابل"، مجلة كلية التربية الاساسية-جامعة بابل، العدد 2، الصفحة 310-317.

4. بيثون، نغم نافع (1992) م،"خواص قوة الاختبار وحدود الثقة لمعاملات الانموذج اللوجستي الخطي دراسة مقارنة "، رسالة ماجستير في الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد.

5. علي، مهدي كاظم، ورشيد، وسن سعيد (2013) م،"استجابات عدد من مكونات الدم والإدرار لعدو ١٠٠٠٠متر وفي فترة استعادة الشفاء لدى عدائي المسافات الطويلة"، مجلة كلية التربية الرياضية-جامعة بغداد، المجلد 25، العدد 2.الصفحة 31-52.

6. يحيى، مزاحم محمد، وعبدالله، محمود حمدان (2007) م،"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. *Albert, A & Anderson J. A (1984).“On the Existence of Maximum Likelihood Estimates in Logistic Regression Models”, Biometrika , Vol. 71, No. 1 (Apr., 1984), pp. 1-10.*

8. *Albert, A & Lesaffre, E (1986).“Multiple Group Logistic Discrimination”, Comp & Mths, Vol. 12A, No. 20, PP. 209-224, Printed in Great Britain.*

9. *Allison, P. D (1999).“Logistic Regression Using the SAS® System: Theory and Application”, Copyright © 1999 by SAS Institute Inc., Cary, NC, USA.*

10. *Allison, P. D (2008).“Convergence Failures in Logistic Regression”, University of Pennsylvania, Statistics and Data Analysis, Paper 360.*

11. *Bull, S. B & Mak, G & Greenwood, M. T (2002). “A modified score function estimator for multinomial logistic regression in small samples”, Computational Statistics & Data Analysis 39, pp. 57–74.*

12. *Heinze, G & Schemper, M (2002).“A solution to the problem of separation in logistic regression” Statist. Med. , 21:2409–2419 (DOI: 10.1002/sim.1047).*

13. *Konis, K (2007).“Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models”, A thesis submitted for the degree of Doctor of Philosophy in Statistics, Worcester College, University of Oxford.*

14. *LE Cessie, S & Van, J. C (1992).“Ridge Estimators in Logistic Regression”, Appl. Statist. , Vol. 41, No. 1, pp. 191-201.*

15. *Lesaffre, E & Marx, B. D (1993). “Collinearity in Generalized Linear Regression”, COMMUN. STATIST.-THEORY METH., Vol. 22, No. 7, pp. 1933-1952.*

16. *Magnac, T (2005).“Logit models of individual choices”, Université de Toulouse, prepared for the New Palgrave, First version.*

17. *Marx, B. D & Smith E. P (1990).“Weighted multicollinearity in logistic regression : diagnostics and biased estimation techniques with* *an example from lake acidification", journal Canadian des sciences aquatiques, volume 47, No. 6, pp. 1128-1135.*

18.*Shahmandi, M & Farmanesh, F & Gharahbeigi M (2013)."Data Analyzing by Attention to Weighted Multicollinearity in Logistic Regression Applicable in Industrial Data", British Journal of Applied Science & Technology3(4), PP. 748-763.*

19.*Shen, J & Gao, S (2008)."A Solution to Separation and Multicollinearity in Multiple Logistic Regression", Indiana University School of Medicine, Journal of Data Science* ***6****, PP. 515-531.*