المجلة العراقية للعلوم الإحصائية

http://stats.uomosul.edu.iq

مقارنة طريقتي تقدير معلمات انموذج انحدار لوجستي ثنائي

فرح هيثم فتحي و صفاء يونس الصفاوي

قسم الاحصاء والمعلوماتية ، كلية علوم الحاسوب والرياضيات ، جامعة الموصل ، الموصل ، العراق

معلومات النشر

تاريخ المقالة:
تم استلامه في 17 أب 2021
تم القبول في 24 ايلول 2021
متاح على الإنترنت في 1 كانون الاول 2021

الكلمات الدالة:
انموذج الانحدار اللوجستي الثنائي
تقدير معلمات
طريقة المربعات الصغرى الموزونة

المراسلة:
فرح هيثم فتحي
farah.csp140@student.uomosul.edu.iq

الخلاصة

يتناول هذا البحث أهم نماذج الانحدار غير الخطية الواسعة الاستخدام في نمذجة التطبيقات الاحصائية وهو انموذج الانحدار اللوجستي الثنائي (the binary logistic regression model)، وتم تقدير معلمات هذا الأنموذج باستخدام طرائق التقدير الاحصائية. شملت المقارنة بعض من طرائق التقدير المهمة وهي طرائق التقدير الاعتيادية التي تضمنت طريقة المربعات الصغرى الموزونة (Weighted Least Squares Method WLS)، وطريقة بيز(Bayes Method BM)، ومن اجل اختيار الطريقة الفضلى في التقدير وذلك من خلال أخذ عدد من النماذج واحجام عينات مختلفة وباستعمال المعيار الاحصائي متوسط مربعات الخطأ (MSE) لمقدرات الانموذج اللوجستي لغرض المقارنة بين أفضلية طرائق تقدير معلمات الأنموذج، وقد تم التوصل بشكل عام الى انه طريقة المربعات الصغرى الموزونة (WLS) هي الأفضل لامتلاكها اقل قيمة لمعيار المقارنة MSE مقارنة بالطرائق الاخرى. تم ايضا في الجانب العملي استخدام هاذين النموذجين لنمذجة البيانات الخاصة بالمصابين بأمراض السكري, الذين تم الحصول على بياناتهم من مستشفى ازادي الواقع في مدينة دهوك لعام 2021, و تم تقدير المعلمات باستخدام الطرائق الفضلى. تم مقارنة المرضى الذين لديهم استعداد للاصابة بمرض السكري مع الذين لم يصابو بمرض السكري للتوصل لمدى ملائمة الأنموذج في نمذجة هذا النوع من البيانات واستخلاص السبب الرئيسي لحالات حدوث مرض السكري وهو الانسولين، وكذلك دقة الطرق في تقدير معلمات الانموذج.

DOI: 10.33899/IQJOSS.2021.169971 , ©Authors, 2021, College of Computer Science and Mathematics, University of Mosul
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

1- مقدمة Introduction

يعد انموذج الانحدار اللوجستي الثنائي احد النماذج اللاخطية التي يمكنها وصف العلاقة بين متغير تابع ثنائي القيمة اي بمعنى ياخذ قيمتان هما الصفر لاحتمال عدم حدوث حدث معين والواحد الصحيح لاحتمال حدوث ذلك الحدث اما المتغيرات المستقلة فانها تاخذ قيم وصفية او كمية، وعليه تستعمل ستراتيجية تحويل الاستجابات الى فئتي او فئات المتغير التابع من نوع لوجت ووقتها فقط سوف تتحول العلاقات غير الخطية الى علاقات خطية بين متغير الاستجابة التابع والمتغيرات التوضيحية (المستقلة)، حيث تم الاعتماد على بعض طرائق التقدير ومنها طريقة تقدير المربعات الصغرى الموزونة (WLSE) وطريقة تقدير بيز (BE) من اهم الطرائق التي تطبق في تقدير معلماتها من اجل الحصول على قيم المعلمة المثلى اذ يتم الحصول على معادلات الامكان من الدرجة الاولى لمعلمات دالة الامكان من الانموذج اللوجستي الثنائي مما يزيد من احتمالية مشاهدة الفرد وتكون بذلك غير خطية وبالتالي لايمكننا الحصول على تحليل للقيم المثلى للمعلمات (Hosmer, & Sturdivant, 2013)(Groenwald & Mokgatlhe, 2005).

الهدف من البحث هو تقدير معلمات انموذج الانحدار اللوجستي الثنائي باستخدام طرق التقدير الكلاسيكية (الاعتيادية) (BE،WLSE)، مع المقارنة بين هذه الطرق باستعمال المعيار الاحصائي متوسط مربعات الخطأ (MSE)، للوصول الى افضل طريقة، ومن اجل بناء أنموذج الانحدار اللوجستي بالجانب التطبيقي.

2- انموذج الانحدار اللوجستي الثنائي (Binary logistic regression model)

يعد هذا النموذج من نماذج الانحدار اللاخطي وأن متغير الاستجابة (Y) يتبع توزيع برنولي (Bernoulli distribution) والذي يأخذ القيم (0) و (1)، (Hosmer & Sturdivant , 2013) اي أن متغير الاستجابة الفئوي (Y) له حالتين تتمثل الاولى (Y=1) عند وقوع حدث معين والثانية (Y=0) عند عدم وقوع ذلك الحدث (النجاح) هو (π (Xi)) بالاعتماد على قيم المتغيرات التوضيحية للمشاهدات واحتمال عدم وقوع الحدث (الفشل) هو (1- π (Xi)) وبذلك تكون دالة الكثافة الاحتمالية بالصيغة الاتية (Rodriguez,2007):-

$$P(Yi/Xi) = [\pi(Xi)]^{Yi}[1 - \pi(Xi)]^{1-Yi} \quad (1)$$

$$Yi = 0,1$$

$$P(Y = 0 /X_i) = 1 - \pi(X_i) \quad (2)$$

$$P(Y = 1 /X_i) = \pi(X_i)$$

اذ أن:

$\pi(Xi)$ تمثل التوقع الشرطي E(Y/X) (Conditional mean) لمتغير الاستجابة (Y) عند قيمة معينة ل (X) وتكون بحالتين:

$\pi(Xi)$ احتمال حدوث الاستجابة او احتمالة النجاح (success)

$1 - \pi(X_i)$ احتمال عدم حدوث الاستجابة او احتمال الفشل (failure)

ويكون توقع متغير الاستجابة يمثل احتمال النجاح

$$E(Yi) = \pi(Xi)$$

والتباين لمتغير الاستجابة يمثلنا احتمال حاصل ضرب النجاح مع احتمال الفشل

$$Var(Yi) = \pi(Xi)(1 - \pi(Xi))$$

وللتعبير عن انموذج الانحدار اللوجستي في حال احتوائه على متغير توضيحي واحد (X)بالصيغة الرياضية التالية (عبد الرزاق واخرون, 2016): -

$$\pi(X_1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}} = \frac{e^{(\beta_0+\beta_1 X_1)}}{1+e^{(\beta_0+\beta_1 X_1)}} \quad (3)$$

$$1 - \pi(X_1) = \frac{1}{1+e^{(\beta_0+\beta_1 X_1)}} \quad (4)$$

$$\pi(X_1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}} = \frac{e^{(\beta_0+\beta_1 X_1)}}{1+e^{(\beta_0+\beta_1 X_1)}} \quad (5)$$

$$1 - \pi(X_1) = \frac{1}{1+e^{(\beta_0+\beta_1 X_1)}} \quad (6)$$

اذ ان β_0 و β_1: معلمتا انموذج الانحدار اللوجستي.

4- طرائق تقدير معلمات الانموذج اللوجستي الثنائي (Methods of Estimating Parameters of The Binary Logistic Model)

سيتم التطرق الى طريقتين في هذا البحث:

1-طريقة المربعات الصغرى الموزونة Weighted Least Squares Method (WLSM)

2-طريقة بيز Bayes Method (BM)

4-1 طريقة المربعات الصغرى الموزونة Weighted Least Squares Method(WLSM)

انموذج الانحدار اللوجستي الثنائي هو حالة خاصة من النماذج الخطية العامه والتي ممكن ان تعتبر امتدادا للانموذج الخطي البسيط (Muller, 2004) .حيث يمكن كتابة انموذج الانحدار الخطي العام بالشكل التالي (Mccullagh & Nelder, 1983) :-

$$Z_i = \beta_\circ + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

$$Z_i = \beta_\circ + \sum_{i=1}^{p} \beta_j x_{ij} + \varepsilon_i \qquad , i = 1,2,\dots,n\ ;\ j = 1,2,\dots,p \tag{7}$$

قدرت معلمات انموذج الانحدار اللوجستي الثنائي باستخدام طريقة المربعات الصغرى الموزونة (WLS) للحصول على افضل تقدير.

$$W_i = \pi_i(1 - \pi_i) \tag{8}$$

حيث ان:

W_i : مصفوفة التباينات التي تمثل الاوزان المختارة للمستوى i

تقدر المعلمات $(\beta_\circ, \beta_1, \dots, \beta_p)$ حسب (WLS) بايجاد القيم التي تجعل الفرق بين استجابة المشاهدة والاستجابة المقدرة اقل مايمكن بمعنى تصغير مجموع مربعات الخطأ (SSE) (Hosmer, & Sturdivant , 2013) :-

$$SSE_i = \sum w_i (Z_i - \hat{\beta}_\circ - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_p x_{pi})^2 \tag{9}$$

وبهذا نكون قد وصلنا الى قيمة $\hat{\beta}$ المقدرة وكما يلي(Strutz, T. 2016) :-

$$\underline{\hat{\beta}} = (\dot{X}\mathrm{WX})^{-1}\,\dot{X}\mathrm{WZ} \tag{10}$$

2-4 طريقة بيز Bayes Method (BM)

ان اسلوب طريقة بيز تتكون من ثلاث اقسام هي التوزيع السابق والتوزيع اللاحق ودالة الامكان الاعظم ووفقا لهذا الاسلوب يمكن كتابة دالة التوزيع اللاحق لمعلمات الانموذجي اللوجستي وفق الصيغ التالية(Hoff,2009):

$$\pi(\beta/Y,Z) = f(Y/\beta,Z)\, f(\beta/Z) \tag{11}$$

$$\pi(\beta/Y,Z) = l(\beta)\pi(\beta) \tag{12}$$

فإن:

$\pi(\beta/Y,Z)$: تمثل دالة التوزيع اللاحق للمعلمات المراد تقديرها.

$l(B) = f(Y/\beta,Z)$: تمثل دالة الامكان الاعظم كما قد ذكرناها سابقا.

$\pi(\beta) = f(\beta/Z)$: تمثل دالة التوزيع السابق للمعلمات.

وتطبق معاينة جبس من خلال التوزيع الشرطي لـ (Z_i) وبالتكرار يمكن توليد بيانات عشوائية من التوزيع المنتظم حسب الصيغة الريلضية التالية Groenwald (& Mokgatlhe2005):

$$\beta_j/\ \beta_{j-p}, Y, Z \sim U(a_j, b_j) \qquad ; j = 0,1,\dots,P \tag{13}$$

اذ $U(a_j,b_j)$: تمثل التوزيع المنتظم للفترة (a) و (b) ان

$$a_j = \max\left[\frac{1}{X_{ip}}\log\left(\frac{Z_i}{1-Z_i}\right) - \sum_{j\neq p}^{N} BX_{ij}\right]$$

$$b_j = \min\left[\frac{1}{X_{ip}}\log\left(\frac{Z_i}{1-Z_i}\right) - \sum_{j\neq p}^{N} BX_{ij}\right]$$

ونحصل على قيم($\hat{\beta}$) التقديرية بأخذ متوسط العينة المسحوبة (المتولدة) من التوزيع اللاحق وحسب الصيغة الرياضية التالية (Chatterjee,2012) :

$$\overset{(i)}{\beta} \sim \hat{\pi}_i(\beta/Y) = \frac{1}{N}\sum_{j=1}^{N} P\left(\beta / Y, \overset{(i-1)}{Z_j}\right) \quad (14)$$

5- معيار المقارنة بين طرائق التقدير المستخدمة (Comparison Criterion Between Used Estimation Methods)

لقد تم الاعتماد على مقياس متوسط مربعات الخطأ (MSE) $Mean\ Square\ Error$ لكافة محاولات (تكرارات) الظاهرة قيد الدراسة مقسما على عدد المحاولات كمقياس للمقارنة بين طرائق التقدير المستخدمة والتي تشير الى مدى دقة التقدير وان قيمته تتناقص وتشير الى دقه وجودة المقدرات والهدف المراد تحقيقه هنا هو للحصول على المقدر $[(\hat{Y}_i) = \hat{\pi}(X_i)]$

في الصيغة (10-2) ويكون حسابة بالشكل التالي (Kubsch ,2021):-

$$MSE = \frac{1}{n-p}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \quad (15)$$

حيث أن: $\hat{Y}_i$: تمثل القيمة التقديرية للمشاهدات, n : تمثل حجم العينة, p : تمثل عدد المعلمات.

6- وصف البيانات التطبيقة

اعتمدت الدراسة على البيانات الطبية لمحافظة دهوك لسنة 2020 م الصادرة من مركز امراض السكر في مستشفى ازادي وتم استخدام خمس متغيرات منها متغير واحد تابع (متغير الاستجابة y) والذي يأخذ قيمتين هما الصفر للمصاب بمرض السكري (اي لدية استعداد للاصابة بالمرض) والواحد الصحيح للسليم (اي ليس لدية مرض السكري) وستة متغيرات توضيحية، وقد تم تعريف كل متغير في الجدول (1) والجدول (2) في الملحق كما واخذت المتغيرات التالية:

6-1متغير الاستجابة (Response Variable):

Yi: المصابين بأمراض السكري، i =0,1 ، $Y = 0$: لدية استعداد للاصابة ، $Y= 1$: ليس لدية استعداد للاصابة.

6-2 المتغيرات التوضيحية (variable explanatory):

$X1$: (الجنس) : متغير وصفي يمثل جنس المريض.

$X2$: (العمر): متغير كمي يمثل عمر المريض ويقاس بوحدة (السنة).

$X3$: (مؤشر كتلة الجسم): متغير كمي يمثل كتلة الجسم حيث يتم قياسة عن طريق المعادلة التالية:

الوزن (كغم) \ [الطول (م)] ^{2}

$X4$: (الكلوكوز): متغير كمي يمثل مستوى السكر بالدم حيث تتراوح النسب الطبيعية من (72-99) ووحدة قياسة mg/dl وتم قياسه بعد صيام ثمان ساعات.

$X5$: (الانسولين): متغير كمي يمثل مستوى الانسولين في الدم ووحدة قياسة mu/ml

$X6$: (مقاومة الانسولين): متغير كمي يمثل مقاومة الجسم للانسولين ويقاس عن طريق المعادلة التالية:

(الكلوكوز* الانسولين)/ 405. حيث أن النسب الطبيعية تتراوح من (0.5-1.4)، واذا زادت النسبة عن 1.9 فهذا يعني ان الشخص لدية استعداد للاصابة بالمرض.

7- تحليل النتائج التطبيقية (Analysis of Applied)

سيتم عرض نتائج الجانب التطبيقي وتحليلها ليتم قياس مدى ملائمة البيانات الحقيقية مع انموذج الانحدار اللوجستي الثنائي المقدر وذلك من خلال اجراء اختبارات خاصة بمقدرات الانموذج اللوجستي.

فقد تم التوصل الى كافة النتائج التطبيقية باستخدام برنامج (MATLAB) ، الجدول 1 يمثل نتائج تحليل طرائق الانحدار اللوجستي الثنائي

جدول (1) تقدير معلمات انموذج الانحدار اللوجستي الثنائي بكلا الطريقتين

Estimated Parameters	Binary Logistic Regression Analysis	
	WLS	BM
β_0	0.503	0.00163
β_1	0.07	0.0023
β_2	0.004	0.0018
β_3	−0.007	0.0024
β_4	0.002	0.0022
β_5	0.0015	0.00229
β_6	0.0005	0.00284

نلاحظ من الجدول (1) قيم مقدرات انموذج الانحدار اللوجستي الثنائي من خلال تحديد متجه مقدر المعلمات $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ \dots \ , \hat{\beta}_{10})$ ويتم بتعويضها مع المتغيرات التوضيحية لنحصل على مقدرات انموذج الانحدار اللوجستي الثنائي وبدورها تقلل مجموع مربعات الخطأ الى أصغر مايمكن حسب طرق التقدير طريقة المربعات الصغرى الموزونة (WLS)، طريقة بيز (BM) والنتائج موضحة في الجدول اعلاه. والجدول 2 يوضح نتائج معيار المقارنة الاحصائي MSE

جدول (2) متوسط مربعات الخطأ (MSE)

الطرق	MSE
WLS	0.2462
BM	0.3485

نلاحظ من الجدول (2) ان طريقة WLS تعطي اقل نسبة خطا MSE .

8- معيار جدول التصنيف (Classification Table Criterion (CTC))

يظهر جدول (3) التصنيف امكانية ملائمة الانموذج وهذه المقارنة تتمثل بمقارنة النتائج المتوقعة من خلال الانموذج مع نتائج المشاهدة ومن خلال التصنيف للقيم المتوقعة على اساس القيم الاحتمالية المعينة.

جدول (3) تصنيف بيانات

حالة المريض		التنبؤ		
		$\hat{\pi} < 0.5$ حدوث المرض(0)	$\hat{\pi} \geq 0.5$ عدم حدوث المرض (1)	المجموع
المشاهدة	حدوث المرض positive Y = 0	d True positive	c False negative	a+c
	عدم حدوث المرض Negative Y = 1	b False positive	a True negative	d+b
	المجموع	d+b	c+a	a+b+c+d

تمثل قيمة d قيمة true positive: اي بمعنى انه تمت مشاهدة المريض مصاب (Positive) وان التنبؤ وفق الطريقة تعطي positive ايضا. اي من خلال البيانات نخبر المريض بانه مصاب وهو بالفعل مصاب. فهذه القيمة تعتبر صحيحة

تمثل قيمة a قيمة true negative: اي بمعنى انه تمت مشاهدة المريض غير مصاب negative وان التنبؤ وفق الطريقة تعطي negative ايضا. اي من خلال البيانات نخبر المريض بانه غير مصاب وهو بالفعل غير مصاب. فهي قيمة صحيحة.

تمثل قيمة c قيمة false negative : اي بمعنى انه تمت مشاهدة المريض بانه مصاب positive والتنبؤ وفق الطريقة اعطى نتيجة negative. اي من خلال البيانات نخبر المريض بانه غير مصاب ولكنه مصاب بالمرض. فهي نتيجة غير صحيحة وقد تكون لها عواقب.

تمثل قيمة b قيمة false positive: اي ان تخبر المريض بانه مصاب وهو غير مصاب بالحقيقة. هذه النتيجة قد تكون مؤثرة ولكن ليست بخطورة القيمة c.

ولمعرفة اداء الطرق نقوم بحساب القيم التالية:

1- **دقة النموذج** Accuracy (AC)

وهي تعرف بان تكون القيم المتوقعة صحيحة نسبة الى كافة التوقعات

$$AC = = \frac{a+d}{a+d+c+b} = = \frac{\text{True positive+ True negative}}{all\ the\ samples}$$

2- **معدل الخطاء** Error Rate (ER)

وهي معدل القيم المتوقعة الخاطئة الى بقية القيم

$$ER = = \frac{c+b}{a+d+c+b} = = \frac{False\ \text{positive+ false negative}}{all\ the\ samples}$$

3- **حساسية النموذج** Sensitivity (SE)

وهي توقع ان يكون التصنيف المتوقع موجبا نسبة للحالة التي تكون موجبا

$$SE = \frac{a}{a+c}$$

9- النتائج

9-1 نتائج طريقة (WLS)

نلاحظ في الجدول (4) ان انموذج الانحدار اللوجستي الثنائي قام بتصنيف المتغيرات المعنوية من خلال ايجاد قيم انموذج الانحدار اللوجستي الثنائي , حيث صنف الانموذج 13 من اصل 40 حالة تصنيفا صحيحا لتنبؤ وجود المرض عند وجود المرض بالفعل. و 27 حالة تم التنبؤ بها بعدم وجود المرض عند وجوده. اي بلغت نسبة التصنيف الصحيح بوجود المرض حولي 35% وهي نسبة منخفضة جدا. ومن جهة اخرى تنبؤات الطريقة ب 359 من 360 حالة صحيحة بعدم وجود المرض اي بنسبة 99%. و بما انه من الضروري ان نتنبا بوجود المرض عند وجوده بالفعل فان هذه الطريقة قد تكون غير محبذة لكشف وجود المرض.

في الجدول (5) نلاحظ ان نسبة الدقة بحدوث المرض قليلة ونسبة الصحيحة لتنبوء بوجود المرض ما قليلة وهي حوالي 88%. نسبة حساسية النموذج ونسبة التصنيف الصحيح مقبولة نوعا ما.

جدول (4) تصنيف بيانات العينة باستخدام طريقة (WLS)

حالة المريض		التنبؤ		
		حدوث المرض(0) $\hat{\pi} < 0.5$	عدم حدوث المرض (1) $\hat{\pi} \geq 0.5$	المجموع
المشاهد	حدوث المرض Y = 0	13	27	40
	عدم حدوث المرض Y = 1	1	359	360
	المجموع	14	394	400
دقة النموذج		AC=%93		
حساسية النموذج		SE=%93		
معدل الخطاء		ER=%7		

9-2 نتائج طريقة (BM)

جدول (5) تصنيف بيانات العينة باستخدام طريقة (BM)

حالة المريض		التنبؤ		
		حدوث المرض(0) $\hat{\pi} < 0.5$	عدم حدوث المرض (1) $\hat{\pi} \geq 0.5$	المجموع
المشاهدة	حدوث المرض0 = Y	21	19	40
	عدم حدوث المرض1 = Y	28	332	360
	المجموع	49	351	400
دقة النموذج		AC=%88.5		
حساسية النموذج		SE=%94		
معدل الخطاء		ER= %11.5		

10-الترابط بين البيانات correlation وتحليل الارتباط

في الجدول (5) الذي ولدناه من برنامج (IBM spss) نلاحظ الترابط (Pearson Correlation) بين قيم x الستة مع بعضها و مع قيمة ال y. كما نلاحظ قيمة ال p-value المتمثلة بقيمة Sig (2-tailed) في الجدول.

نلاحظ في الجدول ان قيمة p-value اكبر من 0.05 بين y والادخال x6 مما يخبرنا انه لا توجد علاقة بين ثبوت الاصابة بالمرض و قيمة مقاومة الجسم للانسولين. كما توضح قيمة Pearson correlation علاقة ضعيفة بين y و قيمة x6 . مما يعني انه ممكن ان نتخلى عن x6 في الحسابات في الطرق الاربعة تحت الاختبار.

10-1 الفحوصات الجراحية .

من المرغوب جدا ان يتم الفحص بابسط الطرق لمعرفه حالة المريض. هنا بعض الادخالات التي اخذناه بالاختبار تحتاج الى تدخل جراحي نوعا ما كما يتطلب فحصه وجود اجهزه مختبرية قد تكون مكلفة. مثال على ذلك الادخال x4 وهو يمثل ((الكلوكوز): متغير كمي يمثل مستوى السكر بالدم). ولفحص هذا الكلوكوز نحتاج جهاز خاص في المختبر يدعى spectrophotometer . صورة هذا الجهاز موجودة في الملحق B. المتغير x5 الذي يمثل مستوى الانسولين يمكن ان ياخذ في البيت بواسطة جهاز محمول ورخيص يدعى Glucometer وهو موجود في الصيدليات ولا يحتاج الى جراحة فقط وخزة صغيرة ويعطي النتيجة.

10-2 النتائج بدون وجود x4 و x6

رغم ان x4 عامل مهم لكشف المرض كما موضح في الجدول (6) من خلال عامل ترابط عالي. الا اننا ممكن ان نسغني عنه لغرض تقليل الكلفة . كما انه يمكن الاستغناء عن x6 لان عامل الارتباط قليل مع ال y.

لذلك سنقوم فحص الطرقتين بوجود اربع ادخالات فقط :x1,x2,x3, x5 والتي تمثل الجنس العمر الوزن والانسولين.

		y	x1	x2	x3	x4	x5	x6
y	Pearson Correlation	1	.179**	.298**	-.289**	.350**	.206**	.019
	Sig. (2-tailed)		.000	.000	.000	.000	.000	0.699
	Sum of Squares and Cross-products	36.000	10.700	641.300	-382.600	1983.877	866.400	11.674
	Covariance	.090	.027	1.607	-.959	4.972	2.171	.029
	N	400	400	400	400	400	400	400

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

11- النتائج بوجود البيانات المنزلية فقط

جدول (7) تصنيف بيانات العينة باستخدام طريقة (WLS) بالطريقة غير الجراحية

حالة المريض x1,x2,x3,x5		التنبؤ: حدوث المرض(0) $\hat{\pi} < 0.5$	التنبؤ: عدم حدوث المرض (1) $\hat{\pi} \geq 0.5$	المجموع
المشاهدة	Y = 0 حدوث المرض	8	32	40
	Y = 1 عدم حدوث المرض	4	356	360
	المجموع	4	388	400
دقة النموذج		AC=%91		
حساسية النموذج		SE=%91.3		
معدل الخطاء		ER=%9		

نلاحظ في الجدول (7) ان انموذج الانحدار اللوجستي الثنائي قام بتصنيف المتغيرات المعنوية من خلال ايجاد قيم انموذج الانحدار اللوجستي الثنائي صنف الانموذج باستخدام الادخالات الاجراحية الاربعة : x1,x2,x3,x5. حيث نلاحظ عند مقارنته مع الجدول (3) ان الانحدار اللوجستي الثنائي اعطى نتائج جيده حيث زادت نسبة الخطاء من 7 الى 9 بالمئة فقط.

جدول (8) تصنيف بيانات العينة باستخدام طريقة (BM)

حالة المريض X1,x2,x3,x5		التنبؤ: حدوث المرض(0) $\hat{\pi} < 0.5$	التنبؤ: عدم حدوث المرض (1) $\hat{\pi} \geq 0.5$	المجموع
المشاهدة	حدوث المرض Y = 0	40	0	40
	عدم حدوث المرض Y = 1	308	51	360
	المجموع	348	51	400
دقة النموذج		AC=%23		
حساسية النموذج		SE=%100		
معدل الخطاء		ER= %77		

في الجدول (8) وضح انه لا يمكن استخدام طريقة BM بالطريقة غير الجراحية, اي عند استخدام BM يجب ان تتم النتائج في المختبر. لان الخطاء زاد من 11 الى 77% وهذا يدل على فشل الطريقة.

ملحق A

عينة من البيانات المستخدمة : y=0 اي ان المريض مصاب

Y	x1	x2	x3	x4	x5	x6
0	1	30	46	47.78784	171	8.84566
1	1	51	19	37.75323	186	0.620592
1	1	34	46	139.5184	77	13.00878
1	1	44	30	161.1248	114	2.123871
1	1	35	25	125.0661	114	8.92781
1	2	60	53	118.2466	147	4.928079
1	2	70	37	127.9066	168	4.994892

الملحق B

12- الاستنتاجات

1- اثبتت طريقة المربعات الصغرى الموزونة (WLS) كفاءتها على كافة طرائق التقدير الاعتيادية إذ حققت المرتبة الاولى من الأفضلية في تقدير المعلمات لانموذج الانحدار اللوجستي الثنائي ولحصوله على اقل قيمة ل MSE.

2- تناقص قيمة (MSE) لمقدرات الأنموذج اللوجستي باستخدام كافة طرائق التقدير الاعتيادية بازدياد حجم العينة، وهذا مايطابق النظرية الاحصائية.

3- نستنتج انه ممكن ان نتنبأ بوجود مرض السكري عند مريض بنسبة خطاء معقولة من خلال معرفة جنسه و عمره ووزنه ونسبة الانسولين غيرالجراحية (كلها ممكن ان تقاس في البيت) باستخدام طريقة WLE. طريقة BM لايكن استخدامها الى من خلال نتائج الكلوكوز التي يجب ان تتم في المختبر.

4- عند وجود نتائج مختبرية من الافضل استخدام BM والا فيمكن استخدام طريقة WLE

التوصيات

1- استعمال متغيرات اخرى غير التي ذكرت في البحث مثل متغير ببتايد السي (C-peptide) أو هرمون الرزستين (Resistin) أو هرمون اللبتين (Leptin) وغيرها من العوامل التي تؤثر على مرض السكري.

2- اجراء الطرائق التي تم عرضها باتباع منهجية الخوارزمية الجينية في مجالات اخرى من غير الطبية مثلا تكون بيانات سلوكية أو اجتماعية.

3- استخدام طرق ال machine learning في عمليات تنبؤ البينات.

Refference

1- Abdul-Razzaq, Muhammad Sadiq and Zaalan, Risan Abdul-Imam, (2016) "Using the logistic regression method to analyze the effect of psychological stress on the incidence of blood pressure", (Basra, Gulf Economics Journal, pages 48-66

2- Talib, H. R., & Mazloum, S. A. (2020). The Use of Binary Logistic Regression Method to Analyze The Factors Affecting Heart Disease Deaths: An Applied Study on A Sample of Patients in Dhi Qar Governorate. Al-Rafidain University College For Sciences, (46).

3-Azen, R., & Walker, C. M. (2021). *Categorical data analysis for the behavioral and social sciences*. Routledge.

4-Akkus ,Ö. , Demir , E. , (2016), " Comparison Som Classical And Meta-Heuristic Optimazation Techniques in The Estimation Of The Logit Model Parameters", ,(2017) , "The Factors Affecting Eye Patients (Cataract) In Jordan by Using the Logistic Regression Model", MAS, pp. 38-42.

5- Strutz, T. (2016). Data fitting and uncertainty. A practical introduction to weighted least squares and beyond.

6-Chatterjee, S., Hadi, A., (2012) ,"Regression Analysis By Example", John Wiley ,INC.

7-Demir , E. , Akkus , Ö., (2015), " An Introductory Study on How the Genetic Algorithm Works in the Parameter Estimation of Binary Logit Model", IJS:BAR, pp.162-180.

8-Gelman, A. & et al., (2014)," Bayesian Data Analysis", Texts in Statistical Science , CRC , LC , Chapman and Hall Book.

9-Kubsch, M., Stamer, I., Steiner, M., Neumann, K., & Parchmann, I. (2021). Beyond p-values: Using Bayesian Data Analysis in Science Education Research. Practical Assessment, Research, and Evaluation, 26(1), 4.

10- Groenewald, P. C., & Mokgatlhe, L. (2005). Bayesian computation for logistic regression. Computational statistics & data analysis, 48(4), 857-868.

11- Acquah, H. D. G. (2013). Bayesian logistic regression modelling via Markov Chain Monte Carlo algorithm.

12- Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.

13-Hosmer, D., Lemeshow, S. & Sturdivant , R. ,(2013)," Applied Logistic Regression", 3rd edition ,New York: willey,WSIPS, http : // ihmsi.org.

14- McCullagh, P., & Nelder, J. A. (2019). Generalized linear models. Routledge.

Comparison of Two Methods for Estimating Parameters of the Model Binary Logistic Regression

Farah Haitham Fathi Safa Younis Al-Safawi

College of Computer Science and Mathematics, University of Mosul

Abstract :In this paper, we deal with one of the most important nonlinear regression models widely used in modeling statistical applications, which is the binary logistic regression model, and then estimating the parameters of this model using statistical estimation methods. In this paper, some ordinary estimation methods are employed to fit the estimation of the parameters of this type of non-linear regression model, and then we compare these estimation methods. Further, the comparison includes some of the important estimation methods, which are the ordinary estimation methods that included the Weighted Least Squares Method (WLS), and the Bayes Method (BM). In order to choose the best method for estimating, by taking several models and different sample sizes and using the statistical standard mean error squares (MSE) for the logistic model estimations for the purpose of comparison. Among the preferred methods for estimating model parameters, and it was generally concluded that the WLS method provides the MSE of estimators compared to the other methods. On the practical side, this model was also used to model data for people with diabetes (that we obtained from Azadi hospital in Duhok) and to estimate parameters using the best methods, and it was reached by comparing patients with diabetes. A census of diabetes with those who did not have diabetes with the appropriateness of the model in modeling this type of data and extracting the main cause of diabetes incidence, which is insulin, as well as the accuracy of the methods in estimating the model parameters.

Keyword: Binary logistic regression model Estimate parameters Weighted least squares method