مقارنة بعض المقدرات الحصينة لتقدير معلمات أنموذج الانحدار اللوجستي ثنائي الاستجابة بأستعمال المحاكاة

أ.م .ايمان حسن احمد / كلية الادارة والاقتصاد / جامعة بغداد
الباحث / ضمياء حامد شهاب

تاريخ التقديم:2017/4/9
تاريخ القبول:2017/5/28

المستخلص

يعد أنموذج الانحدار اللوجستي من نماذج الانحدار اللاخطية الذي يهدف الى الحصول على مقدرات تمتلك كفاءة عالية، والذي يأخذ طابعاً اكثر تقدماً في عملية التحليل الاحصائي لكونه من النماذج الملائمة للبيانات الثنائية (Binary Data).

ومن بين المشاكل التي تظهر نتيجة استخدام بعض الطرائق الاحصائية هي مشكلة عدم تحقق بعض الشروط المطلوبة او كلها مثل مشكلة وجود القيم الشاذة بين البيانات حيث تكون بيانات الظاهرة المدروسة ملوثة, اي وجود بعض المشاهدات تنحرف وبشكل ملحوظ عن المشاهدات الاخرى تدعى بالشواذ.

ومن هنا جاء هدف هذا البحث لتقدير معلمات أنموذج الانحدار الوجستي من خلال دراسة بعض طرائق التقدير الحصينة المتمثلة بطريقة مقدرات الامكان الاعظم الموزونة الحصينة (WMLE), طريقة مقدرات المسافة التربيعية الحصينة (QDE)، وقد تم استخدام اسلوب المحاكاة للمقارنة بين الطريقتين بأختلاف احجام العينات ونسب التلوث المختلفة من خلال متوسط مربعات الخطأ(MSE) للانموذج للوصول الى الطريقة الأفضل في تقدير المعلمات.

تم التوصل من خلال هذا البحث الى افضلية طريقة (WMLE(W_1)) في تقدير معلمات انموذج الانحدار اللوجستي ثنائي الاستجابة بأختلاف حجوم العينات.

المصطلحات الرئيسة للبحث: الانحدار اللوجستي، البيانات الثنائية، مقدرات الامكان الاعظم الحصينة الموزونة (WMLE) ، مقدرات المسافة التربيعية الحصينة (QDE)، المشاهدات الشاذة.

مجلة العلوم
الاقتصادية والإدارية
العدد 102 المجلد 24
الصفحات 423ـ 440

* بحث مستل من رسالة ماجستير

1-1 المقدمة Introduction

يعد أنموذج الانحدار اللوجستي من نماذج الانحدار اللاخطية القابلة للتحويل الى نماذج خطية، حيث يتم من خلاله توضيح العلاقة بين المتغير التابع ثنائي الاستجابة والمتغيرات التوضحية (التفسيرية).

يستخدم أنموذج الانحدار اللوجستي في المجالات المتعلقة بعلوم الحياة والعلوم الزراعية والطبية والاجتماعية اي بشكل عام في الدراسات ذات الطابع التجريبي، نظراً الى خصوصية البيانات في هذه المجالات والتي هي غالباً ما تكون من النوع (Binary) ثنائية الاستجابة ، لكونه من النماذج الملائمة للبيانات الثنائية (Binary Data).

ان طرائق التقدير التقليدية لمعلمات أنموذج الانحدار اللوجستي عند تحليل البيانات الثنائية (Binary Data Response) تكون ضعيفة في معالجة المشكلات بين البيانات، ومن بين احدى تلك المشاكل وجود القيم الشاذة، وأن وجودها ضمن مجموعة البيانات يؤثر وبشكل كبير في نتائج التحليل الاحصائي، من هنا دعت الحاجة الى أهمية استخدام طرائق اخرى لمعالجة تلك المشاكل وتكون اكثر كفاءة في التقدير تدعى (بطرائق التقدير الحصينة) (Robust estimation methods) تتمثل بكونها طرائق بديلة عن الطرائق التقليدية حيث تتصف بأنها قليلة الحساسية تجاه الشواذ أي لا تتأثر كثيرا بوجود القيم الشاذه، اذ يتم الحصول من خلالها على مقدرات حصينة تمتلك كفاءة عالية.

أن ألاهتمام المتزايد من قبل الباحثين بطرائق التقدير الحصينة دفع الكثير منهم الى اقتراح العديد من التقديرات والدوال لهذه الطرائق، وعلى الرغم من اختلاف صيغ هذه الطرائق الا ان اغلبها تشترك في هدف واحد هو استخدام أسلوب الموازنة بين المشاهدات من خلال اعطاء أوزان للمشاهدات الشاذة وبوزن أقل من التي تقرن مع بقية المشاهدات وان الغرض من إعطاء الوزن لتلك المشاهدات هو للتقليل من تأثيرها.

2-1 الاستعراض المرجعي The Review of Literature

في عام (2001) م قام الباحثان (Flores and Garrido) [9] بتقدير معلمات أنموذج الانحدار اللوجستي في حالة كون الاستجابة (ثنائية Binary ، متعددة Maltinomial) بأستعمال طريقة مقدرات الامكان الأعظم الاعتيادية (MLE) وطريقة مقدرات المسافة التربيعية الحصينة (QDE) في حالة وجود وعدم وجود القيم الشاذة في البيانات وبافتراض نسب تلوث مختلفة وتمت المقارنة بين الطريقتين وتم التوصل الى ان طريقة (QDE) هي الأفضل في حالة وجود القيم الشاذة وذات ألاداء القريب من طريقة (MLE) في حالة عدم وجودها.

في عام (2005) م قام الباحث (Simeckova)[17] بتقدير معلمات أنموذج الانحدار اللوجستي مستخدماً طريقة مقدرات الامكان الاعظم(MLE) وطريقة (WMLE) وبدوال وزن مختلفة وبأفتراض نسب تلوث مختلفة وبأستخدام اسلوب المحاكاة ومن خلال المؤشر الاحصائي (MSE) تم التوصل الى افضلية طريقة (WMLE) في تقدير معلمات الانموذج.

في عام (2009) م درس الباحث (حسين)[4] المقارنة بين مقدرات الامكان الاعظم الموزونة الحصينة (WMLE) مع طرائق حصينة أخرى (مقدرات طريقة M، مقدرات طريقةLP) لأنموذج الانحدار اللوجستي وقد تم استخدام أسلوب المحاكاة للمقارنة بين طرائق التقدير المدروسة في حالة تلوث بيانات الظاهرة المدروسة وعلى افتراض ثلاثة مستويات للتلوث (30%،10%،0%) وبالاعتماد على المقاييس الاحصائية للتحيز (Bias) ومتوسط مربعات الخطأ (MSE) للمعلمات وللأنموذج توصل الباحث الى افضلية طريقة (WMLE) مقارنة مع بقية طرائق التقدير المدروسة في تقدير معلمات أنموذج الانحدار اللوجستي.

3-1 هدف البحث Object of research

يهدف البحث الى تقدير معلمات أنموذج الانحدار اللوجستي بأستعمال طريقتي تقدير حصينة والمتمثلة بطريقة مقدرات الامكان الاعظم الموزونة (WMLE) وطريقة مقـــدرات المسافة التربيعية الحصينة (QDE)، في حالة وجود تلوث البيانات (اي احتوائها على قيم شاذة) وعدم وجود تلوث آخذين بنظر العناية نسب مختلفة للتلوث وبأحجام عينات مختلفة، ومن ثم المقارنة بين طريقتي التقدير من خلال معيار المقارنة متوسط مربعات الخطأ (MSE) للوصول الطريقة الأفضل في التقدير وللحصول على تقديرات كفوءة لمعلمات الانموذج.

1-4 مشكلة البحث Problem of the Research

ان استعمال طـرائق التقدير الاعتيادية في تقدير معلمات أنمـوذج الانحـدار اللوجستـي والذي يكون فيه متغير الاستجابة من النوع الثنائي (Binary) تكون ضعيفة في معالجة المشكلات بين البيانات ومن بين احدى تلك المشاكل وجود القيم الشاذة، وأن ظهورها ضمن البيانات يعد غير مرغوب فيه لانها تسبب صعوبة في محاولة لتمثل المجتمع ومن ثم تؤثر في دقة النتائج المتوخاة من عملية التقدير، من هنا دعت الحاجة الى أهمية استخدام طرائق اخرى لمعالجة تلك المشاكل وتكون اكثر كفاءة في التقدير تدعى (بطرائق التقدير الحصينة) (Robust estimation methods) حيث تتصف بأنها قليلة الحساسية تجاه الشواذ أي لا تتأثر كثيرا بوجود القيم الشاذة.

2-الجانب النظري

1-2 أنموذج الانحدار اللوجستي Logistic Regression Model

يعرف أنموذج الانحدار اللوجستي بأنه احد نماذج الانحدار اللأخطية والذي تكون فيه العلاقة بين المتغير التابع(y) متغير الاستجابة والمتغيرات التوضيحية ($x_1, x_2, x_3 \ldots\ldots, x_k$) غير خطية ، فضلا عن كون تقديرات المعلمات وفق هذا النموذج تعد مقبولة في ظل غياب بعض القيود المفروضة على نماذج الانحدار الخطي . [10]

حيث يهتم هذا الانموذج بتحليل البيانات التي يكون فيها متغير الاستجابة (y)من النوع المتقطع ويأخذ أحدى القيمتين (1,0) أما النجاح (Success) بأحتمال (p_i) او الفشل(Failure) بأحتمال (1- p_i) لذلك يكون المتغير (y) يتوزع توزيع برنولي ($Ber \sim (1, p_i)$).

ومن ثم فأن دالة الكثافة الاحتمالية تكون وفق الصيغة الآتية[10]:

$$P_r(Y = 1) = p_i^{y_i} (1 - p_i)^{1-y_i} \qquad \ldots.(1)$$

$$i = 1,2, \ldots\ldots, m$$

$$y_i = 0,1$$

أذ أن :

y_i متغير تابع ثنائي الاستجابة (0,1)

p_i أحتمال حدوث ألاستجابة عندما y_i =1

1-p_i أحتمال عدم حدوث الاستجابة عندما y_i =0

لذلك فأن توقع المتغير y_i يمثل أحتمال حدوث الاستجابة (p_i) وكالآتي:

$$E(y_i) = P_r(Y = 1) = p_i \qquad \ldots.(2)$$

أما تباين المتغيرy_i بالنسبة لتوزيع برنولي كالآتي:

$$v(y_i) = E(y)^2 - [E(y)]^2$$

$$v(y_i) = p_i(1 - p_i) \qquad \ldots.(3)$$

يمكن التعبير عن أنموذج الانحدار اللوجستي بالصيغة الاتية :

$$y_i = p_i + \varepsilon_i \qquad \ldots.(4)$$

حيث أن

p_i تمثل دالة الانحدار اللوجستي أو تمثل دالة الاستجابة اللوجستية

εi يمثل الخطأ العشوائي حيث يكون له متوسط يساوي صفر .

اما تباين حد الخطأ العشوائي فأنه يكون مساوياً الى تباين المتغير y_i ثنائي الاستجابة$p_i(1-p_i)$

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

$$p_i = \frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}} \quad \ldots..(5)$$

نلاحظ من المعادلة (5) ان شكل العلاقة بين المتغيرات التوضيحية (x_{ij}) واحتمال الاستجابة p_i لايمكن ان يكون خطياً وأنما تأخذ شكلاً منحنياً أي على شكل حرف (S). [6]

$$1-p_i = 1-\frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}}$$

$$1-p_i = \frac{1}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}} \quad \ldots..(6)$$

2-2 التحويل الخطي لدالة الانحدار اللوجستي

Linear Transformation to the Logistic Regression Function

يعد هذا النموذج من النماذج القابلة للتحويل الى نماذج خطية بالنظر لكون العلاقة بين المتغيرات التوضيحية (x_{ij}) وأحتمال حدوث الاستجابة (p_i) علاقة غير خطية اي على شكل أنحناءات حيث يميل العديد من الاحصائين الى أزالة انحناءات دالة الانحدار اللوجستي وذلك من خلال أجراء تحويل دالة اللوجت (logit function) حيث أن وجود هذه الانحناءات لها تأثير سلبي على خصائص مقدرات المعلمات.

ومن خلال أيجاد تحويلة لوغارتمية لتحويل العلاقة بين المتغيرات التوضيحية واحتمال حدوث الاستجابة (p_i) الى علاقة بصيغة خطية وتسمى نسبة الافضلية $\frac{p_i}{1-p_i}$ ، وبأخذ اللوغارتم للنسبة $\frac{p_i}{1-p_i}$ فأنها تسمى لوغارتم نسبة الافضلية وعليه فأن مجال قيمه تصبح محصورة بين $(-\infty, +\infty)$. [14]

وعليه فأن التحويل الخطي لدالة اللوجستي هو:

$$\ln\frac{p_i}{1-p_i} = \beta_0+\sum_{j=1}^{k}\beta_j x_{ij} = x'_i\beta = Z \quad \ldots...(7)$$

حيث أن

$\beta = \beta_0, \beta_1, \beta_2, \ldots \ldots, \beta_j$ يمثل متجه من المعلمات

$$x'_i = [1, x_{i1}, x_{i2}, \ldots \ldots, x_{ik}]$$

$$i = 1,2, \ldots \ldots, m$$

x'_i يمثل متجه من المتغيرات التوضيحية بدرجة (1*k)

وأن $x_{i0}=1$ وهو جزء من مصفوفة المتغيرات التوضيحية X

(Z) تمثل العلاقة الخطية الناتجة من أخذ اللوغارتم للنسبة $\frac{p_i}{1-p_i}$

ومن ثم فأن توقع Z_i يكون كالاتي:

$$E(Z_i) = x'_i\beta \qquad \ldots..(8)$$

أما تباين Z_i يكون وفق الصيغة الآتية:

$$var(Z_i) = \frac{1}{n_i p_i(1-p_i)} = [n_i p_i(1-p_i)]^{-1} \qquad \ldots..(9)$$

أذن فأن المقدار (Z_i) الذي هو التحويل الخطي لدالة اللوجستي يتوزع توزيعاً طبيعياً تقاربياً (Asymptotically Normal) بمتوسط $(x'_i\beta)$ وتباين مقداره $[n_i p_i(1-p_i)]^{-1}$ [6]

أي أن :

$$Z_i \sim Asym.\ N\ [x'_i\beta, [n_i p_i(1-p_i)]^{-1}]$$

3-2 طرائق التقدير Methods of Estimation

نستعرض في هذا البحث بعض طرائق تقدير معلمات أنموذج الانحدار اللوجستي والمتمثلة بطرائق التقدير الحصينة بهدف ايجاد مقدرات لمعلمات النموذج المدروس تتصف هذه المقدرات بصفات جيدة تؤهلها لتكوين نموذج تقديري يتم الاعتماد عليه في أغراض مختلفة والوصول الى نتائج اكثر دقة، ومن هذه الطرائق هي:

1-3-2 مقدرات الامكان الاعظم الموزونة Weighted Maximums Likelihood Estimates

تعد طريقة (WMLE) من الطرائق الحصينة في التقدير، واقترح الباحثان (Carroll and Pederson) في عام (1993) هذه الطريقة ، وأن فكرة هذه الطريقة تقوم على اساس تحوير طريقة الامكان الاعظم (MLE) من خلال اجراء تحويل لطريقة (MLE) الى مقدر حصين اكثر كفاءة من خلال اعطاء المشاهدات اوزان للتقليل من تأثير المشاهدات الشاذة، وأن هذه الطريقة تعتمد على دالة الكثافة الاحتمالية للمتغيرات العشوائية والذي يتبع توزيع برنولي وبحسب الصيغة الآتية[5]:

$$P(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i} \qquad \ldots..\ (10)$$

وبأخذ اللوغارتم لطرفي المعادلة (10) نحصل على الصيغة الآتية[17]:

$$\log P(Y_i = y_i) = y_i \log p_i + (1-y_i)\log(1-p_i) \qquad \ldots\ldots(11)$$

وبالنظر لوجود علاقة بين المتغيرات التوضيحية (x_{ij}) واحتمال الاستجابة p_i كما موضح في المعادلتين (5) و (6) يتم التعويض عنهما في المعادلة (11) نحصل على الآتي:

$$\log P(Y_i = y_i) = y_i \log\left[\frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}}\right] + (1-y_i)\log\left[\frac{1}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}}}\right] \qquad \ldots.(12)$$

حيث أن التقدير في المعادلة (12) يتطلب اتباع الاسلوب الذي تعتمده طريقة (WMLE) وهو تصغير المقدار وكالآتي[19]:

$$\min \sum_{j=1}^{m} W_i L_i(\beta) \qquad \ldots..(13)$$

حيث أن

$$L_i(\beta) = \log P(Y_i = y_i)$$

$L_i(\beta)$ يمثل لوغارتم الدالة في المعادلة (11)

W_i تمثل دالة الوزن

ولغرض الحصول على مقدرات (WMLE) والتي تصغر المقدار في المعادلة المذكورة آنفاً (13) يتم استخدام طريقة المربعات الصغرى الموزونة او احدى الطرائق العددية (Numerical Methods) [15]وكالآتي

$$\hat{\beta} = (X'WX)^{-1} X'WZ \qquad \ldots..(14)$$

اذ أن

$$Z = \begin{bmatrix} Ln\ \dfrac{p_1}{1-p_1} \\ Ln\ \dfrac{p_2}{1-p_2} \\ \vdots \\ \vdots \\ Ln\ \dfrac{p_m}{1-p_m} \end{bmatrix}$$

X تمثل مصفوفة المتغيرات التوضيحية وبدرجة (m* k)

W تمثل مصفوفة قطرية عناصر القطر الرئيسة فيها دالة الوزن W_i

حيث تعتمد حصانة المقدر الناتج من هذه الطريقة على دالة الوزن W_i ، وان لهذه الدالة صيغ عديدة حيث تم استخدام الصيغة المقترحة من قبل الباحثان (Meuller and Neykov) في عام (2003) وذلك بأستخدام دوال للاوزان[17] وكالآتي:

$$\left.\begin{aligned} W_1(t) &= (at + b), \\ W_2(t) &= (at^2 + b), \end{aligned}\right\} \qquad \ldots\ldots..(15)$$

a=0.8 , b=0.2

حيث أن

a,b تمثل ثوابت قيمها معروفة يتم الاعتماد عليها لكونها تعطي اقل وزن ويتم تعويضها في المعادلة (15) للحصول على الاوزان.

t تمثل دالة يمكن ايجادها بحسب الصيغة الآتية[15]:

$$t = h(X) = [\,(X - \hat{\mu})' \hat{\textstyle\sum}^{-1} (X - \hat{\mu})]^{½} \qquad \ldots..(16)$$

$$X = [X_{i1}, X_{i2}, \ldots \ldots, X_{ik}]'$$

$h(X)$ تمثل دالة لقياس المسافات الحصينة للموقع والتشتت
وبتكرار المعادلة (16) لـ m من المرات نحصل على t_i وكالآتي:

$$t_i = h(x_i) = [\ (x_i - \hat{\mu})' \hat{\Sigma}^{-1} (x_i - \hat{\mu})]^{½} \qquad \ldots..(17)$$
$$i = 1,2, \ldots \ldots, m$$

$\hat{\mu}$ تمثل التقديرات الاعتيادية لموجه الوسط الحسابي بدرجة k*1
$\hat{\Sigma}$ تمثل التقديرات الاعتيادية لمصفوفة التباين المشتركة بدرجة k*k
أن استخدام هذه الطريقة يكمن بأتباع الخطوات وكالآتي[7]:
1- حساب كل من $\hat{\mu}$ و $\hat{\Sigma}$ وبالصيغ الموضحة فيما يأتي

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \ldots\ldots(18)$$

$$\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}[(\ x_i - \hat{\mu})(x_i - \hat{\mu})' \qquad \ldots..(19)$$

2-أن التقديرات في المعادلتين (18) (19) تستخدم في حساب الصيغة في المعادلة (17) لكل مجموعة
3-اعادة احتساب كل من المتوسطات ومصفوفة التباينات حيث تبدل $(\hat{\Sigma}\ ,\ \hat{\mu})$ بـــ $(\hat{\Sigma}^*\ ,\ \hat{\mu}^*)$ وبأستخدام اوزان مقدرة تعتمد هذه الاوزان على قيم t_i وكالآتي[18] [7]:

$$\hat{\mu}^* = \frac{\sum_{i=1}^{m} w_{0i} x_i}{\sum_{i=1}^{m} w_{0i}} \qquad \ldots..(20)$$

$$\hat{\Sigma}^* = \frac{\sum_{i=1}^{m} w_{0i}^{\,2}[(x_i - \hat{\mu}^*)(x_i - \hat{\mu}^*)'}{\sum_{i=1}^{m} w_{0i}^{\,2}} \qquad \ldots..(21)$$

حيث أن
$\hat{\mu}^*$ تمثل متجه موقع حصين بدرجة (k*1)
$\hat{\Sigma}^*$ تمثل مصفوفة قياس حصين بدرجة (k*k)
w_{0i} تمثل دالة الوزن
لدالة الوزن w_{0i} عدة صيغ حيث تم استخدام احدى هذه الصيغ وهي صيغة دالة هوبر (Huber) حيث يتميز المقدر الناتج من هذه الصيغة بكونه مقدر كفوء وقليل الحساسية [11] وكما في الصيغة الآتية:

$$w_{0i} = \min\{1, \frac{g}{|t_i|}\} \qquad \ldots..(22)$$

حيث أن
g=1.37 وتمثل قيمة معروفة يتم الاعتماد عليها بتعويضها بالمعادلة (22) للحصول على اقل وزن للتقليل من تأثير القيم الشاذة.
وتمثل الخطوات المذكورة آنفاً التكرار الاول للطريقة، ثم يعاد احتساب الخطوة (3) وبصورة تكرارية معتمدين على نتائج التكرار السابق ويتم التوقف عن العملية التكرارية عندما يكون الفرق بين نتائج عمليتين تكرارين متعاقبين (عندما يصبح الفرق بين التقديرات المتعاقبة (اللاحقة والسابقة)) في تقدير المعلمات قليل او غير واضح.

2-3-2 مقدرات المسافة التربيعية الحصينة (QDE) **Quadratic Distance Estimates**

تعد طريقة (QDE) احدى طرق التقدير الحصينة حيث اقترحت من قبل الباحثان (**Luong and Garrido**) في عام (1992)، وأن الهدف الرئيس من هذه الطريقة هو تصغير مجموع الصيغ التربيعية، حيث تستخدم طريقة(QDE) في تقدير المعلمات في حالة وجود شواذ او نسبة من التلوث في البيانات، سوف نستخدم أسلوب الانحرافات لتقدير معلمات الانموذج[13].

وأن P_i تمثل نسبة الاستجابة وتقدر كالآتي:

$$P_i = \begin{cases} \dfrac{1}{2n_i} & if \quad y_i = 0 \\ \dfrac{y_i}{n_i} & if \ 1 \le y_i \le n_i - 1 \\ 1 - \dfrac{1}{2n_i} & if \quad y_i = n_i \end{cases} \qquad \dots \dots (23)$$

بالاعتماد على المعادلة (7) التي تمثل التحويل الخطي لدالة اللوجستي فأن خطوات أستخدام أسلوب الانحرافات في التقدير يكون [9]كالآتي:

$$r_i = y_i - x'_i \beta \qquad \dots \dots (24)$$

حيث أن

y_i تمثل المشاهدة (i) للمتغير المعتمد(الاستجابة)

$$X_i' = v_i x_i' \qquad \dots \dots (25)$$

$$X' = [\, X_1', X_2', \dots \dots, X_m' \,]$$

X' تمثل مصفوفة المتغير المستقل بدرجة (m*k) للتحويل اللوجستي

$$x_i' = [\, x_{i1}, x_{i2}, \dots \dots, x_{ik}]$$

$$i = 1, 2, \dots \dots, m$$

x_i' يمثل متجه صفي من المتغيرات التوضيحية حيث أن x_{i0} =1 وهو جزء من مصفوفة المتغيرات التوضيحية X

$$[n_i P_i \ (1 - P_i)]^{½} \quad v_i =$$

$$Y_i' = v_i Ln \frac{P_i}{1 - P_i} \qquad \dots \dots (26)$$

$$Y' = [\, Y_1', Y_2', \dots \dots, Y_m']$$

Y' يمثل متجه بدرجة m*1

β يمثل متجه معلمات الميل الحدي بدرجة (K*1) حيث يتم تقديره بطريقة المربعات الصغرى الاعتيادية.

β_0 يمثل الحد الثابت حيث نفترض معادلة للبواقي بالاعتماد على المعادلة (24) وكالآتي:

$$r_i = y_i - x'_i \beta_0 \qquad \dots \dots (27)$$

لكون الدالة التجميعية للبواقي(F) غير معلومة يتم الاعتماد على المعادلة (24) للبواقي لايجاد الدالة التجميعية وكالآتي:

$$\hat{F}_{\beta j}(r_i) = \sum_{i=1}^{m} W_{ij} I(y_i - x'_i\beta \le y) \qquad .....(28)$$

$$j = 1, 2, \dots \dots, k$$

حيث أن

W_{ij} تمثل أوزان معلومة

$$W = X(X'X)^{-1}$$

$$w_{ij} = [w_{1j}, w_{2j}, \dots\dots, w_{mj}]$$

$$W' = [w_1', w_2', \dots \dots, w_m']$$

W' تمثل مصفوفة الاوزان بدرجة (m*k)

$$w_i' = [w_1', w_2', \dots \dots, w_m']$$

$I(y_i - x'_i\beta)$ تمثل دالة المؤشر (indicator function) اذ ان لها تأثير في حصانة المقدر وتضرب هذه الدالة في مصفوفة الاوزان للحصول على مقدرات حصينة وكفوءة.

نلاحظ من المعادلة (28) انه يمكن الحصول على ($\hat{F}_{\beta j}$) بالاعتماد على البواقي في المعادلة (24).

أما بالنسبة للحد الثابت فأن الصيغة تكون كالآتي:

$$F_{\beta 0j}(r_i) = \sum_{i=1}^{m} W_{ij} F_{\beta 0}(y) \qquad .....(29)$$

حيث أن

$F_{\beta 0j}$ تمثل الدالة التجميعية لمتغير الاستجابة عند الحد الثابت حيث يمكن الحصول عليها أستناداً الى التوزيعات النظرية.

كذلك يمكن تعريف كل من ($Z_{\beta j}$)، ($Z_{\beta 0j}$)[9] وكالآتي:

$$\left.\begin{aligned} Z_{\beta j} &= [\int_{-\infty}^{\infty} h_1(x) d\hat{F}_{\beta j}(x), \dots, \int_{-\infty}^{\infty} h_k(x) d\hat{F}_{\beta j}(x)]' \\ &= [\sum_{i=1}^{m} W_{ij} h_1(y_i - x'_i\beta), \dots, \sum_{i=1}^{m} W_{ij} h_k(y_i - x'_i\beta)]' \end{aligned}\right\} \qquad \dots(30)$$

$$Z_{\beta 0j} = [\int_{-\infty}^{\infty} h_1(x) dF_{\beta 0j}(x), \dots, \int_{-\infty}^{\infty} h_k(x) dF_{\beta 0j}(x)]' \qquad \dots(31)$$

حيث ان

$$h_i(x) = -h_i(-x) \qquad \forall x \neq 0$$
$$h_i(0) = 0$$

h_i تمثل دوال مؤشر مختلفة وأن لها تأثير مهم اذ تعتمد حصانة المقدر الناتج من هذه الطريقة على اختيار تلك الدالة[8] وكالآتي:

$$h_1(x) = \begin{cases} 1 \ if \ x > 0 \\ 0 \ if \ x = 0 \\ -1 \ if \ x < 0 \end{cases} \quad , \quad h_2(x) \begin{cases} x & if \ |x| \leq M \\ sign(x)M & if \ |x| \ > M \end{cases}$$

أن طريقة المسافة التربيعية يتطلب تصغير مجموع الصيغ التربيعية وكالآتي:

$$\min = d(\beta) = (Z_{\beta 1} - Z_{\beta 01})' \ Q(Z_{\beta 1} - Z_{\beta 01}) + .. + (Z_{\beta k} - Z_{\beta 0k})' Q(Z_{\beta k} - Z_{\beta 0k}) \ .. (32)$$

حيث أن

$$Q = \Sigma^{-1}$$

Q تمثل مصفوفة ثوابت متماثلة وغير سالبة وبدرجة k*k

يمكن الحصول على افضل مصفوفة ثوابت Q بحسب الصيغة[8] الآتية:

$$(\tilde{\beta}) = (X'X)^{-1}(S_0' \Sigma^{-1} S_0)^{-1} \qquad .....(33)$$
$$Q = \Sigma^{-1} = var$$

حيث أن

$$S_0' = [\mathrm{E}(h_1(r)), ..., \mathrm{E}(h_k(r))] \qquad ......(34)$$

وبما أن $Z_{\beta 0j} = 0$ فأنه يمكن أن نختزل المعادلة (32) بالصيغة الآتية:

$$\min = d(\beta) = (Z_{\beta 1})' \ Q(Z_{\beta 1}) + \cdots + (Z_{\beta k})' Q(Z_{\beta k}) \qquad ....(35)$$

ولتبسيط المعادلة (35) وبأستخدام ضرب (Kronecker) نحصل على مقدرات طريقة المسافة التربيعية وكالآتي:

$$\tilde{\beta} = (Z_\beta)' \ (I_k \times Q) Z_\beta \qquad ......(36)$$

حيث أن I_k مصفوفة أحادية

$$Z_\beta = [(Z_{\beta 1})', ..., (Z_{\beta k})' \ ]'$$

3-الجانب التجريبي

1-3 المقدمة Introduction

يعد اسلوب المحاكاة من الاساليب العلمية الرصينة بوصفها اسلوباً للاختبار قبل تطبيق التجربة على بيانات واقعية.[1]

تعددت استعمالاته في المجالات المختلفة وتطور هذا الاسلوب نتيجة التقدم الذي حصل في مجال البرمجة والحاسوب حيث يوجد هناك العديد من الحالات التي يصعب تحليلها رياضياً لكونها معقدة من ناحية الفهم والتحليل وأن الكثير من العمليات الرياضية في الجوانب النظرية تحتاج الى جهد نظري لاشتقاقها فضلاً عن صعوبة ايجاد حلول لبعض التكاملات والمعادلات التفاضلية المعقدة.[3]

وغالباً ما يستخدم اسلوب المحاكاة في حالة فشل جميع الطرائق المستخدمة لايجاد حل لمشكلة ما او في حالة صعوبة الحصول على البيانات اللازمة لدراسة تللك الظاهرة او عدم توفرها بشكل كاف، وأن استخدامه يوفر على الباحثين الكثير من الوقت والجهد والمال كما أن استخدام هذا الاسلوب يؤدي الى تطوير أنموذج النظام من خلال ملاحظة التغييرات التي تطرأ على صياغة المشكلة عند تنفيذها عملياً.[4]

2-3 الانموذج المستعمل في المحاكاة simulation experiments

أن الانموذج الذي تم الاعتماد عليه في هذا البحث يكون وفق المعادلة (7) الواردة في الجانب النظري من البحث.

3-3 وصف تجارب المحاكاة Description simulation experiments

- لتحقيق الهدف من البحث فقد تم استعمال اسلوب المحاكاة من خلال صياغة تجارب لتوليد البيانات بأختيار أربع احجام مختلفة للعينات المفترضة وهي (n=15,25,50,100)، واعتماد نسب مختلفة لتلوث البيانات وهي $(\tau = 0\%, 10\%, 20\%)$ وكل تجربة تم تكرارها(1000) مرة، وعدد المتغيرات التوضيحية (K=5) وقد تم كتابة عدد من البرامج بلغة الـــ (Matlab).

- وكذلك تتضمن تجارب المحاكاة تعيين القيم الافتراضية للمعلمات، وتعد هذه الخطوة مهمة واساسية تعتمد عليها الخطوات اللاحقة، اذ تم فيها اختيار قيم المعلمات والنموذج المفترض، علماً بأن هذه القيم تم تحديدها من تقدير البيانات الحقيقية للبحث بطريقة المربعات الصغرى الاعتيادية كما مبين في أدناه:

β_0	β_1	β_2	β_3	β_4	β_5
-0.40	0.014	-0.011	0.022	0.025	-0.004

- تقدير معلمات انموذج الانحدار اللوجستي وفق الطريقتين التي تم عرضها في الجانب النظري من البحث وهي كالآتي:

1-طريقة مقدرات الإمكان الأعظم الحصينة الموزونة (WMLE).

2-طريقة مقدرات المسافة التربيعية الحصينة (QDE).

- وتمت المقارنة بين الطريقتين بالاعتماد على المقياس الاحصائي متوسط مربعات الخطأ(MSE) للانموذج وبحسب الصيغة الآتية:

$$MSE = \frac{\sum_{i=1}^{m} (\ Y_i - \hat{Y}_i)^2}{m-k} \qquad \ldots.(37)$$

4-3 **تحليل نتائج المحاكاة**

أولاً : **في حالة عدم وجود تلوث τ) (=0%)**

جدول (3-1) تقدير المعلمات بجميع الطرائق، وحجوم العينات عندما تكون نسبة التلوث (τ=0%)

N	Param. / Methods	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
15	WMLE(w_1)	0.0599	0.0051	-0.0049	0.0060	0.0088	-0.0015
	WMLE(w_2)	-0.1657	0.0047	-0.0048	0.0292	0.0090	-0.0018
	QDE	0.0599	0.0112	0.0421	0.0059	0.0311	0.0113
25	WMLE(w_1)	0.1640	0.0058	-0.0059	0.0180	0.0084	-0.0015
	WMLE(w_2)	-0.1326	0.0057	-0.0059	0.0262	0.0087	-0.0016
	QDE	0.1640	0.0154	0.0321	0.0008	0.0372	0.0115
50	WMLE(w_1)	0.0385	0.0050	-0.0036	-0.0025	0.0080	-0.0012
	WMLE(w_2)	-0.2750	0.0046	-0.0036	-0.0021	0.0083	-0.0012
	QDE	0.0385	0.0112	0.0234	0.0013	0.0711	0.0233
100	WMLE(w_1)	0.0478	0.0049	-0.0037	0.0127	0.0079	-0.0013
	WMLE(w_2)	-0.2993	0.0046	-0.0034	0.0094	0.0082	-0.0012
	QDE	0.0478	0.0211	0.0344	0.0155	0.0766	0.0455

من خلال النتائج المبينة في الجدول (3-1) نلاحظ ما يلي:

1- أظهرت طريقة ((WMLE(w2)) افضليتها من حيث اقتراب القيم التقديريـة من القيم الافتراضية فـي حالـة حجوم العينات الصغيرة.

ثانياً: في حالة وجود تلوث(τ=10%)

جدول (3-2) تقدير المعلمات بجميع الطرائق، وحجوم العينات عندما تكون نسبة التلوث (τ=10%)

N	Param. / Methods	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
15	WMLE(w_1)	1.9792	-0.0008	-0.0107	-0.0750	0.0079	-0.0027
	WMLE(w_2)	1.8881	-0.0005	-0.0101	-0.0549	0.0078	-0.0024
	QDE	1.9792	0.0188	0.0115	0.0067	0.0466	0.0229
25	WMLE(w_1)	0.9455	0.0011	-0.0091	-0.0756	0.0066	-0.0026
	WMLE(w_2)	0.0145	0.0015	-0.0091	-0.0663	0.0067	-0.0028
	QDE	0.7455	0.0511	0.0119	0.0008	0.0366	0.0228
50	WMLE(w_1)	0.6539	0.0012	-0.0084	-0.0737	0.0060	-0.0021
	WMLE(w_2)	0.5409	0.0015	-0.0084	-0.0809	0.0060	-0.0020
	QDE	0.5539	0.0212	0.0434	0.0013	0.0323	0.0441
100	WMLE(w_1)	0.4738	0.0013	-0.0083	-0.0562	0.0055	-0.0022
	WMLE(w_2)	0.4409	0.0014	-0.0088	-0.0553	0.0054	-0.0023
	QDE	0.3738	0.0113	0.0522	0.0355	0.0622	0.0321

من خلال النتائج المبينة في الجدول (3-2) نلاحظ ما يلي:

1- أظهرت طريقة ((WMLE(w_1)) افضليتها من حيث اقتراب القيم التقديرية من القيم الافتراضية في حالة حجوم العينات الكبيرة وعند هذه النسبة من التلوث.

ثالثاً: في حالة وجود تلوث(τ =20%)

جدول (3-3) تقدير المعلمات بجميع الطرائق، وحجوم العينات عندما تكون نسبة التلوث (τ=20%)

N	Param. / Methods	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
15	WMLE(w_1)	2.0287	0.0006	-0.0107	-0.0690	0.0073	-0.0023
	WMLE(w_2)	1.9589	0.0010	-0.0107	-0.0702	0.0074	-0.0021
	QDE	2.0287	0.0121	0.0785	0.0062	0.0107	0.0264
25	WMLE(w_1)	0.8251	0.0006	-0.0076	-0.0394	0.0057	-0.0020
	WMLE(w_2)	0.9170	0.0010	-0.0081	-0.0195	0.0058	-0.0022
	QDE	0.8251	0.0116	0.0130	0.0007	0.0154	0.0233
50	WMLE(w_1)	0.6516	0.0022	-0.0066	-0.0589	0.0047	-0.0020
	WMLE(w_2)	0.3357	0.0025	-0.0072	-0.0540	0.0046	-0.0023
	QDE	0.6516	0.0202	0.0185	0.0014	0.0211	0.0336
100	WMLE(w_1)	0.4251	0.0018	-0.0066	-0.0444	0.0044	-0.0020
	WMLE(w_2)	0.4296	0.0022	-0.0072	-0.0490	0.0043	-0.0023
	QDE	0.4251	0.0078	0.0054	0.0021	-0.0032	0.0018

من خلال النتائج المبينة في الجدول (3-3) نلاحظ ما يلي:

1- أظهرت الطرائق افضليتها من حيث اقتراب القيم التقديرية من القيم الافتراضية في حالة حجوم العينات الكبيرة الا ان طريقة (WMLE(w_1)) كانت الافضل من حيث اقتراب القيم التقديرية من القيم الافتراضية عند هذه النسبة من التلوث.

جدول(3-4) الخاص بقيم (MSE) للانموذج ولجميع الطرائق وحجوم العينات وعند نسب معينة من التلوث

نسب التلوث	n	WMLE(w_1)	WMLE(w_2)	QDE	Best
0%	15	0.2843	0.2884	0.7369	WMLE(w_1)
	25	0.2430	0.2393	0.6076	WMLE(w_2)
	50	0.2267	0.2266	0.5458	WMLE(w_2)
	100	0.2164	0.2169	0.5213	WMLE(w_1)
10%	15	0.2850	0.2793	0.4958	WMLE(w_2)
	25	0.2493	0.2507	0.3850	WMLE(w_1)
	50	0.2260	0.2287	0.3217	WMLE(w_1)
	100	0.2195	0.2197	0.2878	WMLE(w_1)
20%	15	0.2834	0.2866	0.4806	WMLE(w_1)
	25	0.2456	0.2465	0.3588	WMLE(w_1)
	50	0.2257	0.2279	0.2905	WMLE(w_1)
	100	0.2157	0.2174	0.2679	WMLE(w_1)

من الجدول (3-4) نلاحظ مايلي:

أولاً : في حالة عدم وجود تلوث(τ =0%)

أظهرت النتائج ومن خلال المقياس (MSE) للانموذج أن هناك تقارب في قيم MSE لطريقة((WMLE(w_1)) و((WMLE(w_2)) في حالة عدم وجود تلوث اذ تتنافس الطريقتين من حيث الأفضلية بأختلاف احجام العينات.

ثانياً: في حالة وجود تلوث (τ =10%)

أظهرت النتائج ومن خلال المقياس (MSE) للانموذج بأن طريقة ((WMLE(w2)) اثبتت كفاءتها عند حجم العينة (n=15) وذلك لانها حققت اصغر(MSE) للانموذج ، بينما كانت طريقة((WMLE(w_1)) هي الافضل لحجوم العينات (n=25,n=50,n=100).

ثالثاً: في حالة وجود تلوث τ (=20%)

أظهرت النتائج ومن خلال المقياس (MSE) للانموذج بأن طريقة ((WMLE(w_1)) هي الأفضل عند هذه النسبة من التلوث اذ اثبتت كفاءتها في تقدير المعلمات ولجميع حجوم العينات الصغيرة والمتوسطة والكبيرة وذلك لانها حققت اصغر(MSE) للانموذج.

4- الاستنتاجات والتوصيات

4-1 الاستنتاجات

بعد تنفيذ تجارب المحاكاة وما تم عرضه من نتائج وتحليل في الجانب التجريبي استنتج الباحث ما يأتي:

1- اثبتت طريقة (W_1)WMLE كفاءتها في تقدير معلمات أنموذج الانحدار اللوجستي في حالة تلوث البيانات ولجميع حجوم العينات، ونسب التلوث العالية وذلك لانها حققت اقل (MSE) للانموذج.

2- اثبتت طريقة (W_1)WMLE بأنها الانسب في حالة حجوم العينات الكبيرة نسب التلوث العالية وذلك لانها حققت اقل (MSE) للانموذج .

3-اشارت نتائج المحاكاة أن طريقة ((WMLE(w_2 تكون افضل في حالة حجوم العينات الصغيرة وعند عدم وجود تلوث، كما اثبتت كفاءتها في حالة نسب التلوث المنخفضة.

4- اشارت النتائج الى ان قيم متوسط مربعات الخطأ(MSE) للانموذج تتناقص بزيادة حجم العينة.

5- ظهر أن مقدرات (QDE) اقل كفاءة في تقدير معلمات أنموذج الانحدار اللوجستي من حيث قيم (MSE) للانموذج بالمقارنة مع طريقة (W_1)WMLE و(W_2)WMLE .

4-2 التوصيات

في ضوء الاستنتاجات التي تم التوصل اليها في الجانب التجريبي يمكن ادراج التوصيات وكالآتي:

1- استعمال طريقة (W_1)WMLE في تقدير معلمات أنموذج الانحدار اللوجستي وبأختلاف احجام العينات لما تبديه من كفاءة ومرونة في التطبيق.

2- استعمال طريقة (W_2)WMLE في تقدير معلمات أنموذج الانحدار اللوجستي في حالة حجوم العينات الصغيرة .

3-استخدام طرائق التقدير الحصينة بدلاً من الطرائق الاعتيادية في التقدير ولاي نوع من النماذج لكون مقدراتها تمتلك كفاءة عالية في حالة وجود القيم الشاذة.

5-المصادر

1-الجشعمي ، حسين علي عبد الله (2007) م ،"مقارنة لبعض المقدرات الحصينة لمعالم النماذج اللاخطية "، اطروحة دكتوراه في الاحصاء ، كلية الادارة والاقتصاد ، الجامعة المستنصرية.

2-العزاوي ، احمد ذياب (2005) م ،"المقارنة بين بعض طرائق تقدير انموذج الانحدار اللوجستي والطرائق الحصينة للتجارب الحياتية ذات الاستجابة الثنائية باستخدام اسلوب المحاكاة " ، رسالة ماجستير في الاحصاء ، كلية الادارة والاقتصاد ، جامعة بغداد.

3-بيثون، نغم نافع (1992) م، "خواص قوة الاختبار وحدود الثقة لمعاملات نموذج اللوجستك الخطي دراسة مقارنة"، رسالة ماجستير في الاحصاء ،كلية الادارة والاقتصاد، جامعة بغداد.

4-حسين، شرين علي (2009) م ،"مقدرات الامكان الاعظم الموزونة الحصينة ومقارنتها مع طرائق اخرى لانموذج اللوجستك مع تطبيق عملي " ، رسالة ماجستير في الاحصاء ، كلية الادارة والاقتصاد ، جامعة بغداد.

5-Carroll,R.J.and Pederson,S.(1993),"On Robustness In The Logistic Regression Model",Journal of the Royal Statistical Society .B,Vol. 55,No. 3,pp.693-709.

6-Chatterjee,S.,Hadi,A.S.(2012),"Regression Analysis By Example",Fifth Edition John Wiley & Sons.

7-Donoho,D.L.(1982),"Breakdown Properties of Multivariate Location Estimators". ph. D. Qualifying Paper, Dept. Statistics, Harvard University.
8-Doray ,L. G.and Luong ,A.(1995),"Quadratic distance estimators for the Zeta family .Insurance :Mathematics and Economics,16,pp. 225-260.
9-Flores,E.and Garrido, J.(2001),"Robust Logistic regression for Insurance risk classification", Universidad Carlos III de Madrid Calle Madrid,126,(spain).
10-Hosmer, D. W., Lemeshow, S., Sturdivant, R.X. (2013),"Applied Logistic Regression" University of Massachusetts, Third Edition.
11- Huber, PJ.(1984),"Finite Sample Breakdown of M-P- Estimators", Harvard University The Annals of Statistics,Vol. 12,No. 1,pp.119-126.
12- Luong ,A.(1991), "Minimum Distance Methods based on Quadratic Distances for transforms in Simple Linear Regression Model",Journal Royal Statistical Society B,Vol. 53,No. 2,pp. 465-471.
13-Luong,A.and Garrido.J.(1993),"Minimum quadratic distance estimation for a parametric family of discrete distributions defined recursively",Australian Journal of Statistics ,35,pp. 59-67.
14-Magnac,T.(2005),"Logit Models of individual Choices", Université de Toulouse ,Prepared for the New Palgrave ,First Version
15-Maronna,R.A.Martin,R.Dand Yohai, V.J.(2006)."Robust Statistics", Theory and Method, Jone Wiley & Sons ,Ltd.
16-Muller,ch.H and Neykov, N.(2003)."Breakdown Points of Trimmed Likelihood Estimators and Related Estimators in Generalized Linear Models", J.Statistic. Planning Inference 116,pp. 503-519.
17-Simeckova,M.(2005),"Maximum Weighted Likelihood Estimator in Logistic Regression", Charles University, Faculty of Mathematics and Physics,Part I,pp.144-148.
18- Stahel,W.A.(1981),"Breakdown of Covariance Estimators",Research Report No. 31,Fachgruppe Fure Statistic .ETH.Zuerich.
19-Vandev,D.Neykov,N.(1998)."About Regression Estimators with High Breakdown Point" Statistics 32,pp. 111-129.
20- Visek, J. A.(2000)."On The Diversity of Estimates",Comp .Stat.and Data Anl.34,pp. 67-89.

Comparison Some Robust Estimators for Estimate parameters logistic regression model to Binary Response – using simulation)).

Abstract

The logistic regression model of the most important regression models a non-linear which aim getting estimators have a high of efficiency, taking character more advanced in the process of statistical analysis for being a models appropriate form of Binary Data.

Among the problems that appear as a result of the use of some statistical methods Is not to achieve some or all the requirements including the presence of abnormal values between data, appears when the data of the studied phenomenon are contaminated ,it means some of the observations variety clearly from other observations called outliers.

From this point was the goal of this research to estimate parameters of logistic regression model through study some of Robust estimation methods The representing of the Robust weighted maximum likelihood estimators(WMLE), Quadratic Distance Estimators(QDE) We Use Simulation to comparison between two methods for different sample sizes and for difference proportions of contamination through mean square error (MSE) of the model, to reach the best method to estimate the parameter.

It was Concluded in through this Research to advantage of the method (WMLE(W_1)) in estimate parameters of binary response logistic regression model for different of samples sizes.

Keywords: Logistic Regression, Binary data, the Robust weighted maximum likelihood estimators(WMLE), Quadratic Distance Estimators(QDE), Outlier Observations.