

مقارنة بين طريقة الامكان الاعظم التكرارية وطريقة بيز لتقدير معاملات انموذج الانحدار اللوجستي مع تطبيق عملي

م.م. علي محمد علي جيجان
م.م. مصطفى علي فخري
م.م. زانا نجم عبد الله
الجامعة التقنية الوسطى/معهد الادارة الرصافة
جامعة ابن سينا للعلوم الطبية والصيدلانية
جامعة كركوك/كلية الادارة والاقتصاد

Comparison Between Iterative Maximum Likelihood Estimators Method And Bayesian Method For Estimating Logistic Regression Model Parameters With Practical Application

Assis. Lec. Ali Mohammed Ali Chichan

Middle Technical Uni. / Institute Of Management / Rusafa

Assis. Lec. Mustafa Ali Fakhri

Ibn Sina Uni.of Medical & Pharmaceutical Sciences

Assis. Lec. Zana Najm Abdallah

Uni. of Kirkuk/ College of Economic and Administration

المستخلص:

هنالك عدده شروط لاستخدام أنموذج الانحدار بصورة عامه في بعض الاحيان لا تتحقق بعض شروط استخدام الانحدار الخطي لذلك لابد من ايجاد طرق بديلة لتحليل البيانات حتى يمكننا من التنبؤ بالظاهرة المدروسة ومن اهم هذه الطرق هي طريقة الانحدار اللوجستي حيث يعتبر الانحدار اللوجستي من طرائق شائعة الاستخدام وخاصة في البيانات الثنائية، حيث استخدمنا في هذا البحث طريقتين لتقدير معلمات انموذج الانحدار اللوجستي وهي طريقة الامكان الاعظم التكرارية وطريقة بيز وكذلك استخدمنا معيارين للمقارنة وهو متوسط مربعات الخطأ ومتوسط مطلق الخطأ النسبي حيث تم استخدام بيانات حقيقية في هذا البحث المتمثلة بمرض سرطان الرئة بحجم عينة (30) مأخوذة من مستشفى مدينة الطب / مستشفى الاورام السرطانية حيث اظهرت النتائج ان طريقة بيز كانت الافضل من خلال معايير المقارنة في تقدير معلمات انموذج الانحدار اللوجستي.

الكلمات المفتاحية: انحدار اللوجستي، الامكان الاعظم التكرارية، بيز، متوسط مربعات الخطأ، متوسط مطلق الخطأ النسبي.

Abstract:

There are many conditions for using regression in general, sometimes the conditions of using regression are not fulfilled in this case we should find alternative methods of data analysis so that we can predict the phenomenon studied and the most important of these methods is the logistic regression method where logistic regression is one of the methods commonly used, especially in binary data, where we used in this research, two methods to estimate the parameters of the logistic regression model, a method iterative maximum Likelihood Estimators and Bayesian Method We used two criteria for comparison, which is the Mean Square Error, and the mean Absolute Percentage error. Real data was used in this research, which is represented by lung cancer with a sample size (30) taken from the City of Medicine Hospital / Cancer Hospital where the results showed that the method Bayesian the best by Comparison criteria in estimating logistic regression model parameters.

Keyword: logistic regression, iterative maximum likelihood Estimators, Bayesian, mean square error, mean absolute percentage error.

1-1 المقدمة

يعد الانحدار اللوجستي من الطرق الشائعة الاستخدام وخاصة في البيانات الفئوية (الثنائية) وكما انه يعطي فكرة سريعة وواضحة لتأثير المتغيرات المستقلة على المتغير التابع (الثنائي الاستجابة) حيث يمكن الاستنتاج من خلال أنموذج الانحدار اللوجستي بأن احد المتغيرات المستقلة يكون ذو أهمية اكبر من المتغيرات الأخرى، وتظهر أهمية الانحدار اللوجستي في حالة لا يمكن استخدام الانحدار الخطي لان من اهم مميزاته انه يجتاز الكثير من الفرضيات لطريقة انحدار المربعات الصغرى الاعتيادية ، مما يجعله الأسلوب الأفضل في حال وجود متغير معتمد يحتوي على بيانات ثنائية.

1-2 مشكلة البحث

في حالة وجود بيانات ثنائية خاصة بالمتغير التابع (المعتمد) لا يمكننا من استخدام طريقة الانحدار الخطي لذلك لابد من إيجاد طرائق أخرى يمكن استخدامها في مثل هذه المتغيرات ومن اهم هذه الطرائق وصف العلاقة بين المتغيرات هي طريقة الانحدار اللوجستي.

1-3 هدف البحث

تقدير انموذج الانحدار اللوجستي (Logistic Regression) المتمثلة بالمقدرات الكلاسيكية والبيزية وتطبيق هذه الطرائق على بيانات حقيقية لمرض سرطان الرئة والمقارنة بينهما.

1-4 الدراسات السابقة

في عام 2005م استخدم الباحث (Srivastava) [7] أنموذج الانحدار اللوجستي للتنبؤ بحدوث العواصف المغناطيسية حيث توصل الباحث الى إمكانية التنبؤ باستخدام لنموذج الانحدار اللوجستي. وفي عام 2017م درس الباحثين (Bassinello، Borries،Oliveira) [5] انموذج الانحدار اللوجستي لبيان جودة الأرز من حيث وقت الطبخ وسعر الأرز المطبوخ حيث استخدموا طريقة بيز لتقدير معلمات الانموذج والمقارنة بين عدده توزيعات سابقة باستخدام طريقة (MCMC).

وفي عام 2018م درس الباحثان (شمال،جيجان) [11] تقدير معلمات، انموذج الانحدار اللوجستي بطريقة مقدرات الإمكان الأعظم التكرارية على بيانات حقيقية تخص مرض عجز القلب حيث توصلوا الى ان وزن المريض وعمر المريض لهما تأثير كبير على مرض عجز القلب.

2- الجانب النظري

1-2 المقدمة

يعتبر أسلوب الانحدار اللوجستي من الأساليب الإحصائية الشائعة الاستعمال في العديد من المجالات التي تكون متغيراتها نوعية او وصفية (متغيرات ثنائية الاستجابة) مثال على ذلك (استجاب للعلاج او لم يستجب)، وغيرها من الأمثلة، حيث يستخدم الانحدار اللوجستي لوصف البيانات وشرح العلاقة بين المتغير الثنائي التابع والعديد من المتغيرات المستقلة.

في مثل هذه الحالات يكون المتغير التابع ثنائي الاستجابة اما يساوي (واحد) لحدوث الاستجابة أو (صفر) لعدم حدوث الاستجابة.

2-2 نموذج الانحدار اللوجستي (Logistic Regression)

كما ذكرنا سابقاً ان نموذج الانحدار اللوجستي يأخذ احدى القيمتين (اما صفر او واحد) حيث تكون القيمة (صفر) لعدم حدوث الاستجابة والقيمة (واحد) لحدوث الاستجابة وان دالة الكثافة الاحتمالية لأنموذج الانحدار اللوجستي تكتب بالصورة التالية [3,6,7]

$$f(g_i) = \tau_i^{g_i} (1 - \tau_i)^{1-g_i} \quad \dots (1)$$

أذ أن:

• g_i : متغير تابع ثنائي الاستجابة يأخذ احدى القيمتين (0، 1).

• τ_i : احتمال وقوع الاستجابة عند $g_i = 1$.

حيث يمكن كتابة انموذج الانحدار اللوجستي كالآتي:

$$g_i = \tau_i + \varepsilon_i, i = 1, 2, \dots, n \quad \dots (2)$$

اذ ان ρ_i تشير الى دالة الانحدار اللوجستي والتي يمكن صياغتها بصورة التالية:

$$\rho_i = P(g = 1) = P(Z\gamma) = \frac{\exp \exp(Z\gamma)}{1 + \exp \exp(Z\gamma)} \quad \dots (3)$$

حيث ان Z تمثل المتغيرات المستقلة.

وان $(1 - \rho_i)$ تشير الى احتمال عدم وقوع الاستجابة والتي يمكن صياغتها بالصورة التالية:

$$1 - \rho_i = P(g = 0) = 1 - P(Z\gamma) = \frac{1}{1 + \exp \exp(Z\gamma)} \quad \dots (4)$$

اما بالنسبة لحد الخطأ يتوزع توزيع برنولي (Bernoulli) بمتوسط صفر وتباين $\tau_i(1 - \tau_i)$.

3-2 مقدرات الامكان الاعظم التكرارية Iterative Maximum Likelihood Estimators

تعد طريقة الإمكان الأعظم احدى الطرائق الشائعة الاستخدام لما لها من خصائص تميزها عن بقية الطرائق الإحصائية، حيث يمكن تعريف هذه الطريقة بانها الطريقة التي تجعل قيم المعلمات في نهايتها العظمى، في هذا البحث سوف نستخدم طريقة الإمكان الأعظم التكرارية لتقدير معلمات انموذج الانحدار اللوجستي.

دالة الإمكان الأعظم لأنموذج الانحدار اللوجستي الذي يتبع توزيع برنولي تكون بالصيغة الاتية [1,4,8]:

$$L(\gamma, Z) = \prod_{i=1}^n \tau_i^{g_i} (1 - \tau_i)^{1-g_i} \quad \dots (5)$$

وحسب خاصية التحويل اللوجستي (دالة اللوجت)

$$\log \frac{\tau_i}{1-\tau_i} = z_i \gamma \quad \dots (6)$$

$$L(\gamma, Z) = \prod_{i=1}^n (1 - \tau_i) \exp[\sum_{i=1}^n g_i (z_i \gamma)] \quad \dots (7)$$

وبأخذ اللوغاريتم إلى دالة الإمكان في المعادلة (7).

$$\log L(\gamma, Z) = \sum_{i=1}^n \log (1 - \tau_i) + [\sum_{i=1}^n g_i (z_i \gamma)] \quad \dots (8)$$

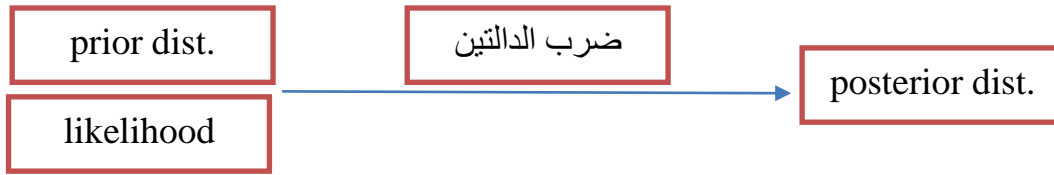
وبعد اخذ المشتقة الاولى والثانية الى لوغاريتم دالة الامكان سوف تكون مقدرات الإمكان الأعظم التكرارية حسب خوارزمية (Newton Raphson) بالصيغة التالية:

$$\hat{\gamma}^{(t+1)} = \hat{\gamma}^{(t)} + \{Z^T W Z\}^{-1(t)} \{Z^T (g_i - \tau_i)\}^{(t)} \quad \dots (9)$$

نلاحظ من المعادلة (9) انه لا يمكن تقدير معلمات الانموذج الا بإعطاء قيم أولية (افتراضية) وعادة ما تكون هذه القيم مساوية للصفر ولأجل تطبيق ذلك سوف نستخدم طريقة (Newton Raphson) التكرارية حيث يكون ناتج هذه الطريقة الفرق بين العمليات التكرارية لتقدير معلمات الانموذج مساوا او قريب من الصفر.

2-4 الطريقة البيزية Bayesian Method

تعد الطريقة البيزية من الطرائق المهمة لما لها من أهمية كبيرة في الكثير من التطبيقات العملية، حيث يتم اعتبار المعلمات في أسلوب بيز متغيرات عشوائية وليس قيم ثابتة كما في الطرائق الإحصائية الأخرى ، حيث تعتمد هذه الطريقة على وجود معلومات أولية (سابقة) حول الظاهرة المراد دراستها مأخوذة من مجتمع قيد الدراسة حيث يمكن تمثيل هذه المعلومات على شكل دالة كثافة احتمالية تسمى التوزيع الاحتمالي السابق (prior dist.) وبعد الحصول على دالة التوزيع السابق نقوم بضرب هذه الدالة بدالة الإمكان الأعظم عندها سوف نحصل على دالة احتمالية جديدة تسمى بدالة الكثافة الاحتمالية اللاحقة (posterior dist.) حيث يمكن تعريف هذه الدالة بأنها تمثل كافة المعلومات الجديدة حول المعلمات المراد تقديرها [2]، ويمكن تلخيص ما ذكر أعلاه بصورة التالية:



ويمكن تلخيص طريقة تقدير معلمات انموذج الانحدار اللوجستي كما يلي [5,9]:

ان دالة الامكان الاعظم لدالة الانحدار اللوجستي كما يلي:

$$L = \left(\frac{e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}}{1 + e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}} \right)^{g_i} \left(1 - \frac{e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}}{1 + e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}} \right)^{1-g_i} \dots (10)$$

وبما ان المتغيرات مستقلة فيما بينها نحصل على:

$$L = \prod_{i=1}^n \left(\frac{e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}}{1 + e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}} \right)^{g_i} \left(1 - \frac{e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}}{1 + e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}} \right)^{1-g_i} \dots (11)$$

يمكن فرض التوزيعات السابقة للمعلمات انها تتبع التوزيع الطبيعي وكما يلي:

$$\gamma_j \sim N(\mu_j, \sigma_j^2) \quad , j=1,2,\dots,p \quad \dots(12)$$

في معظم الدراسات يتم فرض قيمة المتوسط مساوي للصفر ($\mu_j = 0$) واما قيمة الانحراف المعياري تتراوح قيمته بين (10 to 100) $\sigma =$ ، هنا في هذا البحث نفترض ان قيمة الانحراف المعياري مساويه لـ (20) أي ان $\sigma = 20$

وباستعمال طريقة بيز يمكن ايجاد التوزيع اللاحق (posterior dist.) وكما يلي:

$$\text{posterior} = \prod_{i=1}^n \left(\frac{e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}}{1 + e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}} \right)^{g_i} \left(1 - \frac{e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}}{1 + e^{\gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip}}} \right)^{1-g_i} \\ * \prod_{j=1}^p \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left[-\frac{1}{2} \left(\frac{\gamma_j - \mu_j}{\sigma_j} \right)^2 \right] \dots (13)$$

3- الجانب العملي

1-3 المقدمة (introduction)

يعد سرطان الرئة المسبب الرئيس لوفيات مرضى السرطان بين الجنسين، إذ تزيد الوفيات بسبب هذا النوع على إصابات السرطان المختلفة الأخرى مجتمعة، سرطان القولون والبروستاتا والعقد اللمفاوية الثديي معا.

وبما ان السبب الرئيس والمسؤول الأول عن حدوث تلك السرطانات هو التدخين، والذي يسبب (90%) من جميع حالات الإصابة، لذا يمكن منع معظمها من خلال الحد من ظاهرة التدخين، كما ويمكن تقليل خطر الإصابة بسرطان الرئة من خلال تجنب التعرض لعوامل أخرى عادة ما تكون سببا بالإصابة به مثل التعرض للاسبست وغاز الرادون والتدخين السلبي.

3-2 أنواع سرطان الرئة [10]

يصنف الأطباء المختصون السرطان الى نوعين رئيسيين طبقا لشكل الخلايا السرطانية وشكل ظهورها تحت المجهر، ويتخذ الأطباء قراراتهم بشأن طرق معالجة تلك الامراض اعتمادا على ذلك التصنيف، وهما على النحو الآتي:
وهما على النحو التالي:

1. سرطان الرئة ذو الخلايا الصغيرة (Small cell lung cancer): وهو النوع الذي يسمى الورم الخبيث والذي يأخذ شكل سنبله الشوفان في معظم حالاته، ويظهر بكثرة فقط عند المدخنين، وهو اقل انتشارا من سرطان الرئة الذي يتميز بالخلايا غير الصغيرة.
2. سرطان الرئة ذو الخلايا غير الصغيرة (Non-small cell lung cancer): ويعد هذا النوع الأكثر انتشارا، والذي يشمل على أنواع عديدة من السرطانات الوسفية (Squamous cell carcinoma) والسرطانات الغدية (Adenocarcinoma) والسرطانات كبيرة الخلايا.

3-3 علاج سرطان الرئة [10]

عادة ما يتم التشاور بين المريض والطبيب المختص حول طريقة معالجة او نظام علاج سرطان الرئة اعتمادا على مجموعة من العوامل والتي أهمها؛ نوع السرطان ودرجة الإصابة والوضع الصحي العام للمريض، مع الاخذ بالحسبان خيارات المريض الشخصية، ومن اهم طرق المعالجة؛ العلاج الجراحي والعلاج الكيميائي والاشعاعي او الدوائي المركز.

3-4 الجدول الزمني للعلاج الكيماوي [10]

بعض الادوية تعمل بفاعلية أكبر عند إعطائها بشكل متتالي ولأيام عدة، ولكن عادة ما تتلو الجرعات الكيماوية فترة راحة للجسم وبشكل يسمح للخلايا السليمة بالتعافي من مشاكل العلاج الكيماوي، إذ ان كل جرعة تعطى للمريض يجب ان تحقق اكبر فائدة منها وياقل ضرر ممكن، وفي حال تناول ادوية أخرى يجب ان يتم تحديد وقت كل دواء وحجم جرعته، وعادة ما يتم وضع جدول لطريقة وأسلوب العلاج قبل البدء به وذلك بناءً على نوع السرطان ودرجة الإصابة ومرحلتها، وغالبا ما يكون العلاج الكيماوي ذات اثار جانبية قاسية، لذا يكون هذا الجدول قابل للتعديل من حيث الوقت او حجم الجرعة، ويفضل ان يتم الالتزام بجرعات متكاملة من العلاج الكيماوي للحصول على اعلى فائدة ممكنة من العلاج.

3-5 وصف البيانات

استخدمنا في هذا البحث بيانات حقيقة حول مرض سرطان الرئة حيث تم اخذ متغيرين يؤثران على علاج مرض سرطان الرئة وهما المتغير الاول (Z1) والذي يمثل (عمر المريض) والمتغير الثاني (Z2) والذي يمثل (متوسط جرعة العلاج الكيماوي بعد الانتهاء العلاج الكيماوي) اما بالنسبة للمتغير المعتمد (g) يمثل (استجابة المريض للعلاج) وان (1 شفي من المرض، 0 لم يشفى) تم الحصول على البيانات من مستشفى مدينة الطب / مستشفى الاورام السرطانية حيث تم سحب عينة بمقدار (30) مريض مصاب بمرض سرطان الرئة والجدول ادناه يمثل البيانات الحقيقية التي حصل عليها الباحث.

جدول رقم (1)

يمثل البيانات الحقيقية لمرض سرطان الرئة

ت	استجابة المريض (g)	العمر (z1)	متوسط الجرعة (z2)
1	0	67	650
2	1	60	570
3	0	57	530
4	1	56	430
5	1	59	450
6	1	58	540
7	0	36	350
8	1	58	570
9	0	68	600
10	1	33	300
11	0	27	320
12	0	70	530
13	0	62	600
14	1	56	530
15	1	51	480
16	0	67	650
17	1	69	680
18	1	64	600
19	1	25	200
20	1	70	550
21	0	56	300
22	1	45	450
23	1	64	630
24	1	47	300
25	1	43	520
26	0	53	500
27	1	44	430
28	0	33	270
29	1	48	450
30	1	52	510

3-6 تحليل نتائج

في هذا الجزء سيتم تقدير معاملات الانموذج من خلال البيانات الحقيقية لمرض سرطان الرئة للنموذج التالي:

$$g_i = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \varepsilon_i \dots (14)$$

باستخدام الطرائق المذكورة في هذا البحث وهي طريقة الامكان الاعظم التكرارية وطريقة بيز وسوف يتم المقارنة بين الطرائق المستخدمة باستعمال معياري المقارنة متوسط مربعات الخطأ (MSE) ومقياس متوسط مطلق الخطأ النسبي ($MAPE$) ولإجراء ذلك قام الباحث بكتابة برنامج لتقدير المعلمات والمقارنة بين طرائق التقدير من خلال معايير المقارنة باستعمال برنامج WinBUGS.

جدول رقم (2)

يمثل مقدرات الامكان الاعظم التكرارية وطريقة بيز

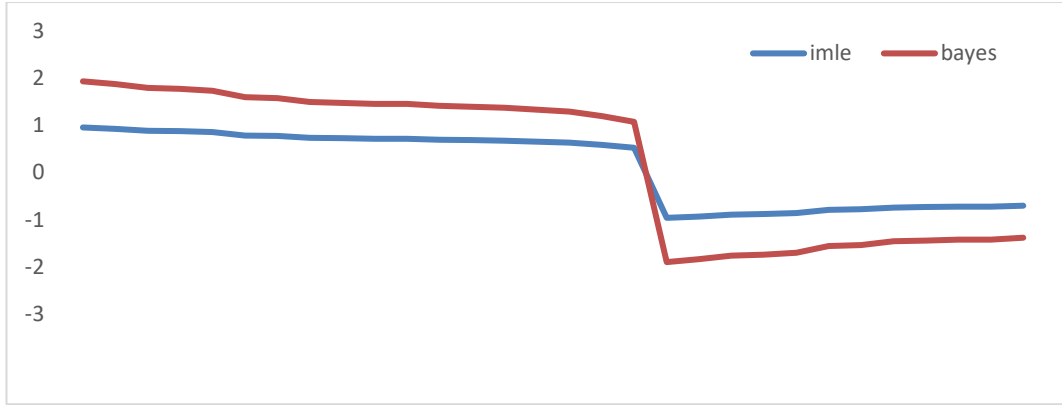
معلمت	طريقة $IMLE$	طريقة Bayes
$\hat{\gamma}_0$	-3.724	-4.675
$\hat{\gamma}_1$	2.732	5.673
$\hat{\gamma}_2$	7.425	9.324

جدول رقم (3)

يمثل معايير المقارنة

المعايير	طريقة $IMLE$	طريقة Bayes
MSE	0.625	0.251
$MAPE$	0.0133	0.0052

حيث يبدو واضحاً من الجدول رقم (3) واستناداً الى لمتوسط مربعات الخطأ (MSE) ومتوسط مطلق الخطأ النسبي ($MAPE$) ان طريقة Bayes كانت الافضل حسب معايير المقارنة.



شكل رقم (1)

دالة الانحدار اللوجستي للطرائق المستخدمة

يلاحظ من الشكل أعلاه العلاقة بين احتمال الاستجابة ρ_i والمتغيرات المستقلة (التوضيحية)

(Z_i) حيث تم رسم دالة الانحدار اللوجت باستخدام الصيغة $logit\left(\frac{\tau_i}{1-\tau_i}\right)$

4- الاستنتاجات والتوصيات

4-1 الاستنتاجات

- 1- من خلال معايير المقارنة المستخدمة نلاحظ ان طريقة Bayes كانت الافضل.
- 2- من خلال قيم المعلمات نلاحظ ان المتغير الثاني (متوسط العلاج الكيميائي) له تأثير أكبر من المتغير الاول (عمر المريض).

4-2 التوصيات

- 1- استعمال طريقة (Bayes) في تحليل انموذج الانحدار اللوجستي لكفاءتها وسرعة العالية في التطبيق.
- 2- يوصي الباحث باستخدام متغيرات اخرى غير مذكورة في هذا البحث مثل (التدخين و الوراثة).
- 3- يوصي الباحث الجهات المسؤولة بأهمية ترتيب تبويب البيانات لسهولة عملية جمعها من قبل الباحثين وتحليل هذه البيانات للوصول الى الهدف المنشود.
- 4- استعمال طرق اخرى في تقدير معلمات انموذج الانحدار اللوجستي مثل (طريقة انحدار الحرف، طريقة المركبات الرئيسية التكرارية وطريقة M الحصينة).

المصادر

أولاً: العربية

- 1- شمال، اياد حبيب، جيجان، علي محمد علي ، 2018م، " تحليل إثر بعض المتغيرات لاستجابة علاج مرض عجز القلب باستعمال انموذج الانحدار اللوجستي " مجلة الكوت الجامعة، المجلد 1، العدد 4.
- 2- كاظم، اموري هادي، مسلم، باسم شلبية، 2002م، "القياس الاقتصادي المتقدم النظرية والتطبيق"، مطبعة الطيف.

ثانياً: الأجنبية

- 3- Berkson. J. (1944) "**Application Of The Logistic Function To Bioassay**" JASA Vol .39, PP . 357 -365.
- 4- Neykov, N; Maeller, Ch.H.; (2003); "**breakdown points of trimmed likelihood estimators and related estimators in generalized linear models**"; J. Statist. Plann .Inference 116, 503 – 519.
- 5- Oliveira,Geiziane, Borries,George Von, Bassinello,Priseila Zaczuk ;2017 ;" **comparing priors in bayesian logistic regression for sensorial classification of rice** ", SAS 1018.
- 6- Shaeffer, R.L 1979 ; "**multi collinearty and logistic regression**", ph.D. dissertiation, university of Michigan, USA.
- 7- Srivastava. N. (2005). " **A Logistic Regression Model For Predicting The Occurrence Of Intense Geomagnetic Storms**" . Annales Geophysicae, 23, 2969 -2974.
- 8- Wang, X.; Van Eeoden; C.; Zideak, J.V. (2004). " **asymptotic properties of maximum weighted likelihood estimators**". 1929, 37-54. University Of British Columbia.

ثالثاً: المواقع الالكترونية

- 9- <http://www.medicine.mcgill.ca/epidemiology/joseph/courses/EPIB-621/bayeslogit.pdf>
- 10- www.arageek.com / الجرعات الكيماوية