Warith Scientific Journal

المقارنة بين طريقتي انحدار النواة و انحدار الشريحة في تقدير انموذج انحدار بواسون مع تطبيق عملي

Comparison of the kernel regression and spline regression methods in estimating the Poisson regression model with a practical application

اسامة عبد العزيز كاظم القريشي
osama Abdulazeez kadhim Al-Quraishi
ousamastat@gmail.com
وزارة التربية
Ministry of Education

الاء فلاح حسن عطوان
Alaa Falah Hassan Atwan
alaaousama29@gmail.com
وزارة التربية
Ministry of Education

المستخلص
ينتمي توزيع بواسون(**Poisson Dist.**) الى التوزيعات الرياضية المتقطعة والمهمة جدا في الكثير من الظواهر الحياتية المختلفة , ومنها النادرة كانقلاب القطارات وحالات الانتحار ...الخ, ويوجد الكثير من النماذج الرياضية وابرزها معادلات الانحدار(regression models) وبالاخص المعادلات المحولة من التوزيعات الاحصائية التي تستخدم للتنبؤ والتقدير, وبسبب التفاوت في جودة معادلة الانحدار المقدرة وعدم صلاحية استخدام بعض المعادلات منها لكونها لاتمتلك خصائص المقدرات الجيدة مما يؤدي الى عدم الثقة بدقتها التنبؤية او التقديرية , الامر الذي استوجب الى تقدير معادلة انحدار بواسون المعمم باستخدام طرائق لامعلمية وهي طريقتي انحدار النواة وانحدار الشريحة في نمذجة العلاقة بين عدد مرات تعرض المريض الى الجلطات القلبية متغير الاستجابه (Y) وعمر المريض المتغير التوضيحي (X) لعينة حجمها (70) في محافظة ذي قار وتبين من خلال النتائج التي تم التوصل اليها الى ان طريقة انحدار الشريحة هي الافضل لامتلاكها اقل MSE , وأن قيمة معامل التفسير الزائف Pseudo R- square بلغت (0.8206) .
الكلمات المفتاحية : انحدار بواسون , الطرائق اللامعلمية , معامل التفسير الزائف

Abstract
The Poisson Dist. belongs to the intermittent mathematical distributions, which are very important in many different phenomena, and rare phenomena such as train rollovers, suicides, etc., and there are many mathematical models, the most prominent of which are regression models, specifically the equations converted from statistical distributions that are used for prediction and estimation, As a result of the variation in the quality of the estimated regression models and the invalidity of using some of the models because they do not have the characteristics of good estimators, it leads to a lack of confidence in their predictive or estimating accuracy, which necessitated the estimation of the generalized Poisson regression equation using nonparametric methods, which are the two methods of nucleus regression and slide regression in modeling the relationship between the number of times the patient was exposed to heart attacks, the response variable (Y) and the patient's age, the exudative variable (X) for a sample size (70) in Dhi Qar governorate, and it was shown through the results that It was concluded that the slide regression method is the best because it has the lowest MSE, and that the value of the pseudo interpretation coefficient R-square was (0.8206)
Key words: Poisson regression, non-parametric methods , pseudoR-square measures

1 - المقدمة

ينتمي توزيع بواسون(Poisson Dist.) الى التوزيعات الرياضية المتقطعة والمهمة جدا في الكثير من الظواهرالحياتية المختلفة , ويوجد الكثير من النماذج الرياضية وابرزها معادلات الانحدار(regression models) وبالاخص المعادلات المحولة من التوزيعات الاحصائية التي تستخدم للتنبؤ والتقدير ,لذا أرتأى الباحث استعمال انموذج انحدار بواسون (Poisson regression model) لمعرفة العلاقة بين متغير الاستجابة (المعتمد) الذي يمثل عدد مرات تعرض المريض الى الجلطة القلبية والمتغير التوضيحي (المستقل) الذي يمثل عمر المريض بعد خضوع البيانات لتوزيع بواسون ,مستخدما الطرق اللامعلمية لتقدير معادلة الانحدار لان هذه الطرائق تمتلك المرونة العالية في تعاملها مع فرضيات الخطا العشوائي بحيث لايؤثر ذلك على كفاءتها , اذ انها تستند بصورة رئيسية على البيانات وان نوع البيانات يفسر الشكل الفعلي لمنحنى الانحدار.

2 - منهجية البحث

2-1 مشكلة البحث :

التفاوت في جودة نماذج الانحدار المقدرة وعدم صلاحية استخدام بعض النماذج منها لكونها لاتمتلك خصائص المقدرات الجيدة الامر الذي يؤدي الى عدم الثقة بدقتها التنبؤية او التقديرية .

2-2 هدف البحث:

يهدف البحث الى استخدام طريقتين لامعلمية وهما طريقة انحدار النواة وطريقة انحدار الشريحة في نمذجة العلاقة بين عدد مرات تعرض المريض الى الجلطات القلبية وعمر المريض باستخدام انموذج انحدار بواسون وتحديد الافضل من بين الطريقتين بالاعتماد على معيار المقارنة MSE .

2-3 عينة البحث :

تمثلت عينة البحث بالمرضى المصابين بالجلطة القلبية.

2-4 حدود البحث :

حدود البحث زمانيا سنة 2022 م للاشهر الثلاثة الاخيرة , مكانيا محافظة ذي قار.

2-5 مصدر البيانات :

تم الحصول على بيانات العينة من قاعدة البيانات الاحصائية لوزارة الصحة / دائرة صحة ذي قار .

3-الجانب النظري:Theoretical aspect

3-1 توزيع بواسون: [8],[1]

ينتمي توزيع بواسون الى شجرة التوزيعات الرياضية المتقطعه المهمة في العديد من التطبيقات الحياتية الاحصائية , وبعض الاوقات يدعى بتوزيع الوقائع النادرة , كوقائع انقلاب القطارات واصابات الامراض النادرة وغيرها من الوقائع النادرة, وسمي بتوزيع بواسون نسبة الى عالم الفيزياء والرياضيات (Simeon Poisson) الفرنسي , اذ قام باشتقاق التوزيع ونشره عام (1873 م) كحالة تقاربية لتوزيع ثنائي الحدين, ويمكن تعريف هذا التوزيع كالاتي:

يقال ان المتغير العشوائي Y هو ذي توزيع بواسون اذا كانت دالة الكتلة الاحتمالية (probability mass function) لـ (Y) تاخذ الشكل التالي:

$$\Pr(Y/\mu) = \begin{cases} \frac{e^{-\mu}\mu^{Y}}{Y!} & Y = 0,1,2,\dots \\ 0 & o.w \end{cases} \qquad \dots\dots\dots\dots(1)$$

اذ ان :

μ: معلمة التوزيع ($\mu > 0$) .

يمتاز توزيع بواسون بالخصائص الاتية:

أ- ($E(y_i) = \mu$).

ب- ($V(y_i) = \mu$).

ان تساوي قيمة المتوسط والتباين تدعى *بخاصية* تساوي التشتت (Equidispersion), وتكون دائما بالتطبيق العملي غير متحققة , وعندما تكون قيمة المتوسط اصغر من قيمة التباين تدعى بخاصية فوق التشتت (over dispersion).

ج- ملتويا باتجاه اليمين شكله البياني .

2-3 انموذج انحدار بواسون : [8] [9]

يمكن تعريف انحدار بواسون بانه الاسلوب الذي بواسطته يتم نمذجة المتغير التابع كونه متغير استجابة ($Response\ Variable$) عندما تكون قيم المتغير بصفة قيم معدودة ($Count\ Data$) او بصفة معدلات ($Rate\ Data$) , لنفرض عينة بحجم (n) من المشاهدات ($y_1 , y_2 , \dots , y_n$) قيم الاستجابه تتوزع بواسون , بمعلمة (μ) وان ($\mu = V(y_i) = E(y_i)$), تعتمد موجه (X) المتغيرات التوضيحيه, ولبناء العلاقة الانحدارية يجب مراعاة قيم (μ) موجبة فقط, الامر الذي يستوجب اللجوء الى الانحدار الخطي المعمم بتوظيف الدالة اللوغاريتمية لبناء علاقة خطية بين (Log μ) و(x), وكالاتي:

$$\text{Log}\ \big(E(y_i)\big) = \text{Log}(\mu_i) = \underline{X_i'}\underline{B} \qquad \dots\dots\dots\dots(2)$$

اذ ان :

$\underline{B}$: موجه المعلمات بسعة (K*1).

$\underline{Xi}$: موجه المتغيرات (المستقلة) التوضيحية ذات بعد (K *1) وبضمنها الحد الثابت.

وبتعميم الدالة الاسية على المعادلة رقم (2) نحصل على الانموذج اللا خطي وكما في المعادلة ادناه :

$$\mu_i = \exp(\underline{X_i'}\underline{B}) \qquad \dots\dots\dots\dots(3)$$

بموجب هذا الانموذج فان معاملات الانحدار الاسي(exponentiated regression coefficient) تاثيرها مضاعف , فعند الزيادة وحدة واحدة في المتغير (X_j) سوف يتضاعف المتغير ($Y_i=\mu_i$) بمقدار ($\exp(\beta_j)$).

مما تقدم فان انموذج انحدار بواسون يستند على ثلاثة افتراضات هي :,

أ- الافتراض الاول : متغير الاستجابة (Y) يتوزع توزيع بواسون بمعلمة (μ) كما معرف في الصيغة رقم (1) اذ ان:

(μ >0)

Pr(Y/μ) : الدالة الاحتمالية الشرطية للمتغير (Y) بمعلومية معلمة التوزيع (μ) .

ب- الافترض الثاني: معلمة التوزيع تمثل متوسط وتباين قيم متغير الاستجابة (Y) كما معرف بالصيغة رقم (3) .

ج- الافتراض الثالث : استقلالية ازواج المشاهدات (y_i , x_i) لكل منها توزيع مستقل.

بالاعتماد على الافتراضين الاول والثاني يمكن اعادة كتابة الدالة الاحتمالية الشرطية لمتغير الاستجابة وكالاتي:

$$f(y/x) = \frac{e^{-e^{x_i'\underline{B}}}\, e^{y_i x_i'\underline{B}}}{y!} \quad , y= 0,1,2,\ldots \qquad \ldots\ldots\ldots\ldots(4)$$

ان توزيع بواسون احادي المعلمة المحددة بوقت واحد (متوسط وتباين مشروط) , وعليه انموذج انحدار بواسون في ظل الافتراضات السابقة تكتب بالشكل الاتي :

$$E(Y/X) = Var(Y/X) = \mu_i = \exp(x_i'\underline{B}) \qquad \ldots\ldots\ldots\ldots(5)$$

تعد ميزة المتوسط المشروط والتباين المشروط المتساوي ميزه خاصة بانموذج انحدار بواسون , والتي تسمى بـ (Equidispersion,) وهذه الميزه غير متحققة عمليا بصوره دائمة , وكثير من الحالات نجد بان التباين المشروط اكبر من المتوسط المشروط وتعرف هذه الميزه بفوق التشتت (over dispersion), وقد يحدث العكس في حالات اخرى.

3-3 طرائق تقدير انموذج انحدار بواسون :

تمثلت طرائق التقدير بطريقتان تقدير لامعلمية لتقدير معادلة الانحدار لان هذه الطرائق تمتلك المرونة العالية في تعاملها مع فرضيات الخطا العشوائي بحيث لايؤثر ذلك على كفاءتها التقديرية او التنبؤيه, اذ انها تستند بصورة رئيسية على البيانات وان نوع البيانات يفسر الشكل الفعلي لمنحنى الانحدار.وكالاتي :

1- طريقة انحدار النواة [3] [5] [6]

تعد طريقة انحدار النواة من الطرائق اللامعلمية المهمه وذات الكفاءة العاليه المقترحه من قبل العالمين Nadaraya و Watson سنة 1964 من خلال لوغاريتم دالة الامكان يمكن اعادة كتابته بالصورة التالية :

$$LE(Y_i) = L(\mu_i) = \frac{1}{n}\{e^{\eta(x_i)} - y_i\,\eta(x_i)\}$$

اذ أن :

$$\eta(x_i) = \ln(\mu\ (x_i))$$

ان طريقة تقدير النواة اللاخطية لأنموذج انحدار بواسون تتلخص بتقدير متجه معلمات الانموذج $\eta(x_i)$ وذلك بإستخدام الانموذج الخطي التالي :

$$\eta(x_i) = b_0 + w'\,\phi\,(x_i) \qquad \ldots\ldots(6)$$

إذ أن :

$$w'\,\phi\,(x_i) = K(x_i, x_l)$$

وأن $K(x_i, x_l)$ هي عبارة عن احدى دوال النواة .

نحصل على تقدير متجه معلمات انموذج انحدار بواسون بواسطة طريقة النواة $\eta(x_i)$ وذلك بتصغير لوغاريتم دالة الأمكان الجزائيه السالبة (Penalizied negative Log _ Likelihood) وكمايلي :

$$L(w,b)=\sum_{i=1}^{n}[e^{b_0 + w'\,\phi\,(x_i)} - y_i(b_0 + w'\,\phi\,(x_i))] + \frac{\lambda}{2} + \|w\|^2 \qquad \ldots\ldots\ldots(7)$$

حيث ان λ تشير الى معلمة الجزاء (Penalty parameter) والتي وظيفتها هي السيطرة على الموازنة بين جودة حسن المطابقة (Goodness of fit) وذلك بالأعتماد على عاملين مهمين هما البيانات ومقدار التمهيد (Smoothness $\|w\|^2$) للدالة $\eta(x_i)$, والتي تجعل لوغاريتم دالة الأمكان الجزائية السالبه في نهايتها الصغرى , لذا سيكون متجه معلمات الأنموذج $\eta(x_i)$ وفق الصيغه التالية :

$$\eta(x_i) = b_0 + k_i'\alpha \qquad \ldots\ldots\ldots(8)$$

حيث أن :

b_0 : تشير الى الحد الثابت .

k_i : يشير الى العمود i في مصفوفة النواة ذات الدرجة (n*n) وعناصرها $k((x_i , x_l))$ وان (i, l)= 1 , 2, 3 , ...n

$\underline{\alpha}$: تشير الى متجه معلمات التمهيد ذات الدرجة (n*1) .

وبذلك فأن معادلة لوغاريتم دالة الامكان الجزائية السالبة ستكون كمايلي :

$$L(b_0, \alpha)=\sum_{i=1}^{n}[e^{b_0 + k_i'\alpha} - y_i(b_0 + k_i'\alpha)] + \frac{\lambda}{2}\ \alpha'\, k\alpha \qquad \ldots\ldots\ldots(9)$$

والغرض الحصول على على مقدرات الامكان الاعظم لنماذج الانحدار اللاخطيه والتي يكون متغير الأستجابه Y فيها يتبع توزيع بواسون يمكن ان الحصول على تلك المقدرات وذلك بإستخدام طرائق التقدير العددية مثل طريقة المربعات الصغرى

للأوزان المعادة ذات المراحل المتعددة (i.r.l.s) (Iterative reweighted least squares) والطريقه الثانيه هي طريقة نيوتن – رافسون , إذ تعد الطريقة الأولى (i.r.l.s) اسهل من الطريقة الثانية (نيوتن – رافسون) في اشتقاق معيار العبور الشرعي (c.v.c) (Cross – validation criterion) والتي لها معدل تقارب اقل من معدل التقارب في طريقة نيوتن – رافسون .

ولغرض استخراج مقدرات الأمكان الأعظم للمعالم المجهولة (b_0, α) نقوم بأيجاد المشتقات الجزئية الى لوغاريتم دالة الأمكان الجزائية السالبة بألنسبة للمعلمة α والمعلمة b_0 ومن ثم مساواة المشتقة الى الصفر أي أن :

$$\begin{pmatrix} \frac{\partial l(b_0,\alpha)}{\partial b_0} = 0 \\ \frac{\partial l(b_0,\alpha)}{\partial \alpha} = 0_n \end{pmatrix} \quad \ldots..(10)$$

وبحل الصيغه التالية بأستخدام طريقة المربعات الصغرى للأوزان المعادة ذات المراحل المتعددة (i.r.l.s) :

$$(V'WV + U)\,\beta = V'W\,y^* \quad \ldots\ldots.(11)$$

إذ أن:

y^* : تمثل قيمة المتغير المعتمد المحول الى الصيغه الخطية $(Ln(y))$

V هي مصفوفة ذات درجة $((n+1)*n)$ والمعرفه كمايلي :

$$V = (1_n , k) \quad \ldots\ldots\ldots(12)$$

إذ أن :

1_n : يمثل متجه الواحد الصحيح ذو الدرجة (n*1) .

k : تمثل مصفوفة النواة كما بينا ذلك سابقاً .

W : عبارة عن مصفوفة قطرية ذات درجة (n*n) والتي تعرف كمايلي :

$$W = \text{daig}\,[\exp(b_0 + k_1\alpha') , \ldots\ldots\ldots..,(b_0 + k_n\alpha')] \quad \ldots\ldots (13)$$

$\underline{\beta}$: تشير الى متجه المعلمات المراد تقديرها ذات الدرجة $((n+1)*1)$ والتي تعرف كمايلي :

$$\underline{\beta} = (b_0 , \alpha')' \quad \ldots\ldots\ldots(14)$$

U : تشير الى مصفوفه مربعة ذات درجة [(n*1) *(n*1)] والمعرفة كمايلي :

$$U= \begin{pmatrix} O & O_n' \\ O_n & \lambda k \end{pmatrix} \quad \ldots\ldots(15)$$

حيث أن:

O_n : تمثل متجه صفري ذو درجة (n*1).

y_i^* : تمثل المتجه الخاص بمتغير الأستجابة بطريقة انحدار النواة ذو الدرجة (n*1) وأن y_i يعرف كمايلي :

$$y_i = \frac{y_i - \exp(b_0 + k_i'\underline{\alpha})}{\exp(b_0 + k_i'\alpha)} + \exp(b_0 + k_i'\alpha) \quad \ldots\ldots(16)$$

أن تقدير متجه المعلمات (b_0, α) عند استخدام طريقة نيوتن – رافسون والتي تحدد قيم التقديرات في المراحل المتعددة للطريقة وتكون كمايلي :

$$\begin{pmatrix} \alpha^{(t+1)} \\ b_0^{(t+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(t)} \\ b_0^{(t)} \end{pmatrix} - \begin{pmatrix} k\,w^{(t)}\,k + \lambda k & k\,\mu^{(t)} \\ \mu^{(t)\prime} k & 1'\,\mu^{(t)} \end{pmatrix}^{-1} \begin{pmatrix} k\,\mu^{(t)} - y + \lambda\alpha^{(t)} \\ 1'\,(\mu^{(t)} - y) \end{pmatrix} \quad \ldots(17)$$

أن مقدرات دالة المتوسط عند متجه المدخلات x_0 وللقيم المثلى التقديرية لمتجه المعلمات (b_0, α) وتكون كمايلي :

$$\hat{\mu}(x_0) = \exp\,(\hat{b}_0 + k_0^{(t)}\hat{\alpha}) \quad \ldots\ldots\ldots\ldots(18)$$

حيث ان :

$$k_0 = (k\,(x_1, x_0), \ldots\ldots\ldots, k\,(x_n, x_0))^t \quad \ldots\ldots(19)$$

لغرض اختبار معلمة الجزاء (λ) ومعلمة الدالة اللبية (σ^2) والتي ينتج عنها افضل تمهيد ويستخدم لهذا الغرض معيار العبور الشرعي الاعتيادي (O.C.V) (Ordinary cross validation) وحسب الصيغه التالية :

$$\text{O.C.V}\,(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - e^{\widehat{\eta_\theta}(x_i)}}{1 - s_{ii}}\right)^2 \quad \ldots\ldots\ldots(20)$$

حيث أن :

s_{ii} يمثل العنصر القطري للصف (i) للمصفوفه S والمعرفه كمايلي :

$$S = V\,(\,V^t\,W\,V + U\,)^{-1}\;V^t\,W \quad \ldots\ldots\ldots(21)$$

لذلك عند التعويض عن s_{ii} بمتوسطها الذي يساوي $(\frac{1}{n}\text{tr(s)})$ في الصيغة (18) O.C.V(θ) سوف ينتج عنها معيار العبور الشرعي المعمم (g.c.v) (Generalized cross validation) والذي يتم حسابه بالنسبة ل θ من المعلمات وحسب الصيغه التالية :

$$g.c.v\ (\theta) = \frac{n\sum_{i=1}^{n}(y_i - e^{\widehat{\eta_\theta}(x_i)})^2}{(n-tr(s))^2} \quad \ldots\ldots(22)$$

إذ أن :

$\widehat{\eta_\theta}(x_i)$: تمثل تقدير الى لوغاريتم μ_θ اي ان $(\eta_\theta(x_i) = \log(\mu_\theta)(x_i))$ لكل معلمة من المعالم بعد استبعاد المشاهدة (i) حسب طريقة (Leave – one - out).

2-2-2 طريقة انحدار الشريحة : [4]،[7]،[10]

تعد طريقة انحدار الشريحة من طرائق التقدير اللامعلمية , وتستند فكرتها باستعمال دالة الامكان الجزائية , لبيانات المتغير المعتمد والتي تتوزع بواسون وكما في المعادلة رقم (23) :

$$L(\beta; y_1, y_2, \ldots, y_n, x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(y_i|x_i;\beta) \quad \ldots\ldots\ldots\ldots (23)$$

وباخذ الوغارتم لطرفي المعادلة أعلاه :

$$log(\beta, Y, X) = Log \prod_{i=1}^{n} f(y_i|x_i, \beta) = \sum_{i=1}^{n} Log\, f(y_i|x_i, \beta)$$
$$= \sum_{i=1}^{n}[-\exp(x_i'\beta) + yx_i'\beta - Log(y_i!)] \quad \ldots\ldots\ldots\ldots (24)$$

وباعتماد تمهيد الشرائح فان :

$$Log\mu(x) = X_i'\beta = s(x) \quad \ldots\ldots\ldots\ldots(25)$$

وباستعمال (cubic B-spline) الى دالة s(x) فيمكن كتابتها كالاتي :

$$s(x) = \sum_{k=1}^{K}\theta_k B_k(x) \quad \ldots\ldots\ldots\ldots(26)$$

اذ ان :

B_k :دوال الاساس للشرائح التكعيبية للدالة s(x) .

θ_k : معاملات الشرائح

k : عباره عن (q+4)

q : عدد نقاط الربط بشرط اختيارها على ابعاد متساوية للقيم للمتغير المستقل (x)

$$q = \min.\left[\frac{\text{numberof distinct valueof x}}{6}\ ,\ 30\right] \quad \ldots\ldots\ldots\ldots(27)$$

ويمكن اعادة كتابة المعادلة (27) على هيئة مصفوفات وكالاتي:

$$s(x) = B\theta \quad \ldots\ldots\ldots\ldots(28)$$

اذ ان :

B: مصفوفة (n*K) تمثل دوال الاساس للشرائح التكعيبية وهي:

$$B = [B_1(x_i), \ldots\ldots, B_K(x_i)] \quad \ldots\ldots\ldots\ldots(29)$$

Θ: موجه (K*1) تمثل معاملات الشرائح وهي :

$$\theta = [\theta_1, \ldots\ldots, \theta_K]' \quad \ldots\ldots\ldots\ldots(30)$$

و استنادا الى المعادلة رقم (25) , (27) فان المعادلة رقم (24) لوغارتم الامكان الجزائيه تكون كالاتي :

$$\ell(\theta) = \sum_{i=1}^{n}\left[y_i B_{x_i}'\theta - \exp\left(B_{x_i}'\theta\right) - \log(y_i!)\right] \quad \ldots\ldots\ldots\ldots(31)$$

يمكن استعمال طريقة دالة الامكان الجزائية penalized likelihood لتقدير موجه المعلمات(θ) وكالاتي :

$$\ell_{pen}(\theta) = \ell(\theta) - \frac{1}{2}\int (s''(x))^2 dx \quad \ldots\ldots\ldots\ldots(32)$$

اذ ان :

$\left(\int (s''(x))^2 dx\right)$: مقياس مقدار التمهيد (حد الجزاء) .

$S''(x)$: المشتقة الثانية للدالة (s) لموجه المعلمات (θ).

المعادلة رقم (32) يتم تقريبها لتبسيط حد الجزاء ولتكون بالشكل التالي :

$$\ell_{pen}(\theta) = \ell(\theta) - \frac{\lambda}{2}\theta' P\theta \quad \ldots\ldots\ldots\ldots(33)$$

اذ ان :

(λ) : معلمة التمهيد وتحدد وفق مقياس المقارنة كمقياس(GCV) العبور الشرعي المعمم , او مقياس (AIC) اكاكي.

ان (λ) معلمة التمهيد وظيفتها ضبط دقة التقدير و مقدار التمهيد الذي يساوي $(\theta' P\theta)$

اذ ان:

$$\theta^T P\theta = \sum_{k=1}^{K}(\Delta^2\theta_k)^2 \quad \ldots\ldots\ldots\ldots(34)$$

اذ ان :

$$\Delta^2\theta_k = \theta_k - 2\theta_{k-1} + \theta_{k-2} \quad \ldots\ldots\ldots\ldots(35)$$

و ان:

$$P = D_2^T D_2 \quad \ldots\ldots\ldots\ldots(36)$$

اذ ان :

D_2 : مصفوفة (Δ^2) عامل الفروق سعتها (K-2)*K , وبافتراض (K=5) فان المصفوفة تكتب كالاتي :

$$D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix} \quad \ldots\ldots\ldots\ldots(37)$$

تشتق ($\frac{\partial \ell_{pen}(\theta)}{d\theta}$) لوغاريتم دالة الامكان الجزائية المرقمه (24) بالنسبة لموجه المعلمات(θ) لاجل الحصول على موجه المعلمات(θ) ومن ثم نجعل لـ ($\frac{\partial \ell_{pen}(\theta)}{d\theta} = 0$) وبعد ذلك نجد الحل لمنظومة المعادلات التالية باستخدام طريقة المربعات الصغرى الموزونة ذات المراحل المتعددة (IRWLS) :,

$$(B'WB + \lambda P)\theta = B'Wy^* \quad \ldots\ldots\ldots\ldots(38)$$

اذ ان :

(W=diog[exp(Bθ)) مصفوفة قطرية (n*n) ,

B : مصفوفة دوال الاساس للشرائح التكعيبية

$y^* = (y_1^* , y_2^* , \ldots , y_n^*)^T$: موجه بيانات متغير الاستجابة الخاص بطريقة الشرائح التكعيبية (n*1) يكتب كالتالي:

$$y^* = W^{-1}(y - \mu) + B\theta \quad \ldots\ldots\ldots\ldots(39)$$

وان تقدير موجه معاملات الشرائح الجزائية $(\hat{\theta}_{pen})$ في المرحلة (t+1) باستخدام طريقة (IRWS) هي:

$$\hat{\theta}_{t+1} = \left(B'\widehat{W}_t B + \lambda P\right)^{-1} B'\widehat{W}_t \hat{y}_t^* \quad \ldots\ldots\ldots\ldots(40)$$

أذ ان :

$(\widehat{W}_t)$ و $(\hat{y}_t^*)$ يتم احتسابها وفق تقدير موجه معاملات الشرائح الجزائية في المرحلة السابقة (t) .

ولتقدير موجه $(\hat{\mu}_B)$ دالة المتوسط باستعمال الشرائح التكعيبية وفق التالي:

$$\hat{\mu}_B = \exp\{\hat{s}_B(x)\} = \exp(B\hat{\theta}_{pen}) \quad \ldots\ldots\ldots\ldots(41)$$

وايضا فان تقدير ($\widehat{Var - cov}(\hat{\theta}_{pen})$ مصفوفة التباين والتباين المشترك التقريبية لموجه معاملات الشرائح الجزائية هو:

$$\widehat{Var - cov}(\hat{\theta}_{pen}) \approx \{-Q(\hat{\theta})\}^{-1} \quad \ldots\ldots\ldots\ldots(42)$$

أذ ان:

$Q(\hat{\theta})$: المشتقة الثانية للوغاريتم دالة الامكان الجزائية بالنسبة لموجه معاملات (θ) عند القيم المقدرة لها,,

$(Q(\theta) = \partial^2 \ell_{pen}(\theta)/\partial\theta\ \partial\theta^T)$, ومن العلاقة (24)تم تقدير تباين ($\hat{s}_B(x)$) و كالتالي:

$$\widehat{Var}\left(\hat{s}_B(x)\right) = B_x' \{-Q(\hat{\theta})\}^{-1} B_x \quad .....(43)$$

تختار (λ) *على اساس* درجات الحرية الفعالة d (Edf)

أذ أن :

((Edf = tr(S_λ)-1

(S_λ) : مصفوفة التمهيد:

$$S_\lambda = B(B'WB + \lambda P)^{-1} B'W \quad \ldots\ldots\ldots\ldots(44)$$

ولأجل المقارنة بين قيم λ واختيار المثلى منها نستخدم مقياس أكياكي (AIC) اوبيز (BIC)حسب الصيغة التالية :

$$AIC(\lambda) = Dev(y; \theta, \lambda) + 2tr(S_\lambda) \quad \ldots\ldots\ldots\ldots(45)$$

$$BIC(\lambda) = Dev(y; \theta, \lambda) + Ln(n) * tr(S_\lambda) \quad \ldots\ldots\ldots\ldots(46)$$

أذ ان :

Dev(y;θ,λ) : إحصاءة الانحراف .

4- حسن المطابقة:

يوجد العديد من معايير حسن المطابقة لنماذج الخطية المعممة ومن هذه المعايير معيار الانحراف deviance

4-1 إحصاءة الإنحراف: [11] [7] [5]

تحسب احصاءة الانحراف وفق الصيغة الآتية:

$$D(y, \hat{\mu}) = 2[\ell(y) - \ell(\hat{\mu})] \quad \ldots\ldots\ldots\ldots (47)$$

إذ أن

$D(y, \hat{\mu})$: إحصاءة الإنحراف

$\ell(\hat{\mu})$: لوغاريتم دالة الإمكان للإنموذج المقدر

$\left(\ell(\hat{\mu}) = Log\, L(\hat{\mu})\right)$

$\ell(y)$: لوغاريتم دالة الإمكان الأعظم المتحققة

y و $\hat{\mu}$: موجهين ذو سعة $(n \times 1)$ بمدخلات y_i و $\hat{\mu}_i$ على التتابع.

ان صيغة معيار الانحراف وفق الإنموذج إنحدار بواسون هي :

$$Dev_p = \sum_{i=1}^{n}\left[y_i Log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right] \quad \ldots\ldots\ldots\ldots (48)$$

اذ ان :

$(y_i - \hat{\mu}_i)$: الفرق بين القيم المشاهدة لمتغير الاستجابة و القيم التقديرية لها, وبما ان مقدرات الامكان الاعظم لمعلمات انموذج انحدار بواسون تشترط مجموع بواقي الأنموذج تساوي صفر عند امتلاك الإنموذج الحد الثابت .وعليه فانه باستعمال الدالة الأسية للمتوسط , فإن إحصاءة الإنحراف Dev_p يمكن حسابها بشكل أبسط وكالاتي:

$$Dev_p = \sum_{i=1}^{n}\left[y_i Log\left(\frac{y_i}{\hat{\mu}_i}\right)\right] \quad \ldots\ldots\ldots\ldots (49)$$

ان معيار الإنحراف Dev_p يتوزع تقاربياً وفق توزيع مربع كاي بدرجة حرية (n-k)،[9],[3]،[1]

2-4 مقياس معامل التفسير الزائف:

يتم الحصول على معامل التفسير من خلال تحليل مربعات الإنحرافات الكلية وكالاتي,:

$$\sum_{i=1}^{n}(y_i - \overline{Y})^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{n}(\hat{\mu}_i - \overline{Y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{\mu}_i)(\hat{\mu}_i - \overline{Y})$$
$$\ldots\ldots\ldots\ldots (50)$$

إذ أن:

$\sum_{i=1}^{n}(y_i - \overline{Y})^2$: SSTotal.

$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$:.ssRegression.

$\sum_{i=1}^{n}(\hat{\mu}_i - \overline{Y})^2$: ss.error

وباستعمال طريقة المربعات الصغرى لحصول على المعلمات المقدره لا إنموذج الإنحدار الخطي مع احتواء الإنموذج للحد الثابت فإن:

$$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)(\hat{\mu}_i - \overline{Y}) = 0 \quad \ldots\ldots\ldots\ldots (51)$$

$$R^2 = 1 - RSS/TSS \quad \ldots\ldots\ldots\ldots (52)$$

$$R^2 = ESS/TSS \quad \ldots\ldots\ldots\ldots (53)$$

الا انه المعادلة رقم (51) لاتساوي صفر عند استعمال الطرائق التقديرية الأخرى , وايضا لنماذج الإنحدار غير الخطية لاتساوي صفر، بما في ذلك أنموذج إنحدار بواسون لذا فإن الميزة المرغوبة في مقياس معامل التفسير تكون غير متوفرة في نماذج الإنحدار غير الخطية، الأمر الذي يؤدي إلى استعمال معايير أخرى .

ومن خلال استعمال احد الصيغتين (52) و (62) في ايجاد R^2, فإن مقدرات الإمكان الأعظم لإنموذج الإنحدار غير الخطي لا تشترط زيادة قيمة (R^2)، وعند تضمين الإنموذج للحد الثابت فقد يكون (R^2) حسب معادلة (52) قيمة سالبة او يكون(R^2) وفق الصيغة (53) أكبر من الواحد.لذا اقترح الباحثين Cameron و Unlindeijer (1996 ، 1997)، ، المقياس Pseudo-R^2 باعتمادهم على تحليل الإنحراف Deviance، إذ أن:

$$D(y,\overline{y}) = D(y,\hat{\mu}) + D(\hat{\mu},\overline{y}) \quad \ldots\ldots\ldots\ldots (54)$$

إذ أن:

$D(y,\overline{y})$: الإنحراف للإنموذج المتضمن على الحد الثابت فقط.

$D(y,\hat{\mu})$: الإنحراف للإنموذج المقدر.

$D(\hat{\mu},\overline{y})$: الإنحراف المفسر

وعليه يمكن حساب معامل التفسير الزائف وفق احد المعادلات ادناه :

$$R^2_{DEV} = 1 - \frac{D(y,\hat{\mu})}{D(y,\overline{y})} \quad \ldots\ldots\ldots\ldots (55)$$

$$R^2_{DEV} = \frac{D(\hat{\mu},\overline{y})}{D(y,\overline{y})} \quad \ldots\ldots\ldots\ldots (56)$$

$$R^2_{DEV,p} = \frac{\sum_{i=1}^{n}\left[y_i Log\left(\frac{\hat{\mu}_i}{\overline{y}}\right) - (y_i - \hat{\mu}_i)\right]}{\sum_{i=1}^{n} y_i Log\left(\frac{y_i}{\overline{y}}\right)} \quad \ldots\ldots..(57)$$

كما يمتاز معامل التفسير الزائف بان قيمته بين الصفر والواحد وعند إضافة متغير مستقل إضافي قيمته لا تتناقص.

5- الجانب التطبيقي:

يتناول الجانب التطبيقي تقدير انموذج انحدار بواسون عن طريق استخدام طرائق لامعلمية الاولى هي طريقة انحدار النواة والاخرى طريقة انحدار الشريحة لبيانات عينة البحث لنمذجة العلاقة بين عدد مرات تعرض المريض الى الجلطة القلبية وأعمار المرضى وفق انموذج انحدار بواسون وتحديد اي من الطريقتين هي الافضل بالاعتماد على بعض معايير المقارنة.

1-5 وصف عينة البحث:

البيانات المستخدمة في التطبيق العملي عبارة عن عينة عشوائية مؤلفة من (70) من ازواج المشاهدات لقيم المتغيرين (Y,X) , اذ ان المتغير Y يمثل عدد مرات تعرض المريض الى الجلطة القلبية , اما المتغير X فيمثل أعمار المرضى المناظرة لها.

جدول (1): عينة البحث التي تمثل عمر المرضى مقابل عدد الاصابة بالجلطات القلبية.

18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	I
75	62	72	66	80	63	68	65	76	62	64	84	81	73	68	62	68	72	X
2	1	3	1	5	1	2	2	3	1	2	4	4	2	2	1	2	1	Y
36	35	34	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19	**I**
67	65	72	82	75	65	71	64	81	77	72	76	86	65	85	82	70	71	X
3	1	2	4	3	1	1	1	3	3	2	2	5	1	4	3	2	2	Y
54	53	52	51	50	49	48	47	46	45	44	43	42	41	40	39	38	37	**I**
70	66	67	75	81	73	68	81	84	63	86	74	69	73	77	68	73	87	X
2	2	2	3	2	2	2	2	4	1	4	2	2	3	2	2	2	2	Y
		70	69	68	67	66	65	64	63	62	61	60	59	58	57	56	55	**I**
		67	74	83	64	67	65	81	76	63	83	81	74	63	75	80	62	X
		1	1	3	1	2	1	1	1	1	3	2	3	1	2	3	1	Y

2-5 نمذجة العلاقة بين اعداد الجلطات القلبية المتكررة وأعمار المرضى:

باستعمال معادلة انحدار بواسون الخطي المعمم , الذي يشترط باستعمال دالة اللوغاريتم دالة رابطه نستطيع بناء علاقة خطية بين متغير الاستجابة لوغاريتم القيمة المتوقعة لأعداد الجلطات القلبية المتكررة(E(Y)) وعمر المرضى (X) المتغيرالتوضيحي وكالاتي:

$$\text{Log}\big(\text{E}(y_i)\big) = \text{Log}(\mu_i) = \beta_0 + \beta_1 x_i \quad \ldots\ldots\ldots\ldots(67)$$

3-5 تقدير معلمات الانموذج:

1-طريقة انحدار النواة

تم تقدير معلمات انموذج انحدار بواسون والمستخدم في نمذجة العلاقة بين المتغير (Y) اعداد الجلطات القلبية المتكررة والمتغير (X) اعمار المرضى وذلك باستخدام برنامج SAS باستخدام الايعازات التالية :

```
proc gam data=a;
model Y= kernel (X)
/dist=Poisson method=gcv;
run;
```

2-طريقة انحدار الشريحة

باستخدام الشرائح التكعيبية التي تم توضيحها في الجانب النظري تم تقدير معلمات انموذج انحدار بواسون والمستخدم في نمذجة العلاقة بين المتغير (Y) اعداد الجلطات القلبية المتكررة والمتغير (X)أعمار المرضى, تم تقدير الانموذج باستخدام حزمة البرامج الاحصائية SAS ,التحليل الاحصائي أجري باستخدام الايعازات التالية:

```
proc gam data=a;
model Y=spline(X)
/dist=Poisson method=gcv;
run;
```

لخصت نتائج الطريقتين في الجدولان التاليان :

جدول (2): نتائج تقدير معلمات انموذج انحدار بواسون بطريقتي kernel و Spline.

Method	Parameter	Parameter estimate	Std. Error	t- value	Sig.	F	Sig
kernel	Intercept	-2.6718	0.8530	-3.13	0.0026	2354.46	0.000
	Linear (X)	0.0465	0.0113	4.12	0.0001		
Spline	Intercept	-2.698	0.3947	-3.28	0.0001	1210.23	0.004
	Linear (X)	0.047	0.0052	4.35	0.0001		

المصدر/ اعداد الباحثان

جدول (3): ملخص تقدير معلمة التمهيد.

Component	Smoothing parameter	GCV	DEV	Pseudo R-square	Mse
Spline (X)	0.9801	0.0647	14.4318	0.8206	0.2038
Kernel (x)	0.3928	0.00222	14.3855	0.7006	0.2042

المصدر/ اعداد الباحثان

بالنظر الى الجدول رقم (2) تبين ان الانموذجان المقدران معنويان اذ ظهرت قيمة اختبار F المحسوبة لانموذج انحدار النواة تساوي (2354.46) *بمعنوية* (0.000) *وهي* اقل من مستوى معنوية (0.05) , بينما ظهرت قيمة اختبار F المحسوبة تساوي (1210.23) بمعنوية (0.004) *وهي* اقل من مستوى معنوية (0.05).

بالنظر الى جدول رقم (3) نلاحظ بان طريقة انحدار الشرائح (بالاعتماد على الشرائح التكعيبية) تمتلك معيار متوسط مربعات الخطا اصغر من طريقة انحدار النواة وعليه فان طريقة انحدار الشرائح هي الاكفأ ويمكن كتابة معادلة انحدار بواسون التقديرية كالاتي :

$$\text{Log}(\hat{y}_i) = -2.6718 + 0.0465\text{x}_\text{i}$$

بينما ظهرت قيمة معامل التفسير الزائف Pseudo R- square باستخدام انحدار الشريحه تساوي (0.8206) وهذا مؤشر على جودة معادلة انحدار بواسون التقديرية في تفسير العلاقة بين (y ,x) , وتبين ان متغير عمر المريض يمارس تأثيره في متغير عدد الجلطات القلبيه بنسبة (82.06%) *, وكلما زاد* عمر المريض وحدة واحدة زاد عدد الجلطات القلبيه بقدار الميل الحدي .

6- الاستنتاجات والتوصيات:

6-1 الاستنتاجات:

في ضوء ما تم التوصل اليه من نتائج تقدير انموذج انحدار بواسون بأستخدام طريقة انحدار النواة وطريقة انحدار الشريحة الذي يصف العلاقة بين اعداد الجلطات القلبية المتكررة وأعمار المرضى ندرج ادناه اهم ماتم التوصل اليه من استنتاجات:

أ- افرزت نتائج التقدير كفاءة طريقة انحدار الشرائح (بالاعتماد على الشرائح التكعيبية) مقارنة مع طريقة انحدار النواة في تقدير انموذج انحدار بواسون بالنسبة لعينة البحث بالاعتماد على معيار متوسط مربعات الخطأ.

ب- افرزت قيم معايير جودة حسن المطابقة الى كفاءة انموذج انحدار بواسون في تمثيل العلاقة بين اعداد الجلطات القلبية المتكررة وأعمار المرضى باستخدام الطريقتين (النواة والشريحة).

ج- وجود تأثير معنوي موجب لأعمار المرضى على عدد الأصابات بالجلطات القلبية المتكررة بالنسبة للطريقتين (النواة والشريحة).

د- قيمة معامل التفسير الزائف Pseudo R- square بلغت (0.8206) باستخدام طريقة انحدار الشريحة والتي تشير الى قدرة تفسيرية عالية للانموذج المقدر واقل منها بالنسبة الى طريقة النواة اذ بلغت (0.7006).

6-2 التوصيات:

بناءا على ما تم التوصل اليه من استنتاجات نوصي بالاتي:

أ- الاعتماد على طريقة انحدار الشريحة في تقدير انموذج انحدار بواسون لكونه اكفئ من طريقة انحدار النواة .

ب- مقارنة طرائق اخرى معلمية أو لامعلمية للتوصل الى طرائق اخرى ذات كفاءة عالية في تقدير انموذج انحدار بواسون.

ج- على الجهات الصحية والمعنية بالامراض المزمنة اقامة ورش وحملات تثقيفية توضح تأثير الزيادة في عمر على احتمالية حصول جلطات قلبية متكررة وذلك بأخذ الاحتياطات منها ضرورة استخدام بعض الادوية التي تساهم في تقليلها .

د- اجراء بحوث مستقبلية تأخذ بنظر الاعتبار تأثير العوامل الاخرى على الاصابة بالجلطات القلبية المتكررة .

المصادر:*Sources*

1. Algamal, Z. Y. and Rasheed, K. B., 2012, "Re- sampling Techniques in Count Data Regression Models",Iraqi Journal of statistical Sciences,vol. 22, pp. (15-25).
2. Barrios, E.B. and De Vera, E.A. , 2013, "Semi Parametric Poisson regression model for clustered data", Proceeding 59th ,[S] World Statistics Congers, 25-30 August 2013, Hong Kong(Session Cpso25).
3. Camarda, C. G. , 2008, "Smoothing methods for the analysis of mortality data", PH.D thesis, Carlos III University, Madrid, Spain.

4. Cameron, A. C. and Trived, P.R, 1998, "Regression Analysis of count Data", Cambridge University press, pp.(1-434).
5. Durban, M., Currie, I. and Eilers, P. , 2001, “Using P-Splines to Smooth Two – dimensional Poisson Data” , Proceedings of the 17th International Workshop on Statistical Modeling. Chania, Greece. Eds. M. Stasinopoulos and G. Touloumi, 207-214.
6. Hwang, G and Shim,J., 2008, “Semiparametric Kernel Poisson Regression for longitudinal Count Data”,
7. Li, C. S. and, Tu, W. , 2007, “A Spline - Based lack – of – Fit, Test for independent variable Effect in Poisson Regression”, Notional Institutes of Health, vol. 6, issue 1, pp. (239-247).
8. Paul H. C. Ellers and Brian D. Marx, 1996, "Flexible Smoothing with B-splines and Penalties" , Statistical Science, Vol. 11, No. 2, pp. 89-121.
9. Rodriguez, G. , 2007, “generalized linear Models”, Princeton University, Revised September 2007.
10. Spinelli,J. and others,2002, “Test for The response Distribution in a Poisson regression Model”, Journal of Statistical Planning and Inference, vol. 108,issue 1, pp. (137-154).
11. Winklmann Rainer , 2008, “Econometric Analysis of Count Data”, Fifth edition, Springer – Verlag Berlin Heidelberg.