KUFA STUDIES CENTER JOURNAL
UNIVERSITY OF KUFA
مجلة مركز دراسات الكوفة - جامعة الكوفة
KSCJ OPEN ACCESS
ISSN (P) : 1993-7016
ISSN (E) : 2708-728X
www.journal.uokufa.edu.iq

استخدام نموذج Quasi-Poisson لتحليل بيانات مرضى الثلاسيميا

Using Quasi-Poisson Model to Analyze Thalassemia Patient Data

أ.م.د. حسن سامي عريبي

الباحث محمد حسين نعمة

كلية الإدارة والاقتصاد/ جامعة القادسية

Asst Prof Dr. Hassan Sami Arabi

Researcher Mohammed Hussein Ne'meh

Faculty of Administration and Economics/ University of Al-Qadisiyah

DOI: https://doi.org/10.36322/jksc.v1i74(B).17726

الملخص:

شهد العقد الاخير من القرن العشرين اهتماما كبيرا بنماذج بواسون , الذي يعد احد نماذج العد التي تتعامل مع المتغير متقطع كمتغير معتمد. لقد واجه الباحثون عدد من التحديات لتطبيق هذا النموذج في عالم البيانات الحقيقية. اذ أظهرت بعض البيانات عدم تجانس عالي مما سبب ظهور مشكلة فرط التشتت. أُقتُرِح نموذج شبه بواسون ونموذج ذي الحدين السالب كنماذج بديلة لنموذج بواسون, بوجود مشكلة فرط التشتت.

سعت البحث الى دراسة المقارنة بين نموذج بواسون الكامل ونموذج بواسون بعد حذف عدد من المتغيرات غير المهمة , مع نموذج شبه بواسون وذو الحدين السالب بعد اختيار افضل المتغيرات , تم تنفيذ ذلك من خلال بيانات مرضى الثلاسيميا.

الكلمات المفتاحية: بواسون , شبه بواسون , ذي الحدين السالب , فرط التشتت , اختيار المتغيرات , معيار معلومات اكايكي , اختبار نسبة الاحتمالية , قيمة P.

Abstract:

The last decade of the twentieth century witnessed a great interest in Poisson's models, which is one of the counting models that deal with the discrete variable as a dependent variable. Researchers have faced a number of challenges to apply this model in the real data world. Some data showed high heterogeneity, which caused the emergence of the problem of over-dispersion. The quasi-Poisson model and the negative binomial model have been proposed as alternative models to the Poisson model, with the presence of the problem of over-dispersion.

This search sought to study comparing the full Poisson model with the Poisson model after deleting a number of unimportant variables, with the quasi Poisson and negative binomial model after choosing the best Variables, this was implemented through the data of thalassemia patients.

Keyword: Poisson , Quasi-Poisson , Negative Binomial , Overdispersion , Variable Selection , AIC , Likelihood Ratio Test , P-value.

المقدمة:

نماذج بواسون اكتسبت اهمية كبيرة في العديد من المجالات العلمية التطبيقية ولاسيما في المجالات الطبية. اذ ان نموذج بواسون يستخدم عند حدوث الحالات النادرة , لكن من الصعوبة بمكان ان نجد بيانات حقيقية تتوافق مع نموذج بواسون. لذلك تم اقتراح العديد من النماذج للتعامل مع بعض الحالات التي وجدت في عالم البيانات الحقيقية , مثلا عند وجود حالات فرط التشتت وبالاعتماد على حجم ذلك

التشتت يمكن استخدام نماذج العد , من هذه النماذج , نموذج Quasi–Poisson للحالات خفيفة التشتت و نموذج ذي الحدين عندما يكون التشتت شديدا. في هذه الدراسة تناولنا نموذج انحدار بواسون مع وجود مشكلة فرط التشتت , اذ اننا ركزنا على اختيار افضل المتغيرات عند تطبيق أي نموذج من نماذج العد السابقة , لذلك اعتمدنا على المقارنة بين نموذج بواسون قبل حذف بعض المتغيرات (النموذج الكامل) ونموذج بواسون بعد حذف المتغيرات , بالإضافة الى المقارنة مع نموذجي شبه بواسون وذي الحدين السالب. ان وجود متغيرات كثيرة في بيانات مرضى الثلاسيميا كانت مشكلة الباحث.

تناولت هذه الدراسة في المبحث الاول الاطار النظري للبحث , واحتوى على الاساسيات العامة.

اما المبحث الثاني فقد تضمن المفاهيم النظرية لتوزيع بواسون وافتراضات التوزيع , بالإضافة الى مشكلة فرط التشتت وكيف تنشأ والنماذج التي تعالج هذه المشكلة مثل نموذج شبه بواسون ونموذج ذي الحدين السالب. واحتوى المبحث الثالث على طرق اختيار المتغيرات منها معايير المعلومات. وبالنسبة للمبحث الرابع فقد شمل اختبار نسبة الاحتمالية واختبار P–value. بالنسبة لبيانات عينة الرسالة , ونتائج التحليل الاحصائي لهذه العينة تم تناوله في المبحث الخامس.

المبحث الاول: الاطار النظري:

اولا: مشكلة البحث:

حُدد عدد مرات اعطاء الدم بوصفه متغيراً معتمداً , لذا يجب اختزال المتغيرات الاخرى باستخدام طرق اختيار المتغيرات (variable selection) لمعرفة من هي المتغيرات الاكثر تأثيرا على عدد مرات اعطاء الدم للمرضى في مركز الثلاسيميا في محافظة النجف الاشرف.

ثانيا: فرضية البحث:

معرفة المتغيرات الاكثر تأثيرا على عدد مرات اعطاء الدم للمرضى في مركز الثلاسيميا في محافظة النجف الاشرف.

ثالثا : الحدود المكانية:

بيانات المرضى المسجلين في مركز الثلاسيميا وامراض الدم داخل مستشفى الزهراء التعليمي في محافظة النجف الاشرف. اخذت البيانات من اضابير المرضى الذين تم سحبهم في العينة العشوائية. سُحبت عينة عشوائية حجمها ١٥٥ مريض لـ (١٢) متغيراً توضيحياً ومتغير استجابة واحد الا وهو عدد مرات اعطاء الدم

رابعا: الحدود الزمانية:

الفترة من ٢٠١٨/١٠/١ ولغاية ٢٠٢١/٩/١ , أُخِذت معدلات زيارات المرضى للمركز سنويا.

المبحث الثاني: المفاهيم النظرية لنموذج بواسون وفرط التشتت

١- توزيع بواسون (Poisson Distribution)

يعد توزيع بواسون من التوزيعات المتقطعة المهمة جدا في الكثير من التطبيقات الإحصائية وكان اكتشافه مثمرا ومؤسفاً في الوقت نفسه. لقد كان مثمراً ؛ لأنه يسمح بنمذجة الأحداث النادرة ولكن مؤسفاً ؛ لأن الطريقة التي تم بها اكتشافه تسببت في اعتقاد خاطئ بأن التوزيع ينطبق فقط على الأحداث النادرة. هناك بيانات كثيره تنطبق على توزيع بواسون مثل عدد العملاء الذين يصلون إلى البنك في فترة زمنية معينة ((Padilla, ٢٠٠٣) وعدد الحوادث سنوياً على طول امتداد طريق سريع معين (١٩٩٧) ,Antelman , (١٩٩٨) Vogt and Bared. في مجال الأعمال التجارية كان توزيع بواسون محورياً في قائمة الانتظار ونظرية المخزون ومراقبة الجودة وفقاً لـ (Padilla, ٢٠٠٣). غالبا ما يستخدم الباحثون في العديد من التخصصات العلمية اعداداً صحيحة غير سالبة , وان بيانات هذه المتغيرات لا تتوافق مع التحليلات الإحصائية التقليدية مثل الانحدار الخطي المتعدد, لان شكل توزيع هذا النوع من البيانات يكون منحرفاً انحرافاً موجبا , الأمر الذي يجعل استخدام توزيع بواسون الاحتمالي أكثر ملاءمة لتحليل هذه المتغيرات , (سلمى ثابت ذاكر وانتصار مجيد جاسم (٢٠١٧)).

ISSN (P) : 1993-7016
ISSN (E) : 2708-728X
www.journal.uokufa.edu.iq

لنفرض ان (y) هو متغير عشوائي يشير الى عدد الاوقات لحصول حدث معين خلال فترة زمنية معينة لذا فأن (y) يتوزع بواسون بمعلمة قدرها (μ), شكل توزيع بواسون هو:

$$P(Y=y)=\frac{\mu^{y}e^{-\mu}}{y!} \qquad (2-1)$$

μ: تمثل معلمة التوزيع وهي ذات قيمة موجبة $(\mu \cdot <)$

المعلمة الأساسية لتوزيع بواسون هي المتوسط (μ) وهو متوسط عدد شيء ما لكل وحدة زمنية أو مساحة. بالنسبة للقيم المنخفضة لـ (μ) يكون للتوزيع شكله المنحرف المعتاد ولكن عندما تصبح (μ) كبيرة يتقارب شكل التوزيع مع التوزيع الطبيعي .

ان السبب الرئيس لاكتشاف توزيع الاحداث النادرة من خلال (Siméon–Denis Poisson) (١٨٤٠-١٧٩٠) هو ان توزيع ذي الحدين الذي يأخذ الصيغة الاتية :

$$f(x,n,y,p)=\frac{n!}{y!(n-y)!}p^{y}(١-p)^{(n-y)} \qquad (2-2)$$

إذ ان (y): عدد معين من التجارب الناجحة.

(n) : عدد المحاولات.

(p) : احتمال النجاح.

يكون غير مناسب لأحجام العينات الكبيرة ؛ اذ ان $(n!)$ عندما تكون (n) كبيرة , يصعب السيطرة عليه. لذلك اقترح بواسون علاجاً لهذه المشكلة بالسؤال عما يحدث لتوزيع ذي الحدين في النهاية عندما (n) تقترب من اللانهاية و تقترب (p) من الصفر, النتيجة هي توزيع بواسون. توزيع ملتوٍ وغير سلبي

اقترح بواسون استخدام هذا التوزيع تقريباً لتوزيع ذي الحدين (the binomial distribution) في حالة وجود أحداث نادرة ، أي عندما تكون (p) منخفضة جداً و(n) كبيرة , على سبيل المثال قد يكون احتمال الإصابة بمرض نادر منخفضاً (واحد من كل ألف) وعدد السكان كبيراً جداً (على سبيل المثال ، عشرات

الملايين من الاشخاص) , بحيث يكون عدد الأشخاص الذين يصابون بالمرض سنوياً (أو في فترة معينة من الوقت) نادرا , فيمكن وصفها بشكل مفيد من خلال توزيع بواسون. عندئذ يكون توزيع هذه الظاهرة منحرفاً بسبب وجود احتمال كبير لعدم الإصابة بالمرض وهو غير سلبي لأن الأعداد لا يمكن أن تكون أقل من الصفر.

أشار (Larsen & Marx, ٢٠٠٥) الى ان العديد من توزيعات العد اشتقت من توزيع بواسون بمفهوم جديد يتضمن الأحداث شبه النادرة. ولكن في السنوات الأخيرة تبنى الإحصائيون وجهة نظر أوسع لبواسون بوصفه توزيعاً قابلاً للتطبيق بشكل عام على البيانات العددية وهي حقيقة تم التعرف عليها خلال الـ ٥٠ عاماً الأخيرة من اكتشاف بواسون. يعد توزيع بواسون مناسبًا للاستخدام إذا تم استيفاء الافتراضات الأربعة الاتية:

١. عدد الأحداث يمكن عدها. نفترض أنه يمكن حساب عدد "الأحداث" التي يمكن أن تحدث خلال فترة زمنية معينة , يمكن أن تأخذ قيم ٠ ، ١ ، ٢ ، ٣ ، ... إلخ.

٢. المشاهدات مستقلة عن بعضها بعض.

٣. يكون التوزيع متقطعا وذا معلمة واحدة وهي المتوسط ويرمز له بالرمز λ (lambda) أو μ (mu). يعرف المتوسط على أنه معلمة المعدل , وهو العدد المتوقع لعدد مرات حدوث حدث معين خلال فترة زمنية معينة.

٤. متوسط وتباين النموذج متطابقان أو قريبان جدا من التماثل.

٥. معلمة التشتت بيرسون مربع كاي ($\chi^{٢}$ person) لها قيمة تقترب من (١). تنتج القيمة (١) عندما تكون الفروق الملاحظة والمتوقعة للاستجابة هي نفسها.

٢– فرط التشتت Overdisperision

KUFA STUDIES CENTER JOURNAL
UNIVERSITY OF KUFA
مجلة مركز دراسات الكوفة - جامعة الكوفة
KSCJ OPEN ACCESS
ISSN (P) : 1993-7016
ISSN (E) : 2708-728X
www.journal.uokufa.edu.iq

يبدو أن جاذبية نماذج انحدار بواسون لتحليل البيانات العددية التي تحتوي على أعداد صحيحة غير سالبة موثقة على نطاق واسع في الأدبيات الاقتصادية والطبية. وضح Dean في (١٩٩٢) السبب في ذلك إلى أن انحدار بواسون لنموذج العدد أساسي، ومتغير الاستجابة يكون متقطعا. يتعامل بواسون جيدًا مع البيانات المتقطعة ويفترض أن المتوسط الشرطي لمتغير الاستجابة يساوي التشتت الشرطي لذلك المتغير ((Frome et al, (١٩٧٣) و Frome and Checkoway (١٩٨٥)). واقعا ليس من السهل العثور على مثل هذا الشرط في البيانات الحقيقية ، ولاسيما البيانات المتقطعة التي ربما تخلق مشكلة عدم التجانس. ومع ذلك، عندما يتجاوز التشتت الشرطي في البيانات التشتت الأساس في ظل نموذج مفترض ، تحدث هذه الظاهرة ، التي أصبحت تُعرف باسم التشتت المفرط. قد لا تعكس المتغيرات المشتركة للدالة الشرطية بشكل كاف عدم التجانس الذي تم التغاضي عنه أو عدم ملاحظته من قبل الباحث، مما قد يؤدي إلى التشتت المفرط. ستكون إحصاءة الدرجة التي تم الحصول عليها من النماذج غير كافية وستكون الأخطاء القياسية التي يتم حسابها غالبًا بواسطة طريقة الامكان الاعظم (MLE) متحيزة. ويرجع السبب في ذلك إلى أن منهج MLE لا يمكنه تقديم تقديرات للمعلمات عندما لا يمكن تحديد التوزيع الاحتمالي للمتغير عشوائياً بشكل كافٍ بواسطة أي نموذج افتراضي موجود. تكون بواقي بيرسون أو الانحراف في النموذج المناسب كبيرة جدا في المقام الأول مما يؤدي إلى عدم كفاية إحصاءة حسن المطابقة أيضًا , وفقا لـ ((Manton,et. al ١٩٨١; Breslow ١٩٨٤; Brillinger ١٩٨٦)).

عدم التجانس غير الملحوظ في البيانات ليس هو السبب الوحيد للتشتت المفرط ، ربما بسبب ظهور القيم الشاذة ، او تأثير المتغيرات الأخرى التي تؤدي إلى الاعتمادية بين حدثين احتماليين أو أكثر ، او قد تم تضخيم متغير الاستجابة بواسطة الأصفار الزائدة (Hilbe, (٢٠١٤. قد يؤدي وجود التشتت المفرط إلى استنتاجات غير صحيحة ، ويرجع السبب في ذلك إلى ما يسمى بالتحيز الصاعد والتحيز النازل. التحيز

KUFA STUDIES CENTER JOURNAL
UNIVERSITY OF KUFA
مجلة مركز دراسات الكوفة - جامعة الكوفة
KSCJ OPEN ACCESS
ISSN (P) : 1993-7016
ISSN (E) : 2708-728X
www.journal.uokufa.edu.iq

الصاعد يمكن ملاحظته مع المتغيرات المستقلة لأنه يزداد مع أهميتها. يؤدي التحيز النازل إلى التقليل من شأن الانحراف المعياري لتقديرات المعلمات (under estimate)(٢٠٠٧) Ismail and Jemain. على الرغم من ان التشتت المفرط تناولته الأدبيات الإحصائية لاول مرة من خلال تعليقات Student (١٩١٩) , اقترح (١٩٥٠) Fisher اختبار جودة الملاءمة لتقييم فعالية توزيع بواسون في حالة العينة الواحدة ، أي عندما تؤخذ الاعداد كمتغيرات مستقلة بمتوسط مشترك من توزيع بواسون. مما لا شك فيه أنه في النموذج المصاغ عندما لا يتم وضع معامل التشتت في الاعتبار فبالتأكيد إن النتائج لا يمكن الاعتماد عليها. صنف (٢٠١٤) ,Hilbe التشتت المفرط إلى نوعين حقيقي وظاهري. مصدر التشتت الحقيقي هو إما تضخم صفري أو ارتباط ، في حين أن القيم الشاذة ، والمتغيرات التوضيحية المفقودة ودوال الارتباط غير المناسبة تسبب التشتت المفرط الظاهر (لمزيد من التفاصيل انظر: Hilbe and Hardin, (٢٠٠٧) و Dean and Lundy, (٢٠١٤)). لذلك فإن تركيز الباحثين في الأدبيات الإحصائية يميل إلى تعديل النموذج ليناسب أي نوع من أنواع التشتت المفرط. من الواضح أن قرار تعديل نموذج بواسون لملاءمة البيانات يعتمد على حدوث نقص في التشتت أو فرط في التشتت. هناك نماذج تتعامل مع فرط التشتت وهي :

أ. نموذج شبه بواسون Quasi–Poisson model

هو نموذج يمكنه التعامل مع مشكلة التشتت المفرط ، إذ يولي هذا النموذج الانتباه إلى معامل التشتت الذي يتسبب في عدم تساوي تباين البيانات مع المتوسط. تم تطوير هذا النموذج بواسطة Wedderburn (١٩٧٤) من طرق تقدير شبه الاحتمال. في تقدير الاحتمالية، يجب أن يتبع متغير الاستجابة الشكل التوزيعي بينما يتطلب تقدير شبه الاحتمال فقط العلاقة بين المتوسط والتباين. وفضلا عن ذلك، فإن نموذج شبه بواسون يقدر قيمة احصاءة التشتت بينما لم ينتبه نموذج بواسون لهذه القيمة.

نموذج شبه بواسون يتكون من نماذج خطية معممة مع افتراضات تشبه بواسون (Ver Hoef and Boveng، ٢٠٠٧). إن المتوسط والتباين للمتغير العشوائي y الذي يتبع توزيع شبه بواسون يكون كالاتي:

$$E(y) = \mu$$
$$var(y) = \alpha * E(y) = \alpha\mu \qquad (٢-٣)$$

إذ μ تشير إلى متوسط متغير الاستجابة Y و تشير α إلى معلمة التشتت التي سيتم تقديرها من البيانات , عندما تكون $\alpha > ١$ يتكون لدينا تشتت مفرط.

ب. نموذج ذي الحدين السالب (Negative Binomial Model)

بديل آخر لنمذجة البيانات العددية مفرطة التشتت الا وهو نموذج ذي الحدين السالب , وهو نموذج يصور عدد النجاحات في سلسلة من تجارب برنولي المستقلة والموزعة بشكل متماثل , مشتق من خليط بواسون و كاما. وغالباً ما يستخدم لنمذجة بيانات بواسون شديدة التشتت وفقا لـ (Amaliana and Wardhani,٢٠١٧). ينشأ كتوزيع لعدد حالات الفشل (Y) قبل (r) من حالات النجاح في التجارب المستقلة ، مع احتمالية النجاح (p) في كل تجربة , وبالتالي ($r \geq ٠$ و $٠ \leq p \leq ١$). في مثل هذه الحالة يمكن التعبير عن دالة الكتلة الاحتمالية بما يأتي (٢٠١١) Lindén and Mäntyniemi :

$$f(k; r, p) = \Pr(X = k) = \binom{k + r - ١}{r - ١}(١ - p)^k p^r \qquad (٢-٤)$$

و r هو عدد حالات النجاح ، k هو عدد حالات الفشل ، و p هو احتمال النجاح ,

في هذه الصيغة ، المتوسط هو $(\frac{(١-p)r}{p})$ والتباين هو $\frac{(١-p)r}{p^٢}$.

الكمية الموجودة بين قوسين هي المعامل ذو الحدين وتساوي

$$\binom{k+r-١}{r-١} = \frac{(k+r-١)!}{(r-١)!(k)!} = \frac{(k+r-١)(k+r-٢)\cdots(r)}{(k)!} = \frac{\Gamma(k+r)}{k!\,\Gamma(r)} \qquad (٢-٥)$$

حيث $\Gamma(.)$ ترمز لدالة كاما.

يمكن كتابة دالة الكتلة الاحتمالية بشكل اخر , وكما يأتي:

$$f(k;r,p) = \Pr(X=k) = \frac{\Gamma(k+r)}{k!\ \Gamma(r)}(١-p)^k p^r \qquad (٢-٦)$$

المبحث الثالث: طرق اختيار المتغيرات:

يمكن أن يؤدي اختيار المتغيرات التفسيرية المهمة الموجودة في النموذج إلى تحسين دقة التنبؤ بالنموذج كما عبر عنها Konishi and Kitagawa (٢٠٠٨). إن المجموعة الفرعية الصغيرة من المتغيرات التفسيرية تجعل تفسير النتائج أسهل. لذلك نحتاج إلى توضيح البيانات بطريقة ربما تكون بسيطة، إذ نحدد مجموعات فرعية من مجموعة المتغيرات الأصلية للحصول على أصغر مجموعة فرعية يمكن استخدامها للنمذجة وتقليل التكلفة. لذلك تجب إزالة المتغيرات التفسيرية الزائدة عن الحاجة. يمكن أن يساعدنا استبعاد هذه المتغيرات الزائدة عن الحاجة في توفير الوقت Snipes and Taylor (٢٠١٤) .

ومن اهم معايير اختيار المتغيرات هي :

معيار معلومات اكايكي Akaike Information criterion(AIC)

يُقدر معيار معلومات Akaike (AIC) الجودة النسبية للنماذج الإحصائية لمجموعة معينة من البيانات. تم اكتشافه من قبل الإحصائي الياباني Hirotsugu Akaike في عام ١٩٧٣. تقدر (AIC) قيمة جودة كل نموذج بالنسبة للنماذج الأخرى بالنظر إلى مجموعة من نماذج البيانات , وبالتالي فإن (AIC) يوفر وسيلة لاختيار النموذج. تُستخدم استراتيجية (AIC) للاختيار من بين نموذجين متنافسين أو أكثر. يمثل النموذج الذي يمتلك اقل AIC افضل تقريب للنموذج الحقيقي (انظر (Mutua,١٩٩٤) و Burnham and Anderson,(٢٠٠٤)) , يُعد هذا النموذج صاحب أصغر خسارة متوقعة للمعلومات عندما تُستبدل القيم المعيارية الحقيقية في النموذج بتقديرات MLE. صيغة (AIC) هي:

KUFA STUDIES CENTER JOURNAL
UNIVERSITY OF KUFA
مجلة مركز دراسات الكوفة - جامعة الكوفة
KSCJ OPEN ACCESS
ISSN (P) : 1993-7016
ISSN (E) : 2708-728X
www.journal.uokufa.edu.iq

$$(3-1)\text{AIC} = -٢\square\square\square\square + ٢\square$$

فإن: L هي دالة الامكان الاعظم للنموذج المقدر (MLE).

k هو عدد المتغيرات التفسيرية

يُشار إلى ٢k بوصفه مصطلحاً جزائياً، والذي يتم تعديله وفقًا لأبعاد النموذج. بالنظر إلى أن إضافة المزيد من المعلمات إلى النموذج يجعل البيانات أكثر احتمالية، فعندما نقوم بزيادة عدد المتغيرات التفسيرية، يصبح (٢□□□□−) أصغر ؛ لذا تُضاف العقوبة ٢k إلى اللوغاريتم الاحتمالي للتكيف مع هذا التحيز المحتمل.

المبحث الرابع: الاختبارات:

١. اختبار نسبة الاحتمالية (LRT) Likelihood Ratio Test

هو اختبار مهم لتقييم قيمة النماذج المتداخلة (أي حيث تتم مقارنة نموذج به عدد أقل من المتغيرات التفسيرية مع نفس النموذج لكن مع متغيرات تفسيرية اكثر). يقوم الاختبار بتقييم ما إذا كان يجب الاحتفاظ بالمتغيرات التفسيرية التي تم سحبها من النموذج. تُستخدم اختبارات نسبة الاحتمالية أيضًا لمقارنة النماذج المختلفة إذا كان أحدها مجموعة فرعية أو نسخة مختصرة من النموذج الأخر , انظر (٢٠١١) Hilbe.

لنفرض ان n ، ⋯ ، ١ = i ، X_i ، هي متغيرات عشوائية مستقلة ذات قيم صحيحة غير سالبة. نريد اختبار الفرضية القائلة بأن البيانات تتوزع توزيع بواسون ، أي أن فرضية العدم هي:

$$H_٠ : X_i \sim Poisson(\mu_i) \,,\, \mu_١ = \cdots = \mu_n = \mu$$

بشكل عام عند استخدام نموذج مثل بواسون ، يرغب الباحث عادة في التركيز على البدائل "مفرطة التشتت" , لذا فالفرضية البديلة هي:

$$H_١ : X_i \sim Poisson(\mu_i) \,,\, \mu_١ \neq \cdots \neq \mu_n$$

KUFA STUDIES CENTER JOURNAL
UNIVERSITY OF KUFA
مجلة مركز دراسات الكوفة - جامعة الكوفة
KSCJ OPEN ACCESS
ISSN (P) : 1993-7016
ISSN (E) : 2708-728X
www.journal.uokufa.edu.iq

بالنسبة لفرضية العدم $H_{٠}$ ، تكون طريقة الامكان الاعظم الخاص بـ μ المشترك هو $\bar{\mu} = \bar{X}$ ، وهو متوسط العينة. اما بالنسبة للفرضية البديلة $H_{١}$ ، تكون طريقة الامكان الاعظم لـ μ_i هي $\mu_i = X_i$.

صيغة احصاءة likelihood ratio لاختبار فرضية العدم $H_{٠}$ مقابل الفرضية البديلة $H_{١}$ هي:

$$T_{LR} = ٢\sum_{i=١}^{n} \mu_i ln\left(\frac{\mu_i}{\bar{\mu}}\right) \qquad (٤-١)$$

في ظل فرضية العدم تُوزع هذه الإحصاءة بشكل تقريبي كمتغير مربع كاي (Chi–squared) مع درجة حرية $n-١$. (بشكل تقريبي عندما $\infty \leftarrow n$ لـ μ). ومن ثم فإن هذا الاختبار يرفض $H_{٠}$ عندما تكون $T_{LR} > \chi^{٢}_{n-١}$. في ظل الفرضية البديلة، تتوزع إحصاءة likelihood ratio توزيع مربع كاي غير مركزي بشكل تقريبي بدرجة حرية (n–١) ومعلمة لامركزية , وكالتالي:

$$\psi^{٢} = \frac{\sum_{i=١}^{n}(\mu_i - \bar{\mu})^{٢}}{\bar{\mu}} \qquad (٤-٢)$$

عندما $\bar{\mu} = \frac{\sum_{i=١}^{n}\mu_i}{n}$

يكتب $T_{LR} \sim \chi^{٢}_{n-١}(\psi^{٢})$. هذا التقريب صالح (تقريباً) عندما $\infty \leftarrow \mu$ لـ n مع $\mu_١, \mu_٢, \cdots, \mu_n$ تم اختيارها للاعتماد على n بطريقة تظل $\psi^{٢}$ ثابتة.

٢- اختبار P–value   P–value test

هي احتمال رفض فرضية العدم ($H_{٠}$) , وهي الفرضية التي تنص على أنه (على سبيل المثال ، عدم وجود فرق أو عدم وجود ارتباط). عبر (٢٠٠٦) Blocker, et.al الى ان "P" في P–value تشير إلى الاحتمالية (probability) , وان P–value تقيس قوة الدليل ضد ($H_{٠}$) , وكلما كانت P–value أصغر، زاد احتمال رفضنا لفرضية العدم .

مستوى (α) هو مقدار الخطأ من النوع الأول الذي يرغب الباحث في قبوله , (الخطأ من النوع الأول هو حيث يتم رفض فرضية العدم ($H_{.}$) وهي صحيحة). اكد كل من Thiese و Ott في عام (٢٠١٦) انه يجب تحديد مستوى المعنوية (α) مسبقًا ، أي قبل اجراء الدراسة وجمع البيانات. اقترح فيشر أنه يمكن استخدام مستوى ٥٪ اي ان (α = ٠.٠٥) لاستنتاج وجود دليل قوي إلى حد ما ضد فرضية العدم ($H_{.}$) ، أي مرة واحدة من ٢٠ مرة سيكون الباحث مخطئًا وبالصدفة العشوائية , اذ يكتشف الباحث أن هناك اختلافًا , في حين أنه في الحقيقة لا يوجد اي اختلاف.

تتأثر P-value بحجم العينة اذ كلما زاد حجم العينة يقل تأثير الخطأ العشوائي , فضلا عن تقليل التباين الكلي , وعندئذٍ تصبح القياسات أكثر دقة بالنسبة للمجتمع كله. تسمح هذه الدقة المتزايدة باكتشاف الفروق الصغيرة بين المجموعات.

تقدر P-value احتمالية أن تكون النتيجة ناجمة عن الصدفة , إذا كانت قيمة ($P <$ ٠.٠٥) ولكن حجم التأثير قد يكون منخفضاً جدًا ، من حيث رفض الفرض العدمي ولكنه صحيح , عندئذٍ الاختبار يكون ذا دلالة إحصائية ولكن عند التطبيق ليس كذلك.

المبحث الخامس: عرض نتائج تحليل البيانات وتفسيرها:

اتضح من النتائج المعروضة في جدول رقم (١) ان هناك اربعة متغيرات تفسيرية (الوزن , الحالة الزوجية للمريض , صلة القرابة للوالدين , نسبة الهيموغلوبين في الدم) مؤثرة على المتغير المعتمد (عدد مرات اعطاء الدم), و اعتبار المتغيرات التفسيرية الاخرى (العمر, جنس المريض , صنف الدم , نسبة الحديد في الدم , هل تم رفع الطحال , عدد افراد الاسرة , السكن (ريف- مدينة) , عدد الاخوة المصابين) ليست لها اهمية في النموذج قيد الدراسة.

جدول رقم (١) نتائج اختبارات (AIC , LRT ,P-value) لنموذج بواسون الكامل لبيانات مرضى الثلاسيميا

S	Variable	AIC	Likelihood Ratio	P – value
١	Age	٦٤٠,٧١	٠,٠٠٢٣	٠,٩٦١٧٢٥٨
٢	Sex	٦٤٠,٧٣	٠,٠٢٣٢	٠,٨٧٨٨٨٧٨
٣	Blood groub	٦٤١,٣٨	٠,٦٧٧٥	٠,٤١٠٤٥٠٤
٤	X.Weight	٦٤٩,٠٩	٨,٣٨٩٢	٠,٠٠٣٧٧٤٧ **
٥	S.Ferritin	٦٤١,٧٨	١,٠٧٢٤	٠,٣٠٠٣٩٨٥
٦	Splenectomy	٦٤١,٥٤	٠,٨٣٦٦	٠,٣٦٠٣٦٥٤
٧	X.Marital status	٦٥٤,٨٥	١٤,١٤٥٧	٠,٠٠٠١٦٩٢ ***
٨	Consanguinity	٦٥٦,٠٦	١٥,٣٥٢٩	$8.918\,e^{-٠٥}$ ***
٩	No. family members	٦٤٠,٨٥	٠,١٥٠٢	٠,٦٩٨٣٢٩٥
١٠	Res countryside city	٦٤١,٣٦	٠,٦٦٠٨	٠,٤١٦٢٧٢٧
١١	Affected brothers and sisters	٦٤٢,٦٦	١,٩٦٠٦	٠,١٦١٤٤٧٧
١٢	Hemoglobin	٦٦٥,٤٩	٢٤,٧٨٦٠	$6.406\,e^{-٠٧}$ ****

اتضح للباحث ان المتغيرات (الوزن , الحالة الزوجية , صلة القرابة للوالدين , نسبة الهيموغلوبين في الدم) هي صاحبة اكبر تأثير على المتغير المعتمد (عدد مرات اعطاء الدم).

بعدها تم استخدام اختبار (LRT) واتضح ايضا بأن المتغيرات الاربعة اعلاه كانت قيمها اعلى من باقي المتغيرات , بقية المتغيرات في هذا الاختبار كانت قيمها قليلة جدا. اخر اختبار استخدمه الباحث هو اختبار (P–value) إذ تم اخذ المتغيرات التفسيرية التي قيمها اقل من (٠,٠٥) اي انها معنوية , فوجد الباحث ان المتغيرات التفسيرية الأربعة اعلاه (المؤشرة بعلامة *) قيمها اقل القيم الموجودة في عمود الاختبار , ومن هنا تكون اكثر المتغيرات تأثيرا على متغير الاستجابة.

جدول رقم (٢) يوضح اختبارات z و P–value للمتغيرات التفسيرية الاربعة المهمة في نموذج بواسون

KUFA STUDIES CENTER JOURNAL
UNIVERSITY OF KUFA
مجلة مركز دراسات الكوفة - جامعة الكوفة
KSCJ OPEN ACCESS
ISSN (P) : 1993-7016
ISSN (E) : 2708-728X
www.journal.uokufa.edu.iq

Poisson				
	Estimate	Std.Error	z	p–value
Intercept	4.45742	0.60365	٧,٣٨٤	$1.53\ e^{-١٣}$
Weight	0.01390	0.00286	٤,٨٦١	$1.17\ e^{-٠٦}$
Matrial Status	−0.50197	0.12311	−٤,٠٧٧	$4.55\ e^{-٠٥}$
Consanguinity	0.13266	0.02820	٤,٧٠٣	$2.56\ e^{-٠٦}$
Hemoglobin	−0.12845	0.02186	−٥,٨٧٧	$4.18\ e^{-٠٩}$
AIC	633.05			

جدول رقم (٣) يوضح اختبارات *z* و *P–value* للمتغيرات التفسيرية الاربعة المهمة في نموذج شبه بواسون.

Quasi Poisson				
	Estimate	Std.Error	z	p–value
Intercept	4.45742077	0.52471887	٨,٤٩٤٨٧٤	$1.1811672\ e^{-١٤}$
Weight	0.01390302	0.00248636	٥,٥٩١٧١٦	$1.034897\ e^{-٠٧}$
Matrial Status	−0.50197204	0.10701208	−٤,٦٩٠٧٩٨	$6.074889\ e^{-٠٦}$
Consanguinity	0.13265734	0.02451645	٥,٤١٠٩٥٣	$2.428408\ e^{-٠٧}$
Hemoglobin	−0.12845348	0.01899988	−٦,٧٦٠٧٥١	$2.859551\ e^{-١٠}$
AIC	630.61			

جدول رقم (٤) يوضح اختبارات z و P–value للمتغيرات التفسيرية الاربعة المهمة في نموذج ذي الحدين السالب.

Negative Binomial				
	Estimate	Std.Error	z	p–value

Intercept	4.45746033	0.603673283	7.383895	1.537243 e⁻¹³
Weight	0.01390312	0.002860473	٤,٨٦٠٤٢٦	1.171333 e⁻٠٦
Matrial Status	−0.50197682	0.123113616	−4.077346	4.555266 e⁻٠٥
Consanguinity	0.13265740	0.028205316	٤,٧٠٣٢٧٧	2.560192 e⁻٠٦
Hemoglobin	−0.12845506	0.021858811	−5.876580	4.188284 e⁻٠٩
AIC	635.05			

من النتائج المعروضة في الجداول رقم (٢) , (٣) , (٤) حاول الباحث الاعتماد على اختبار (P-value) لتحديد المتغيرات التفسيرية المهمة , لكنه اكتشف عدم الجدوى من الاعتماد على هذا الاختبار لان معظم المتغيرات سوف تظهر على انها معنوية. لاحظ ايضا ان قيم اختبارات (Estimate , Std.Error) لنموذج بواسون مشابهة لنموذج ذي الحدين السالب , وهذا دليل على استحالة امكانية الاعتماد على هذين الاختبارين لمعرفة النموذج المناسب. اعتمد الباحث على اختبار (Likelihood Ratio) الموجود في جدول رقم (١) لاختيار المتغيرات التفسيرية المهمة.

جدول رقم (٥) يوضح قيم معايير المعلومات لنماذج (بواسون الكامل , بواسون , شبه بواسون , ذي الحدين السالب) للبيانات الحقيقية .

Criteria	Full Model	Poisson	Quasi Poisson	Negative Binomial
AIC_n	4.146475	4.084187	4.068418	4.097104
AIC	642.703693	633.049048	630.604860	635.051173
BIC_{qh}	4.408983	4.123506	4.107737	4.136423
BIC	682.268220	648.266174	645.821985	650.268299

جدول رقم (٥) قام الباحث بإدخال المتغيرات الاربعة المهمة في نماذج (بواسون , شبه بواسون وذو الحدين السالب) , وتم اجراء عملية اختيار المتغيرات على النماذج الثلاثة فضلا عن نموذج بواسون الذي يحتوي على كل المتغيرات (Full Model). تم استخدام معايير معلومات Akaike للتأكد من النموذج

KUFA STUDIES CENTER JOURNAL
UNIVERSITY OF KUFA
مجلة مركز دراسات الكوفة - جامعة الكوفة
KSCJ OPEN ACCESS
ISSN (P) : 1993-7016
ISSN (E) : 2708-728X
www.journal.uokufa.edu.iq

الافضل. لاحظ الباحث ان قيم معايير المعلومات لنموذج بواسون الكامل (Full Model) هي اعلى القيم الموجودة في الجدول فقيمة معيار AIC_n هي (٤,١٤٦٤٧٥) , وقيمة معيار AIC هي (٦٤٢,٧٠٣٦٩٣) , وقيمة معيار BIC_{qh} هي (٤,٤٠٨٩٨٣) , وقيمة معيار BIC هي (٦٨٢,٢٦٨٢٢٠) , هذا دليل على ضعف هذا النموذج. اما في نموذج بواسون فكانت لدينا نتيجة افضل من نموذج بواسون الكامل , فكانت قيمة معيار AIC_n هي (٤,٠٨٤١٨٧) , وقيمة معيار AIC هي (٦٣٣,٠٤٩٠٤٨) , وقيمة معيار BIC_{qh} هي (٤,١٢٣٥٠٦) , و قيمة معيار BIC هي (٦٤٨,٢٦٦١٧٤) , اما بالنسبة لنموذج شبه بواسون فنلاحظ ان قيمه هي اقل القيم الموجودة في الجدول ؛ اذ ان قيمة معيار AIC_n تساوي (٤,٠٦٨٤١٨) , وقيمة معيار AIC هي (٦٣٠,٦٠٤٨٦٠) , وقيمة معيار BIC_{qh} هي (٤,١٠٧٧٣٧) , و قيمة معيار BIC هي (٦٤٥,٨٢١٩٨٥) , وهذا دليل على ان نموذج شبه بواسون هو الافضل مقارنة ببقية النماذج.

الاستنتاجات:

نستنتج مما تقدم بأن اختيار المتغيرات يكون أفضل من اختيار نموذج كامل. لذا نوصي الباحثين باعتماد نموذج شبه بواسون لمعالجة مشكلة عدد مرات اعطاء الدم لمرضى الثلاسيميا في مركز الثلاسيميا وامراض الدم في محافظة النجف الاشرف.

المراجع:

أولاً– المراجع العربية:

١. باستخدام أنموذج بواسون. Journal of Administration and Economics, (١١١), ٢٤١-٢٥٥.

٢. عبد المجيد حمزة الناصر , احلام احمد جمعة (٢٠٠٧) .. المقارنة بين طرائق تحديد رتبة انموذج الانحدار الذاتي الطبيعي باستخدام بيانات مولدّة وبيانات لبعض العناصر المناخية في العراق). Journal of Economics and Administrative Sciences, ١٣(٤٨), ٢٥١-٢٥١.

ثانياً– المراجع الإنكليزية

١. Acquah, H. D. G. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship.

٢. AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In ٢nd International Symposium on Information Theory Akademiai Kiado. (pp. 267-281).

٣. Amaliana, L., Sa'adah, U., & Wardhani, N. W. S. (2017, December). Modeling Tetanus Neonatorum case using the regression of negative binomial and zero-inflated negative binomial. In Journal of Physics: Conference Series . IOP Publishing. (Vol. ٩٤٣, No. ١, p. 012051).

٤. Blocker, C., Conway, J., Demortier, L., Heinrich, J., Junk, T., Lyons, L., & Punzig, G. (٢٠٠٦). Simple facts about p-values. CDF٨٠٢٣.

٥. Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. Journal of the Royal Statistical Society: Series C (Applied Statistics), ٣٣(١), ٣٨-٤٤.

٦. Brillinger, D. R. (1986). A biometrics invited paper with discussion: The natural variability of vital rates and associated statistics. Biometrics, ٦٩٣-٧٣٤.

٧. Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociological methods & research, ٣٣(٢), ٢٦١-٣٠٤.

٨. Dean, C. B. (1992). Testing for overdispersion in Poisson and binomial regression models. Journal of the American Statistical Association, 87(418), 451-457.

٩. Dean, C. B., & Lundy, E. R. (٢٠١٤). Overdispersion. Wiley StatsRef: Statistics Reference Online, ١-٩.

١٠. Fisher, R. A. (1950). 236: The Significance of Deviations From Expectation in a Poisson Series.

١١. Frome, E. L., & Checkoway, H. (1985). Use of Poisson regression models in estimating incidence rates and ratios. American journal of epidemiology , ١٢١(٢), ٣٠٩-٣٢٣.

١٢. Frome, E. L., Kutner, M. H., & Beauchamp, J. J. (1973). Regression analysis of Poisson-distributed data. Journal of the American Statistical Association , ٦٨(٣٤٤), ٩٣٥-٩٤٠.

١٣. Hardin, J. W., Hardin, J. W., Hilbe, J. M., & Hilbe, J. (٢٠٠٧). Generalized linear models and extensions. .Stata press.

١٤. Hilbe, J. M. (٢٠١١). Negative binomial regression. .Cambridge University Press.

١٥. Hilbe, J. M. (٢٠١٤). Modeling count data. .Cambridge University Press.

١٦. Ismail, N., & Jemain, A. A. (2007). Handling overdispersion with negative binomial and generalized Poisson regression models. In Casualty actuarial society forum (Vol. 2007, pp. ١٠٣-٥٨). Citeseer.

١٧. Konishi, S., & Kitagawa, G. (٢٠٠٨). Information criteria and statistical modeling.

١٨. Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. Sociological methods & research, ٣٣(٢), ١٨٨-٢٢٩.

١٩. Larsen, R. J., & Marx, M. L. (٢٠٠٥). An introduction to mathematical statistics. Prentice .Hall.

٢٠. Lindén, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. Ecology, ٩٢(٧), ١٤١٤-١٤٢١.

٢١. Manton, K. G., & Stallard, E. (1981). Methods for the analysis of mortality risks across heterogeneous small populations: examination of space-time gradients in cancer mortality in North Carolina counties ١٩٧٠-٧٥. Demography, ١٨(٢), ٢١٧-٢٣٠.

٢٢. Mutua, F. M. (1994). The use of the Akaike Information Criterion in the identification of an optimum flood frequency model. Hydrological Sciences Journa ٢٤٤-٢٣٥ ,(٣)٣٩ .

KUFA STUDIES CENTER JOURNAL
UNIVERSITY OF KUFA
مجلة مركز دراسات الكوفة - جامعة الكوفة
KSCJ OPEN ACCESS
ISSN (P) : 1993-7016
ISSN (E) : 2708-728X
www.journal.uokufa.edu.iq

٢٣. Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics, 4(2), 199-203.

٢٤. Nussbaum, E. M., Elsadat, S., & Khago, A. H. (٢٠٠٨). Best practices in analyzing count data: Poisson regression. Best practices in quantitative methods, ٣٠٦-٣٢٣.

٢٥. Padilla, D. P. (٢٠٠٣). A graphical approach for goodness-of-fit of Poisson model. .University of Nevada, Las Vegas.

٢٦. Snipes, M., & Taylor, D. C. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. Wine Economics and Policy, 3(1), 3-9.

٢٧. Student. (1919). An explanation of deviations from Poisson's law in practice. Biometrika, ٢١١-٢١٥.

٢٨. Thiese, M. S., Ronna, B., & Ott, U. (2016). P value interpretations and considerations. Journal of thoracic disease, 8(9), E928.

٢٩. Ver Hoef, J. M., & Boveng, P. L. (٢٠٠٧). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?. Ecology, ٨٨(١١), ٢٧٦٦-٢٧٧٢.

٣٠. Vogt, A., & Bared, J. (1998). Accident models for two-lane rural segments and intersections. Transportation Research Record, ١٦٣٥(١), ١٨-٢٩.

٣١. Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. Biometrika, 61(3), 439-447.