Al-Ghari Journal of Faculty
of Administration and Economics

إحصاءات اختبارات الدرجة لاكتشاف تباين بواسون الزائد

أ.م.د. حسن سامي عريبي
جامعة القادسية
كلية الإدارة والاقتصاد

الباحث محمد حسين نعمه
جامعة القادسية
كلية الإدارة والاقتصاد

المستخلص

في الحالات التي يكون فيها التباين التجريبي للبيانات أعلى من التباين المتوقع بواسطة النموذج ، تُظهر البيانات العددية التي تم فحصها بموجب افتراض بواسون أو البيانات التي تكون على شكل نسب وفقًا لافتراض توزيع ذي الحدين في بعض الأحيان تشتتًا مفرطًا. على وجه الخصوص ، يقلل من تقدير الأخطاء المعيارية مما يؤدي الى تخفيض قيمة (p-value) واعطاء نتائج غير صحيحة ، ولكن يجب التأكد اولا فيما إذا كان التشتت الزائد حقيقيًا أم ظاهريًا ومن ثم يمكن للمرء استخدام المنهج المناسب. تتضمن هذه الورقة نظرة عامة انتقائية على بعض احصاءات اختبارات النتيجة الموجودة في الأدبيات والمستخدمة للكشف عن تباين بواسون الإضافي. تمت تغطية منهجيات هذه الاختبارات لمساعدة ممارسي الإحصاء على فهم كيفية التعامل مع البيانات العددية وتحديد ما إذا كان هناك تباين إضافي في بواسون (التشتت الزائد). تم اشتقاق الصيغة الأساسية لجميع إحصاءات اختبار النتيجة من النموذج ذي الحدين السالب ، ثم تم تقديم إحصاءات الاختبار المعدلة المقترحة لجعل توزيع هذه الاختبارات مقاربًا للتوزيع الطبيعي القياسي. ومع ذلك ، فإن نتيجة الاختبارات لا يمكن الاعتماد عليها ما لم يتم الحصول على المتوسط والتباين الشرطين الصحيحين.

Al-Ghari Journal of Faculty
of Administration and Economics

Abstract

When the empirical variance in the data is higher than that expected by the model, count data examined under a Poisson assumption or data in the form of proportions under a binomial assumption sometimes demonstrate overdispersion. Underestimating standard errors of covariate e_ects is one of the implications of neglecting overdispersion. This paper involves a selective overview of score test statistics that are existing literature and used to detect the extra-Poisson variation. The methodologies of these tests have been covered to aid the statistics practitioners in understanding how to deal with count data and identify whether there is extra-Poisson variation. The baseline formula of all score test statistics had been derived from the negative binomial model. Subsequently, the proposed adjusted test statistics have been presented to approximate the distribution of these tests to standard normal distribution. However, the test result is unreliable unless the correct conditional mean and variance are obtained.

الكلمات الرئيسية: فرط التشتت ، بواسون ، اختبار النتيجة ، NB-1 ، NB-2

Al-Ghari Journal of Faculty
of Administration and Economics

المقدمة

يبدو أن جاذبية نماذج انحدار بواسون لتحليل البيانات العددية التي تحتوي على أعداد صحيحة غير سالبة موثقة على نطاق واسع في الأدبيات الاقتصادية والطبية. ويرجع السبب في ذلك إلى أن انحدار بواسون لنموذج العدد أساسي ، ومتغير الاستجابة يكون متقطعا. يتعامل توزيع بواسون جيدًا مع البيانات المتقطعة ويفترض أن المتوسط الشرطي لمتغير الاستجابة يساوي التشتت الشرطي لذلك المتغير (Frome et al.;1973 and Frome;1983). واقعا ليس من السهل العثور على مثل هذا الشرط في البيانات الحقيقية ، وخاصة البيانات المتقطعة التي ربما تخلق مشكلة عدم التجانس. ومع ذلك ، عندما يتجاوز التشتت الشرطي في البيانات التشتت الأساسي 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. حظت مشكلة التشتت باهتمام الباحثين. قد تم اقتراح بعض الخيارات لتقدير الإحصاءة $\widehat{\emptyset}$ في الأدبيات ، مثل (Score Statistic) ، (Likelihood Ratio) ، (person statistic) ، (Wald statistic).

البحث الاول : منهجية الدراسة

اولا: مشكلة الدراسة

تكون إحصاءة الدرجة (Score Statistic) التي تم الحصول عليها من النماذج الخطية غير كافية وان 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 عشوائي بشكل كافٍ بواسطة أي نموذج افتراضي موجود. لذا تكون بواقي بيرسون أو الانحراف في النموذج المناسب كبيره جدا في المقام الأول مما يؤدي

Al-Ghari Journal of Faculty
of Administration and Economics

إلى عدم كفاية إحصاءة حسن المطابقة أيضًا (Manton et al 1981; Breslow 1984; Brillinger 1986).

ثانيا : اهمية الدراسة :

تناقش هذه الورقة تأثير النقاط الرافعة على عملية اتخاذ القرار لتطبيق نماذج معينة من التشتت المفرط بناءً على نتيجة الاختبار. على سبيل المثال قد تشير نتيجة الاختبار إلى حالة التشتت المفرط ، في حين أن الحقيقة هي حالة نقص التشتت والعكس صحيح.

المبحث الثاني : الجانب النظري

1. مفهوم التشتت المفرط

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 and Hardin, 2014). قد يؤدي وجود التشتت المفرط إلى استنتاجات غير صحيحة ، ويرجع السبب في ذلك إلى ما يسمى بالتحيز الصاعد والتحيز النازل. التحيز الصاعد يمكن ملاحظته مع المتغيرات المستقلة لأنه يزداد مع أهميتها. يؤدي التحيز النازل إلى التقليل من شأن الانحراف المعياري لتقديرات المعلمات (under estimate) (Ismail and Jemain 2007).

على الرغم من ان التشتت المفرط تناولته الأدبيات الإحصائية لاول مرة من خلال تعليقات Student (1919) ، واقترح Fisher (1950) اختبار جودة الملائمة لتقييم فعالية توزيع بواسون في حالة العينة الواحدة ، أي عندما تؤخذ الاعداد كمتغيرات مستقلة بمتوسط مشترك من

Al-Ghari Journal of Faculty
of Administration and Economics

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 Hilbe and Hardin (2014) 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. 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.

كما هو مذكور في أعلاه تعد القيم الشاذة أحد أسباب التشتت الواضح والتي حفزت ممارسي الإحصاء على استخدام نموذج شبه بواسون أو نموذج ذي الحدين السلبي في أبحاثهم العملية. من المعروف أن هناك نوعين محددين من القيم الشاذة في نموذج الانحدار ، والتي تظهر مع متغير الاستجابة تسميان بالقيم المتطرفة الرأسية ، والنقاط الرافعة التي تحدث في مصفوفة التصميم X. كلا النوعين من القيم الشاذة يفترض أن يكون لهما تأثير في تقدير معلمة توزيع بواسون والتي ربما أدت إلى استنتاجات مضللة.

بشكل عام يتمثل نهج دالة شبه الاحتمال في أنه بدلاً من تحديد توزيع احتمالي للبيانات ، تم اعتبار التباين دالة للمتوسط. لذلك فإنه يحدد ببساطة العلاقة بين المتوسط والتباين من خلال السماح بتضخيم التباين على النحو التالي ، $Var(Y) = \emptyset.\lambda$ حيث $\emptyset$ هي معلمة المقياس أو معامل التشتت.

افترض أن Y_i هو متغير عشوائي متقطع ، و$\lambda = \mu(X)$ ، ومن ثم يمكن حساب معلمة المقياس المقدرة بواسطة صيغة إحصاءة χ^2 كاي سكوير المعممة.

Al-Ghari Journal of Faculty
of Administration and Economics

$$\widehat{\emptyset} = \frac{\chi^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\left[\frac{\left(Y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i}\right] \qquad (1)$$

عندما $\hat{\lambda}_i = \hat{\mu}(X_i)$

دفع تضخيم التباين المقدر بـواسطة $\emptyset$ المؤلفين إلى اقتراح عدد من أشكال التباين من خلال الجمع بين دالة التباين ومعلمة المقياس لنمذجة التشتت المفرط. لذلك فإن الإحصاءة $\widehat{\emptyset}$ هي تقدير غير متحيز تقريبًا لحساب نسبة بيرسون كاي سكوير χ^2 إلى درجة حريته في اكتشاف التشتت المفرط (Rodriguez,2015). عندما يكون ادعاء نموذج بواسون بأن المتوسط مساوي للتباين صحيحًا ، فإن معلمة المقياس $\widehat{\emptyset} \approx 1$.

2. اشتقاق اختبار الدرجة من نموذج ذي الحدين السالب

لنفترض $Y_i \sim poisson(v_i\lambda_i)$ توزيعًا مختلطًا ، حيث Y_i هو متغير استجابة ، v_i هي متغيرات عشوائية مستقلة وموزعة بشكل متماثل تمتلك قيم موجبة فقط مثل $E(v_i) = 1$ و $Var(v_i) = \tau$ ، وفي هذه الحالة كانت صيغة Collings and Margolin (1985) هي $Var(Y_i/x_i) = \lambda + \tau\lambda^2$ والتي تماثل التباين التقليدي لذي الحدين السالب 2 (NB-2). بعد اذ Cameron and Trivedi (1986) إعادا كتابة صيغة التباين المفرط كالتالي $Var(Y) = \lambda + \tau\,\lambda^q$. لذلك يمكن للمرء استخدام اختبار الفرضيات بالاعتماد على قيمة τ ، وهي الفرضية الصفرية لنموذج بواسون عندما $H_0: \tau\alpha = 0$ ضد الفرضية البديلة لنموذج بواسون الإضافي $H_1: \tau > 0$. عندما يتم توزيع $v_i's$ كتوزيع جاما ، يجب أن يكون توزيع $Y_i's$ هو توزيع ذو الحدين السالب. اذا $q = 1$ فأن $Var(Y) = \lambda(1+\tau)$ ، هذا يعني أن التباين يتناسب خطيًا مع المعنى الذي يتطابق مع تباين ذو الحدين السالب 1 (NB-1) وبما يكافئ $\emptyset.\lambda$

Al-Ghari Journal of Faculty
of Administration and Economics

عنـــدما $\emptyset = (1 + \tau)$. لنفتـــرض أن $L(\tau)$ هـــي لوغـــارتيم احتمـــالي مـــرجح لـــ

$y_i \sim NB(\lambda_i, \tau\alpha)$ كدالة لـ τ

$$L(\tau) = \sum_{i=0}^{n} \left[y_i \log \tau\, \lambda_i - (y_i + \tau^{-1}) \log(1 + \tau\lambda_i) + \sum_{t=0}^{y_i - 1} \log(t + \tau^{-1}) - \log y_i! \right] \quad (2)$$

يمكن إعادة كتابة الحد الثالث من المعادلة رقم (2) على النحو التالي:

$$\sum_{t=0}^{y_i-1} \log(t + \tau^{-1}) = \sum_{t=0}^{y_i-1} \log\, \tau^{-1}(1 + \tau t)$$

$$\Rightarrow \sum_{t=0}^{y_i-1} \log\, \tau^{-1} + \sum_{t=0}^{y_i-1} (1 + \tau t) = -y_i \log\, \tau + \sum_{t=0}^{y_i-1} (1 + \tau t) \quad (3)$$

استبدل نتيجة معادلة رقم (3) بدلا من $\sum_{t=0}^{y_i-1} log(t + \tau^{-1})$ في المعادلة رقم (2)

$$L(\tau) = \sum_{i=1}^{n} \left[y_i \log \tau\, \lambda_i - (y_i + \tau^{-1}) \log(1 + \tau\lambda_i) - y_i \log\, \tau + \sum_{t=0}^{y_i-1} (1 + \tau\, t) - \log y_i! \right]$$

$$L(\tau) = \sum_{i=1}^{n} \left[y_i \log\, \lambda_i - (y_i + \tau^{-1}) \log(1 + \tau\lambda_i) + \sum_{t=0}^{y_i-1} (1 + \tau\, t) - \log y_i! \right]$$

$$L(\tau) = \sum_{i=1}^{n} \left[y_i \log\, \lambda_i - y_i \log(1 + \tau\lambda_i) - \tau^{-1} \left\{ \tau\lambda_i - \frac{(\tau\lambda_i)^2}{2} + \cdots \right\} + \sum_{t=0}^{y_i-1} (1 + \tau t) - \log y_i! \right]$$

$$L(\tau) = \sum_{i=1}^{n} \left[y_i \log\, \lambda_i - y_i \log(1 + \tau\lambda_i) - \lambda_i - \frac{\tau{\lambda_i}^2}{2} + \cdots + \sum_{t=0}^{y_i-1} (1 + \tau t) - \log y_i! \right] \quad (4)$$

من معادلة رقم (4) يمكن للمرء أن يستنتج أن $L(0) = y_i\, log\ \lambda_i - \lambda_i - log\, y_i!$

Al-Ghari Journal of Faculty of Administration and Economics

وفقًا لمفهوم Neyman-Pearson التقليدي ، أقوى اختبار يرفض H_0 لقيم $L(\tau) - L(0)$ الكبيرة إذا أخذنا في الاعتبار أن $\lambda_i's$ معروف.

$$L(\tau) - L(0) = \sum_{i=1}^{n} \sum_{t=0}^{y_i-1} [\log(1 + \tau t) - y_i \log(1 + \tau\lambda_i)]$$

على الرغم من توفر $\lambda_i's$ إلا أن أقوى اختبار ثابت لـ H_0 ضد H_1 غير موجود بطريقة أو بأخرى لأن الاختبار لا يكون مستقلاً عن τ. نتيجة لذلك يمكن إجراء اقوى اختبار لـ H_0 ضد H_1 . بناءً على نهج Ferguson (2014) يرفض الاختبار الأقوى المحلي H_0 لقيم g الكبيرة للنتيجة $\frac{\partial L}{\partial \tau}$ عند $\alpha = 0$.

$$\frac{\partial L}{\partial \tau} = \sum_{i=1}^{n} \left[-\frac{y_i\lambda_i}{(1 + \tau\lambda_i)} - \left\{ -\frac{{\lambda_i}^2}{2} + \frac{\tau{\lambda_i}^3}{3} - \cdots \right\} + \sum_{t=0}^{y_i-1} \frac{t}{(1 + \tau t)} \right]$$

$$\frac{\partial L}{\partial \tau}\Big|_{\tau=0} = \sum_{i=1}^{n} \left[-y_i\lambda_i + \frac{{\lambda_i}^2}{2} + \sum_{t=0}^{y_i-1} t \right] = \sum_{i=1}^{n} \left[\frac{{\lambda_i}^2 - 2y_i\lambda_i}{2} - \frac{y_i(y_i - 1)}{2} \right]$$

$$= \sum_{i=1}^{n} \left[\frac{{\lambda_i}^2 - 2y_i\lambda_i + {y_i}^2}{2} - \frac{y_i}{2} \right] = \frac{1}{2} \sum_{i=1}^{n} [(y_i - \lambda_i)^2 - y_i] \qquad (5)$$

يعتبر اختبار فرط التشتت الأساسي والذي يسمى بإحصاء الدرجات الجزئية ويشار إليه على أنه (S)

$$S = \frac{1}{2} \sum_{i=1}^{n} [(Y_i - \lambda_i)^2 - Y_i] \qquad (6)$$

3. إحصاءات اختبارات الدرجة

Al-Ghari Journal of Faculty
of Administration and Economics

تم تقديم العديد من الأساليب في الأدبيات للتعامل مع تباين بواسون الإضافي أو التشتت المفرط وتأثيره على الاختبارات المستندة إلى بواسون ، انظر: *(Lawless 1987 a,b and Cox 1983)*. مثل بيرسون كاي سكوير *(Pearson chi-square)* والانحراف (*deviance*) وإحصاء اختبار الدرجة (*score test statistic*). في هذا القسم سوف نقدم بعض الإحصائيات المألوفة لاختبارات الدرجات والتي استخدمها على نطاق واسع ممارسو الإحصاء.

3.1 إحصائية اختبار الدرجة المعيارية

أشار *Dean and Lawless (1989)* إلى أنه يمكن حساب توزيع S بشكل تقريبي باستخدام نظرية العينة الكبيرة الكلاسيكية ذات الاحتمالية القصوى. لذا فقد أظهرا ذلك عندما $n \to \infty$ ان $S/\sqrt{\frac{1}{2}\sum_{i=1}^{n}\hat{\lambda}_i^2}$ هو إحصاء معياري قابل للمقارنة تقاربيًا عندما يتقارب التوزيع إلى معيار عادي. ومن ثم فإن النموذج المقترح يعتمد على تقدير دقة المتوسط والتباين.

أ. عندما $E(Y_i) = \lambda_i$ و $Var(Y_i) = \lambda_i + \tau\lambda_i^2$ و τ صغير ، يمكن كتابة إحصاءة اختبار الدرجة على النحو التالي:

$$S_1 = \frac{\sum_{i=1}^{n}\left\{\left(Y_i - \hat{\lambda}_i\right)^2 - Y_i\right\}}{\sqrt{2\sum_{i=1}^{n}\hat{\lambda}_i^2}} \qquad (7)$$

ب. عندما $E(Y_i) \backsimeq \lambda_i$ و $Var(Y_i) = \lambda_i + \tau\lambda_i^2$ يمكن استبدال $\hat{\lambda}_i$ بدلا من Y_i ومن ثم اعادة كتابة S_1 كالتالي:

$$S_1^* = \frac{\sum_{i=1}^{n}\left\{\left(Y_i - \hat{\lambda}_i\right)^2 - \hat{\lambda}_i\right\}}{\sqrt{2\sum_{i=1}^{n}\hat{\lambda}_i^2}} \qquad (8)$$

Al-Ghari Journal of Faculty
of Administration and Economics

ت. عندما $E(Y_i) = \lambda_i$ و $Var(Y_i) = \lambda_i(1+\tau)$ و τ كبير بما فيه الكفاية ، يمكن كتابة إحصاءة اختبار الدرجة على النحو التالي:

$$S_2^* = (2\,n)^{-1/2} \sum_{i=1}^{n} \left\{ \frac{(Y_i - \hat{\lambda}_i)^2 - Y_i}{\hat{\lambda}_i} \right\} \qquad (9)$$

جميع الاختبارات لها توزيع مقارب للتوزيع الطبيعي التقليدي. يوضح شكل هذه الإحصاءة أن التقلبات المرتفعة بين القيم المرصودة والمتوقعة سترفع قيمة إحصاءة الدرجة ، مما يعني ضمناً ان التشتت المفرط ناجما عن عدم تجانس البيانات وكذلك المتغيرات الأخرى. إذا كانت إحصاءة الدرجة S_1 أو S_1^* أكبر من (0.95) للتوزيع $N(0,1)$ ، فقد نرفض فرضية التوزع عند مستوى أهمية (0.05) باستخدام اختبار من جانب واحد (Payne et al. ,2018). ركز Collings and Margolin (1985) على إعداد اختبارات قوية ضد فرضيات بديلة تستند إلى نموذج بواسون المختلط المحدد. لقد اشتقوا اختبارات درجاتهم من حالات خاصة لنموذج ذي الحدين السالب واشتقوا نظرية التوزيع المقارب لهذه الاختبارات. بشكل أكثر تحديدًا تم اقتراح ثلاثة اختبارات للدرجات تعتمد على $(\bar{y})$ إما ثابتًا ، (متغير واحد) مع خط الانحدار محسوبًا بدون الحد الثابت. اعتبر Dean (1989) أن هذه الاختبارات هي مجرد امتدادات لاختبارات Fisher (1950) و Maron (1971)عندما يتم توزيع المشاهدات بشكل مستقل ومتماثل وفقًا لتوزيع بواسون. لقد قدموا نهجًا آخر لـ S القياسي الذي يشار إليه بـ S_2 ويتم الحصول على تحديد توزيعه بواسطة كل من إحصاءات التشتت لـ Fisher (1950) و Collings-Margolin (1985) ،

$$S_2 = \frac{\sum_{i=1}^{n}(Y_i - \lambda_i)^2}{\bar{Y}}$$

Al-Ghari Journal of Faculty
of Administration and Economics

عندما $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$

وجد Dean and Lawless (1989) حالات خاصة لتحديد التوزيع المحدودة من S_2 مساوية إلى تلك التي قدمها Fisher (1950) و Collings and Margolin (1985).

افترض أن n هو قيمة ثابته وأن $\lambda_i's \to \infty$ بحيث يتقارب λ_i/λ_+ إلى ثابت موجب.

$$S_3 = n\,\lambda_+^{-1}\left(Y-\hat{\lambda}\right)'\left(Y-\hat{\lambda}\right)$$

هو شكل آخر من أشكال S_2 حيث y و $\hat{\lambda}$ متجهان و $\lambda_+ = \sum_{i=1}^{n}\lambda_i$.

بما أن $\lambda_+^{-1}\sum_{i=1}^{n} y_i \xrightarrow{p} 1$ عندما $\tau = 0$ ، فإن كلا من إحصائات S_2 و S_3 لها نفس التوزيع المحدد ،

ثم $\sqrt{\lambda_+}\left(Y-\hat{\lambda}\right) \xrightarrow{D} N(0,\Delta)$ عندما $\Delta = \lambda_+^{-1}\sqrt{W}\,(I-H)\sqrt{W}$ هي مصفوفة التغاير المقاربة.

$W = diag(\lambda_1, \lambda_2, \ldots., \lambda_n)$ و $H = \sqrt{W}X(X'WX)^{-1}X'\sqrt{W}$ هي مصفوفة الرافعات لانحدار بواسون. نظرا لتقدير معاملات الانحدار في ظل وجود نقاط رافعة تؤدي إلى جعل $E\{\sum_{i=1}^{n}\left(Y_i-\hat{\lambda}_i\right)^2\} < \sum_{i=1}^{n}\lambda_i$.

Al-Ghari Journal of Faculty
of Administration and Economics

يمكن كتابة $E\left\{\left(Y_i-\hat{\lambda}_i\right)^2\right\}=(1-h_{ii})\lambda_i$ عندما h_{ii} هي i^{th} العناصر القطرية لـ H.

وبالتالي فإن التوزيع الفارغ لـ S_1 و S_2 من $\sqrt{\lambda_+}\left(Y-\hat{\lambda}\right)\xrightarrow{d}N(0,\Delta)$ هي مجموعة خطية من المتغيرات العشوائية المستقلة $\chi^2_{(1)}$ (Rao,1973)

على وجه الخصوص $S_2\xrightarrow{\text{converg}}\sum_{i=1}^{n}\pi_i\gamma_i$ عندما $\pi_i's\sim\chi^2_{(1)}$ هي متغيرات عشوائية مستقلة ، $\gamma_i's$ هي قيم eigne لـ n لـ Δ

3.2 إحصائية اختبار النتيجة القياسية المعدلة

نتيجة لبطء توزيع S_1, S_1^*, S_2^* ليكون تقريبيًا كتوزيع طبيعي قياسي ، عندما $n\longrightarrow\infty$ Dean and Lawless (1989) و Dean and Lundy (2016) قاموا بتعديل هذه الإحصائيات بنفس الخصائص المقاربة للعثور على المستويات الهامة (p-value) والقيم الحرجة.

لنبدأ في تعديل إحصائيات S_1 لتعكس افتراض توزيع أكثر ملاءمة وكالتالي :

$$S_\alpha=\frac{\sum_{i=1}^{n}\{(Y_i-\hat{\lambda}_i)^2-Y_i+h_{ii}\hat{\lambda}_i\}}{\sqrt{2\sum_{i=1}^{n}\hat{\lambda}_i^2}} \qquad (10)$$

باستخدام التوزيع المحدود لـ S_2 عندما $\lambda_i's\rightarrow\infty$ ، أدخل المؤلفون السابقون $\hat{\lambda}_i$ في مصفوفة Δ وبالتالي في القيم الذاتية (eigenvalues) ، γ_i لتقريب $\Pr\{S_2>s\}$. ثم حساب احتمال المتغيرات العشوائية المستقلة $\chi^2_{(1)}$. لهذا الغرض استخدم المؤلفون عزمين تقريبيين $c\chi^2_{(d)}$ للقيام بهذا الإجراء. إن متوسط وتباين التوزيع المحدود لـ S_2 هما :

$$\bar{S_2}=n\,tr(\Delta)=n\,\lambda_+^{-1}(1-h_{ii})\lambda_i$$

Al-Ghari Journal of Faculty
of Administration and Economics

$$Var(S_2) = 2n^2\, tr(\Delta'\Delta) = 2n^2\lambda_+^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j(1-h_{ii})^2$$

اختيار التقريب $S_2 \sim c\chi^2_{(d)}$ بحيث يكون المتوسط المقارب والتباين صحيحين سوف يسمح بتقدير c و d على النحو التالي:

$$\hat{c} = n\,\lambda_+^{-1}\,\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\lambda}_i\hat{\lambda}_j(1-h_{ii})^2}{\sum_{i=1}^{n}(1-h_{ii})\hat{\lambda}_i}$$

$$\hat{d} = \frac{\left\{\sum_{i=1}^{n}(1-h_{ii})\hat{\lambda}_i\right\}^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\lambda}_i\hat{\lambda}_j(1-h_{ii})^2}$$

عندما $\hat{d} \geq 10$ يفضل نقل المتغيرات $\chi^2_{(1)}$ باستخدام تحويل Wilson-Hilferty. ومع ذلك فإن هذا التقريب يعطي التالي :

$$S_b = \sqrt{4.5\,\hat{d}}\left\{\sqrt[3]{S_2/\hat{c}\hat{d}} + 2/9\hat{d} - 1\right\} \qquad (11)$$

تتضمن هذه الصيغة التوزيع الطبيعي القياسي عندما تكون $\hat{c} \doteq 1$ و $\hat{d} \doteq n - p$.

S_b هي إحصائات كافية عندما تكون $\hat{d}$ كبيرة بما يكفي ، وإلا فإن $S_2 \sim c\chi^2_{(d)}$ أكثر ملاءمة منها.

قام Dean and Lundy (2016) بتعديل S_1^* و S_2^* لاختبار الفرضية الصفرية $\tau = 0$ مرتين ، الأولى للنموذج مع $Var(Y_i) = \lambda_i + \tau\lambda_i^2$ وأخرى للنموذج باستخدام $Var(Y_i) = \lambda_i(1+\tau)$ على التوالي ، وكالاتي :

$$S_1^{**} = \frac{\sum_{i=1}^{n}\left\{\left(Y_i - \hat{\lambda}_i\right)^2 - (1-h_{ii})\hat{\lambda}_i\right\}}{\sqrt{2\sum_{i=1}^{n}\hat{\lambda}_i^2}} \qquad (12)$$

Al-Ghari Journal of Faculty
of Administration and Economics

$$S_2^{**} = \frac{1}{(2n)^{1/2}} \sum_{i=1}^{n} \left\{ \frac{\left(Y_i - \hat{\lambda}_i\right)^2 - (1 - h_{ii})\hat{\lambda}_i}{\hat{\lambda}_i} \right\} \qquad (13)$$

ومع ذلك فإن توزيع S_1^{**} و S_2^{**} مقارب $(n \to \infty)$ وموزع $N(0,1)$ ويتقارب مع الحالة الطبيعية (normality) بشكل أسرع من S_1^* وS_2^* . من الملاحظ أنه عندما $\lambda_i \to \infty$ تكون مقاربة للثابت الإيجابي n ، يمكن كتابة S_2^{**} بالشكل التالي :

$$\widetilde{S_2^*} = \sum_{i=1}^{n} \left[\frac{\left(Y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i} \right] \qquad (14)$$

وهو ما يكافئ إحصاء بيرسون في المعادلة رقم (1) مع $\chi^2(n-p)$ كتوزيع محدود.

المبحث الثالث : المناقشة

جميع الاختبارات السابقة نصت على انها اشتقت على افتراض ان تحديد نموذج الانحدار كان دقيقا . وفقًا لذلك هناك فرضيتان بديلتان يجب أخذهما بنظر الاعتبار ، $\lambda_i = \lambda_i(x_i, \beta)$ ، $\mathrm{Var}(y) = \emptyset.\lambda$) هي غير محدد بدقة. وبخلاف ذلك فإن التشتت الظاهر يكون حاضرا على سبيل المثال المتغيرات المشتركة المفقودة (لمزيد من التفاصيل انظر: Hilbe and Hardin, 2014; Dean and (Lundy, 2016. ومع ذلك ما لم يتم وصف λ_i بشكل ملائم ، فأننا لن نستطيع الى حد كبير تفسير إحصاءة بيرسون على أنها تشير إلى التشتت المفرط في النموذج الخطي المعمم. على العكس من ذلك يمكن أن يكون التشتت الظاهر موجودًا ويمكن للمرء أن ينظر إلى الأسباب التي أدت إلى ذلك لأنه قد تكون لا تطابق المنهجية ، أو المتغيرات المفقودة ، او قد لا يكون من المناسب استخدام دالة المتوسط المناسبة ... وغير ذلك.

Al-Ghari Journal of Faculty
of Administration and Economics

المصادر

[1] "Student." An explanation of deviations from PoissonBT)™s law in practice. Biometrika 12, (1919).

[2] Alain N. Detecting Overdispersion in Count Data: Comparison of Tests. (Doctoral dissertation, East West University), (2015).

[3] BreslowN. Extra-Poisson Variation in Log-Linear Models., J R Stat Soc Ser C Appl Stat 33,(1984)

[4] BrillingerD. R. The Natural Variability of Vital Rates and As- sociated Statistics (with discussion). Biometrics 131, (1986).

[5] CollingsB.J., Margolin BH Testing goodness-of-fit for the Poisson assumption when observations are not identically distributed. J. Am. Stat. Assoc 80(930),(1985).

[6] CoxD. R. Some Remarks on Overdispersion. Biometrika, 7011 (1983).

[7] Dean C.,Lawless J.F. Tests for detecting overdispersion in Poisson regression models. J. Am. Stat. Assoc, 84 (1989).

[8] DeanC. Testing for overdispersion in Poisson and binomial regression models. J. Am. Stat. Assoc, 87(1992).

[9] DeanC., Lundy E. R. Overdispersion, Encyclopedia of Clinical Trials. Wiley, N. Balakrishnan, P. Brandimarte, B. Everitt, G. Molenberghs, W. Piegorsch and F. Ruggeri (eds.), Wiley StatsRef: Statistics Reference Online. 19, DOI: 10.1002/9781118445112.stat06788.pub2. , 19 (2016).

[10] FergusonT. S. Mathematical statistics: A decision theoretic approac. Academic

Al-Ghari Journal of Faculty
of Administration and Economics

press.BT‧1I; (2014).

[11] Fisher R. A. The Significance of Deviations From Expectation in a Poisson Series. Biometrics., 6 (1950).

[12] FromeE. L. The Analysis of Rates using Poisson Regression Models. Biometrics, 39, (1983).

[13] HarrisT., Hilbe J., Hardin J. Modeling count data with generalized distributions. Stata J., 14, (2014).

[14] IsmailN., Jemain A. Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. Casualty actuarial society forum, (2007).

[15] Lawless J. F. Regression methods for poisson process data. J. Am. Stat. Assoc, 82, (1987a).

[16] Lawless J. F. Negative Binomial Regression Models. Can. J. Stat., 15, (1987b).

[17] Manton K.,Woodbury M.,Stallard E. Variance components approach to categorical data models with heterogenous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. Biometrics, (1998).

[18] Moran P. A. Maximum Likelihood Estimation in Non-Standard Conditions Proceedings of the Cambridge Philosophical Society, 70, (1971).

[19] Rao C. R. Linear Statistical Inference and Its Application. Wiley, 2, (1973).

[20] Rodriguez G. Models for over-dispersed count data Generalized Linear Models.Princeton University, Web. 14 Oct. 2015. Available at: http://data.princeton.edu/wws509/stata/overdispersion.html, (2015).