الكشف عن القيم الشاذة في نموذج الانحدار الخطي متعدد المتغيرات باستخدام معاينة جبس

د. محمد نذير إسماعيل قاسم **يونس حازم إسماعيل**

قسم الرياضيات / كلية التربية

جامعة الموصل

الاستلام	القبول
2011 / 03 / 22	2011 / 06 / 01

Abstract

This paper deals with finding outliers in a multivariate linear regression model after assuming a model of normal – Wishart distribution. This method is based on the estimation of probability of an outlier for each observation by mixed Bernoulli model with shifting location outlier. We show how to obtain the posterior distribution in the mixed model by Gibbs sampler algorithm. Also the determination of the number of outliers is done by criterion of marginal likelihood distribution. The theoretical results of this research are applied to real data of multivariate linear regression. The results obtained are so encouraging in determining the outliers in these data.

الملخص

يتناول هذا البحث طريقة لإيجاد القيم الشاذة في نموذج الانحدار الخطي المتعدد المتغيرات وذلك بعد افتراض نمذجة التوزيع الطبيعي – وشارت. وأساس هذه الطريقة هو تقدير احتمالية القيمة الشاذة لكل مشاهدة بواسطة نموذج برنوللي المختلط ذو القيمة الشاذة مزاحة الموقع. ونبين كيفية الحصول على التوزيع ألبعدي في النموذج المختلط بواسطة خوارزمية معاينة جبس. يتم أيضا تحديد عدد القيم الشاذة باستخدام معيار توزيع الإمكان الحدي. تم تطبيق النتائج النظرية على بيانات حقيقية لنموذج انحدار خطي متعدد المتغيرات وكانت النتائج مشجعة في تحديد القيم الشاذة في البيانات.

1. مقدمة

يعد موضوع الانحدار متعدد المتغيرات من الأدوات المهمة في تحليل بيانات متعددة المتغيرات في حالة وجود أكثر من متغير استجابة واحد وأكثر من متغير توضيحي واحد. ويشمل تحليل بيانات في العلوم الاقتصادية والاجتماعية والنفسية والهندسية والزراعية والطبيعية والى أخره من التطبيقات.

إن دراسة وكشف القيم الشاذة في بيانات متعددة المتغيرات تعتبر من البحوث النادرة في أدبيات القيم الشاذة. انظر Ekiz (2002). حيث في هذا البحث بشكل خاص ندرس إمكانية استخدام الأسلوب الجديد لطريقة سلسلة ماركوف مونتي كارلو (Monte Carlo Markov Chain) (MCMC) في الكشف عن القيم الشاذة في بيانات متعددة المتغيرات. ومن الرواد الأوائل في اكتشاف القيم الشاذة بواسطة MCMC هما الباحثين Verdinelli and Wasserman (1991).

سوف نستخدم طريقة معاينة جبس لهذين الباحثين في نماذج متعددة المتغيرات ذات القيمة الشاذة مزاحة الموقع Multivariate Location Shift Outlier Models لنبين كيفية الحصول على دوال الإمكان الحدية من خوارزمية معاينة جبس. إن دوال الإمكان الحدية تستخدم في الاختبار البيزي بشكل واسع وكذلك في تشخيص النموذج كما في معيار بيز في السلاسل الزمنية BIC (Bayesian Information Criterion) . كما أن نسبة دوال الإمكان الحدية تعرف عامل بيز Bayes Factor والذي سيأتي ذكره لاحقا.

وكما هو معلوم أن نموذج الانحدار الخطي البسيط ممكن أن يكون ملوثا Contaminated لنموذج فيه قيمة شاذة مزاحة الموقع حيث أن احتمالية الاختلاط تتبع توزيع برنولي بمعلمة غير معروفة (انظر إسماعيل (2005)).

إن نمذجة القيم الشاذة ممكن إيجادها في أدبيات البحوث مثلا (Kitagawa and Akaike(1992), Barnett and Lewis(1984), Pettit and Smith (1985))

والأعمال التي تضمنت دراسة ونمذجة القيم الشاذة في البيانات متعددة المتغيرات مثلا Ekiz (2006), Ekiz (2002), Chaloner and Brant (1988).

يشمل هذا البحث خمسة مباحث الأول هي المقدمة والثاني تحليل الانحدار متعدد المتغيرات ومحتوي على قيمة شاذة والذي يتضمن نموذج الانحدار البيزي بقيم شاذة مضافة. والمبحث الثالث تحديد التوزيعات الاحتمالية بطريقة معاينة جبس لنموذج الانحدار متعدد المتغيرات بقيم شاذة والمبحث الرابع خصص لدالة الإمكان الحدية لنموذج الانحدار متعدد المتغيرات والخامس عامل بيز وأخيرا الجانب التطبيقي.

2. تحليل الانحدار متعدد المتغيرات مع قيم شاذة

لنفرض أن لدينا نموذج تحليل الانحدار متعدد المتغيرات الأتي

$$Y = XB + E \quad \ldots\ldots\ldots\ldots (1)$$

حيث أن $Y' = (Y_1', Y_2', \ldots\ldots, Y_n')$ مصفوفة ذات بعد $n \times p$ وكل صف Y_i ذات بعد p من متغيرات الاستجابة Response Variable وكل متغير استجابة مقاس على n من المشاهدات. X مصفوفة ذات بعد $n \times k$ بحيث أن:

$X' = (X_1', X_2', \ldots\ldots\ldots, X_n')$ ، وكل صف من X تمثل $X_i' = (x_1', \ldots\ldots, x_k')$ المشاهدة i من k من المتغيرات التوضيحية Explanatory Variables و B مصفوفة ذات بعد $k \times p$ والتي تمثل معاملات الانحدار، E مصفوفة الخطأ العشوائي ذات بعد $n \times p$.

نفرض أن Y_i يتوزع توزيع طبيعي، أي أن

$$E_i \sim (0, \Psi) \quad , \quad Y_i \sim N_p[x_i B, \Psi] \quad , \quad i = 1, \ldots\ldots, n$$

حيث Ψ مصفوفة التغاير ذات بعد $p \times p$. سوف نبين الآن كيفية توسيع هذا النموذج إلى نموذج متعدد المتغيرات بقيمة شاذة.

و أول نموذج من هذا النوع هو نموذج القيم الشاذة مزاحة الموقع أحادي المتغير Univariate Location Shift Outlier Model والذي تم تحليله بواسطة معاينة جبس من قبل Verdinelli and Wasserman (1991) أيضا انظر إسماعيل (2005).

إن نموذج متعدد المتغيرات بقيمة شاذة مزاحة الموقع مع n من متغيرات الدليل Indicator Variables $z_1, z_2, \ldots\ldots\ldots, z_n$ والتي تتوزع توزيع برنوللي ممكن كتابته بالشكل الأتي:

$$f(Y_i) = P(z_i = 0) f(Y_i | z_i = 0) + P(z_i = 1) f(Y_i | z_i = 1) \quad \ldots\ldots\ldots\ldots\ldots (2)$$

$$= (1 - \varepsilon) N_p[Y_i | x_i B, \Psi] + \varepsilon N_p[Y_i | a_i + x_i B, \Psi] \quad \ldots\ldots\ldots\ldots\ldots (3)$$

حيث أن Y_i هي الصف i لمصفوفة البيانات Y و a_i تمثل الصف i ذات بعد p من مصفوفة إزاحة الموقع A والتي بعدها $n \times n$ بحيث $A' = (a', \ldots\ldots, a')$ ، وسيكون للقيم الشاذة متعددة المتغيرات المصفوفة $D_z = diag(z_1, z_2, \ldots\ldots, z_n)$. ونفرض أن كل متغير دليل يتوزع توزيع برنولي بالمعلمة ε والتي تمثل احتمالية كون المشاهدة i شاذة.

وإذا كانت المشاهدات مستقلة فان النموذج الاحتمالي للانحدار متعدد المتغيرات بقيم شاذة له الشكل الأتي:

$$\mathrm{Y} \sim N_{n \times p}[\mathrm{X}B + D_z A \, , \, \Psi \otimes I_n] \quad \ldots\ldots\ldots\ldots (4)$$

ومنقول Y يكون

$$\mathrm{Y}' \sim N_{P \times n}[B' \mathrm{X}' + A' D_z \, , \, I_n \otimes \Psi] \quad \ldots\ldots\ldots\ldots (5)$$

حيث أن $\otimes$ تمثل حاصل ضرب كرونيكر. انظر Rao and Kleffe (1988) .

المعلومات الأولية (أو السابقة) عن مصفوفات المعلمة ممكن صياغتها بشكل محكم وكالأتي:

$$B \sim N_{k\times p}\left[B_*, G_* \otimes H_*\right]$$
$$\Psi^{-1} \sim W_p\left[\Psi_*, n_*\right]$$
$$A \sim N_{n\times p}\left[A_*, P_* \otimes I_n\right]$$
$$z_i \sim Ber\left[\varepsilon_i\right] \quad i = 1,2,.....,n$$

حيث أن $0 < \varepsilon_i < 1$ الاحتمالية الأولية لمعلمة توزيع برنولي. المصفوفة A_* ذات بعد $n\times p$ والمصفوفة P_* ذات بعد $p\times p$ تمثلان المعلومات الأولية لمعلمتي الموقع (الوسط) والتباين للقيم الشاذة. و ε_i احتمالية النجاح للمشاهدة i أن تكون شاذة وعادة تكون ثابتة بقيمة ε .

3. معاينة جبس في نموذج الانحدار الاحتمالي متعدد المتغيرات بقيم شاذة

في البداية لابد من الإشارة إلى مقدمة تعريفية إلى معاينة جبس حيث تعود معاينة جبس إلى عمليات جبس Gibbs Processes المأخوذة من علم الفيزياء والتي تمثل طريقة لبناء عمليات رياضية جديدة من العمليات الرياضية القديمة. إذ أننا نبدأ بعملية ابتدائية $Y^{(0)}$ ثم نعرف عملية جديدة Y وذلك بأخذ الدالة الاحتمالية $p(Y)$ لـ Y بالنسبة إلى $Y^{(0)}$ وللتعرف على كيفية عمل طريقة معاينة جبس افرض انه لدينا k من المتغيرات العشوائية والتي نرمز لها بالرمز

$$Y = \left(Y_1, Y_2, ..., Y_k\right)$$

ونفرض لدينا التوزيعات الشرطية التامة بالشكل:

$$p\left(y_i \middle| y_j\right) \ , \ i \neq j \quad , \quad i, j = 1,2,...,k$$

وفي هذه الخوارزمية لا نحتاج الشكل التام للتوزيعات الشرطية

$$p\left(y_i \middle| y_j\right) \ , \ i \neq j$$

لكن نحتاج فقط كتابتهم بشكل يتناسب وتحت شروط معتدلة Casella and George (1992).

خطوات خوارزمية معاينة جبس

وفيما يلي توضيح موجز لخوارزمية معاينة جبس:

1. افرض لدينا المتغيرات العشوائية

$$\{Y_1, Y_2, ..., Y_k\}$$

ولها القيم الابتدائية

$\{Y_1^{(0)}, Y_2^{(0)}, ..., Y_k^{(0)}\}$

2. اسحب $Y_1^{(1)}$ من التوزيع الشرطي

$\{Y_1 | Y_2^{(0)}, Y_3^{(0)}, ..., Y_k^{(0)}\}$

3. اسحب $Y_2^{(1)}$ من التوزيع الشرطي

$\{Y_2 | Y_1^{(1)}, Y_3^{(0)}, ..., Y_k^{(0)}\}$

.

.

4. k. اسحب $Y_k^{(1)}$ من التوزيع الشرطي

$\{Y_k | Y_1^{(1)}, Y_2^{(1)}, ..., Y_{k-1}^{(1)}\}$

وهذا يكمل تكرار واحد لمعاينة جبس وعليه بعد تكرار واحد نكون قد حصلنا على $\{Y_1^{(1)}, Y_2^{(1)}, ..., Y_k^{(1)}\}$

وبعد t مثل هذه التكرارات نكون قد حصلنا على $\{Y_1^{(k)}, Y_2^{(k)}, ..., Y_k^{(k)}\}$

وبتكرار هذه العملية m من المرات فان معاينة جبس ممكن أن تولد مشاهدات مستقلة عددها m للمتغير مثلا Y_1 والذي يعتبر عينة عشوائية من التوزيع $p(y_1)$ بشرط أن تكون التكرارات مستقلة وباستخدام قيم ابتدائية $\{Y_1^{(0)}, Y_2^{(0)}, ..., Y_k^{(0)}\}$

وبعد الحصول على الخصائص الإحصائية لهذا المتغير مثلا الوسط والتباين أو أية خواص إحصائية أخرى لهذا المتغير Y_1 وهكذا بالنسبة للمتغير Y_2 و Y_3 والى Y_k انظر إسماعيل (2005).

وألان نعود إلى النموذج في (1) والذي يمكن كتابته بالشكل:

$$\underset{(n\times p)}{\mathrm{Y}} = \underset{(n\times K)}{\mathrm{X}}\ \underset{(K\times p)}{B} + \underset{(n\times p)}{E} \qquad (6)$$

افرض أن A و D كما معرفة سابقا في (2). ممكن صياغة النموذج بعد افتراض التوزيع الطبيعي بالشكل:

$$\mathrm{Y} \sim N_{n\times p}[\mathrm{X}B + D_z A, \Psi \otimes I_n]$$

نفرض أن $\theta = (B, \Psi, A, D_z)'$ متجه يمثل المعلمات B, Ψ, A, D_z.

التوزيع المشترك للمتغير Y والمعلمة θ يكون

$$p(\mathrm{Y}, \theta) = N_{n\times p}[\mathrm{Y}|\mathrm{X}B + D_z A, \Psi \otimes I_n].N_{k\times p}[B|B_*, G_* \otimes H_*].W_p[\Psi^{-1}|\Psi_*, n_*]$$
$$.N_{n\times p}[A|A_*, P_* \otimes I_n].\sum_{i=1}^{n} Ber(Z|\varepsilon_i)$$

نستخدم في الأتي θ^c لتمثل الحجة (Argument) الشرطية في التوزيع الشرطي التام لمجموعة المعلمة θ بدون المعلمات الحالية.

يمكن إيجاد التوزيعات الشرطية التامة لمعاينة جبس كما يأتي:

أ) لمصفوفة معاملات الانحدار B:

$$p\left(B|Y,\theta^{c}\right)=N_{K\times p}\left[B_{**},C_{**}\right]$$

حيث أن:

$$C_{**}^{-1}=G_{*}^{-1}\otimes H_{*}^{-1}+\Psi^{-1}\otimes X^{'}X,$$
$$vec\,B_{**}=C_{**}\left[vec\left(G_{*}^{-1}B_{*}H_{*}^{-1}+X^{'}\left(Y-D_{z}A\right)\Psi^{-1}\right)\right]$$

حيث أن vec عملية رياضية تحول المصفوفة ذات السعة $n\times p$ إلى متجه ذو بعد np عنصر بشكل عمود واحد تحت الأخر ابتداء من الصف الأول.

ب) لمصفوفة التغاير Ψ:

$$p\left(\Psi^{-1}\middle|Y,\theta^{c}\right)=W_{p}\left[\Psi_{**},n_{**}=n_{*}+n\right]$$

يمثل توزيع وشارت ذات بعد p وبمعلمة قياس

$$\Psi_{**}=\Psi_{*}+\left(Y-XB-D_{z}A\right)\left(Y-XB-D_{z}A\right)^{'}$$

تعريف: إذا كانت المصفوفة $M\left(p\times p\right)$ ممكن كتابتها بالشكل $M=x^{'}x$ حيث أن x مصفوفة ذات بعد $(m\times p)$ وتمثل مصفوفة بيانات من توزيع $N_{p}\left[0,\Sigma\right]$ فان M يتوزع توزيع وشارت بمصفوفة قياس (scale) ومعلمة درجات الحرية m ويكتب $M\sim W_{p}\left[\Sigma,m\right]$ وإذا كانت $U=M^{-1}$ فان $U\sim W_{P}^{-1}\left[\Sigma,m\right]$.

انظر Mardia et al(1997), Johnson and Wichern (2007)

ج) لمصفوفة إزاحة الموقع A:

$$p\left(A|Y,\theta^{c}\right)=N_{n\times n}\left[A_{**},G_{**}\right]$$

توزيع طبيعي متعدد المتغيرات بالمعلمات

$$C_{**}^{-1}=P_{*}^{-1}\otimes I_{n}+\Psi^{-1}\otimes D_{z}^{'}D_{z},$$
$$vec\,A_{**}=G_{**}\left[vec\left(A_{*}\,P_{*}^{-1}+D_{z}\left(Y-XB\right)\Psi^{-1}\right)\right].$$

ولكل مشاهدة فان الوسط ألبعدي Posterior Mean يمكن حسابه وذلك بتجزئة نظام المعادلات إلى معادلات بمتغير واحد:

$$C_{**i}^{-1}=P_{*}^{-1}+z_{i}^{2}\Psi^{-1},\ i=1,.....,n,$$
$$a_{**_{i}}=G_{**_{i}}\left[P_{*}^{-1}a_{*_{i}}+z_{i}\Psi^{-1}\left(y_{i}-Bx_{i}\right)\right].$$

د) لمتغيرات المؤشر z_{i}:

$$p\left(z_{i}\middle|Y,\theta^{c}\right)=Ber\left[\varepsilon_{i*}=\frac{c_{i}}{c_{i}+d_{i}}\right],\quad i=1,.....,n,$$

توزيع برنولي بالمركبات التي نحصل عليها بطريقة تطبيق نظرية ببز، أي بمعنى

$$c_{i}=N_{p}\left[y_{i}|x_{i}B+a_{i},\Psi\right].\varepsilon_{i},$$
$$d_{i}=N_{p}\left[y_{i}|x_{i}B,\Psi\right].\left(1-\varepsilon_{i}\right),\quad i=1,.....n,$$

حيث أن x_{i} هو الصف i في المصفوفة X و a_{i} يمثل الصف i من المصفوفة A.

4. حساب الإمكان الحدي لنموذج الانحدار متعدد المتغيرات بالقيم الشاذة

باستخدام طريقة Chib (1995) من الممكن إيجاد الإمكان الحدي عند النقطة $\hat{\theta}_0$ والتي هي:

$$\hat{\theta}_1 = \left(\hat{\theta}_0, \hat{A} = 0, \hat{D}_z = 0\right) \qquad \text{.....(7)}$$

حيث $\hat{\theta}_0 = \left(\hat{B}, \hat{\Psi}\right)$ نفس النقطة لنموذج الانحدار متعدد المتغيرات بدون قيم شاذة. وألان يكون لدينا التجزئة الآتية:

$$p\left(\hat{\theta}\,\middle|\,Y\right) = p\left(\hat{D}_z\,\middle|\,Y\right).p\left(\hat{A}\,\middle|\,\hat{D}_z, Y\right).p\left(\hat{\theta}_0\,\middle|\,\hat{A}, \hat{D}_z, Y\right) \qquad \text{.....(8)}$$

و

$$p\left(\hat{\theta}\right) = p\left(\hat{\theta}_0\right).N\left[A_*, I_n \otimes P_*\right]\prod_{I=1}^{n} Ber(\varepsilon). \qquad \text{.....(9)}$$

انظر Polasek (1999).

1. نستخدم دورة جبس من J من نقاط العينة لنموذج الانحدار متعدد المتغيرات بالقيم الشاذة لحساب المقدار

$$p\left(\hat{D}_z\,\middle|\,Y\right) = \int \prod_{i=1}^{n} Ber\left(\varepsilon_i\right) p\left(\theta\,|\,Y\right) d\theta$$

$$= \frac{1}{J}\sum_{j=1}^{J}\prod_{i=1}^{n} Ber\left(\varepsilon_i^{(j)}\right) \qquad \text{.....(10)}$$

حيث أن معلمة الكثافة البعدية في الموقع i لتوزيع برنولي معطاة من:

$$\varepsilon_{i**}^{(j)} = \frac{c_i^{(j)}}{c_i^{(j)} + d_i^{(j)}}$$

$$c_i^{(j)} = N_p\left[y_i\,\middle|\,\varepsilon^{(j)} z_i^{(j)} + B_j\, x_i^{(j)}\right],$$

$$d_i^{(j)} = N_p\left[y_i\,\middle|\,B^{(j)}\, x_i^{(j)}\right]$$

2. المركبة الثانية ممكن الحصول عليها بدون معاينة جبس:

$$p\left(\hat{A}\,\middle|\,\hat{D}_z = 0, Y\right) = N_{n\times p}\left[\hat{A}\,\middle|\,A_{**}, I_n \otimes G_{**}\right] = \prod_{I=1}^{n} N\left[\hat{a}_i\,|a_{i**}, G_{**}\right]. \qquad \text{.....(11)}$$

ومن الممكن ملاحظة من التوزيع الشرطي التام لمعلمة إزاحة الموقع A عند وضع $\hat{D}_z = 0$، فان التوزيع الشرطي يساوي التوزيع الأولي:

$$G_{**} = I_n \otimes P_n$$

وكذلك

$$vec\, A_{**} = vec\, A_*$$

لمزيد من التفاصيل انظر Chib (1995)

3. ممكن الحصول على الدالة الحد في المعادلة (6) والذي هو $p\left(\hat{\theta}_0\,\middle|\,\hat{A}, \hat{D}_z, Y\right)$ بواسطة احتساب الإمكان الحدي (الهامش) لنموذج الانحدار المتعدد المتغيرات بدون القيم الشاذة.وبعد أخذ الوغاريتم نحصل على النتيجة التالية دالة الإمكان ينتج:

$$\log p(Y) = \log p\left(Y|\hat{\theta}_1\right) + \log p\left(\hat{\theta}_1\right) - \log p\left(\hat{\theta}|Y\right) \quad \text{.....(12)}$$

حيث أن جزء الإمكان معطى بواسطة

$$p\left(Y|\hat{\theta}_1\right) = N\left[Y|\hat{X}\hat{B}, \hat{\Psi} \otimes I_n\right] \quad \text{.....(13)}$$

والذي له نفس قيمة الانحدار متعدد المتغيرات بدون قيم شاذة بسبب $\hat{D}_g = 0$.

ولاحظ إن الصيغة (10) هي تبسيط بسبب أن مركبات إزاحات الموقع تختصر.

ولمزيد من التفاصيل التي تحتاج الى عمليات جبرية طويلة في الحصول على المعادلة (12) انظر ملحق بحث Polasek and Ren (1998).

5. اختيار النموذج مع عوامل بيز

المتراجحات البعدية تستخدم في التحليل البيزي لاختبار نموذجين أو أكثر لنفس مجموعة البيانات. الصيغة الأساسية للاختبار النموذجين M_1 , M_2 هي

المتراجحات البعدية = عامل بيز × المتراجحات الأولية

Posterior odds = Bayes factor. Prior odds

أو

$$\frac{p(M_1|Y)}{p(M_2|Y)} = BF \cdot \frac{p(M_1)}{p(M_2)}$$

حيث أن $p(M_1|Y)$, $p(M_2|Y)$ هما الاحتمالات البعدية للنموذجين M_1 , M_2 على التوالي. $p(M_1)$, $p(M_2)$ هي الاحتمالات الأولية للنماذج M_1 , M_2 على التوالي.وفي ابسط حالة(اقل المعلومات الأولية) يكونان متساويين. وفي هذه الحالات تكون المتراجحات البعدية تساوي عامل بيز والذي يمكن تعريفه على انه النسبة لدالة الإمكان ألبعدي

$$BF = \frac{p(M_1|Y)}{p(M_2|Y)} = \frac{\int p(Y,\theta_1)d\theta_1}{\int p(Y,\theta_2)d\theta_2} \quad \text{.....(14)}$$

حيث أن θ_1 , θ_2 تمثلان معلمتي النموذجين M_1 , M_2 على الترتيب. فإذا كان $BF > 1$ نختار النموذج M_1 وإذا كان $BF < 1$ نختار النموذج M_2. مثلا عامل بيز لاختيار نموذج يحتوي على قيم شاذة مقابل نموذج بدون قيم شاذة سيعرف على انه:

$$BF = \frac{p(Y \mid outliers)}{p(Y \mid no\ outliers)}$$

إن عامل بيز يمكن حسابه من الصيغة المعرفة

$$BF = \frac{P_i F_p(a)}{(1-P_i)(1-F_p(a))} \quad \text{.....(15)}$$

حيث P_i تمثل الاحتمالية البعدية للصيغة التربيعية للخطأ المشاهد وان $F_p(k)$ هي احتمال أن تكون القيمة المشاهدة شاذة و k تمثل قيمة حرجة تؤخذ من جدول توزيع مربع كاي استنادا إلى عدد المتغيرات المعتمدة وحجم العينة المدروسة. عبد الغني (2010).

من الواضح أن عامل بيز يعتمد بشكل أساسي على الاحتمالية البعدية للصيغة التربيعية للخطأ المشاهد والمعرفة بالشكل:

$$P_i = E_{\Sigma/Y}\left\{ p_r\left(T_i > \frac{a}{\sigma_{(i)}} | \Sigma, Y \right) \right\} \qquad \text{.....(16)}$$

انظر عبد الغني (2010).

المتراجحات البعدية لنموذج القيمة الشاذة نحصل عليه من عامل بيز بواسطة:

$$R = BF \times r \qquad \text{.....(17)}$$

حيث أن R المتراجحات البعدية و r المتراجحات الأولية.

من المعادلة (17) وبأخذ اللوغارتم تصبح:

$$\log R = \log BF + \log r$$

حيث

$$r = \frac{p(outlier)}{1 - p(outlier)}$$

$$\Rightarrow p(outlier) = \frac{r}{1+r} \qquad \text{.....(18)}$$

وهذه المعادلة تمثل احتمالية القيمة شاذة.

الجانب التطبيقي: تم تطبيق الجانب النظري على بيانات حقيقية لنموذج انحدار متعدد المتغيرات والتي تمثل قياسات نسبة السكر في الدم y_1 والمقاس بوحدة الملي مول (m.mol)، وضغط الدم العالي y_2 والواطي y_3 والمقاس بوحدة الملي لتر زئبقي كمتغيرات استجابة (معتمدة)، والعمر x_1 مقاس بالسنة، والوزن x_2 مقاس بالكيلو غرام، والطول x_3 مقاس بالسنتمتر كمتغيرات توضيحية (مستقلة) لعينة من المرضى.إن هذه البيانات أخذت من عبد الغني (2010).والبيانات موضحة في الجدول الأتي:

الجدول (1): قياسات نسبة السكر في الدم وضغط الدم العالي والواطي مع المتغيرات التوضيحية (العمر والوزن والطول) لعينة من المرضى.

التسلسل	السكر	الضغط		العمر	الوزن	الطول
		العالي	الواطي			
.1	5.0	12	8	49	66	164
.2	4.7	11	8	45	107	160
.3	1.9	16	8	18	80	158
.4	6.0	12	8	50	87	162

147	95	69	8	16	23.3	**.5**
155	80	60	8	13	7.2	**.6**
162	80	72	8	13	5.5	**.7**
155	92	59	8	12	10.1	**.8**
166	85	53	7	14	6.2	**.9**
166	68	55	8	10	5.3	**.10**
154	77	50	6	12	3.8	**.11**
163	80	50	6	12	6.8	**.12**
160	90	55	8	12	8.1	**.13**
161	80	50	8	12	13.6	**.14**
156	80	54	8	15	15.7	**.15**
171	72	40	8	12	6.1	**.16**
167	75	59	8	12	5.4	**.17**
160	70	42	7	9	5.0	**.18**
170	90	47	9	12	26.7	**.19**
147	80	64	7	15	28.1	**.20**

لتكوين نموذج الانحدار المتعدد المتغيرات تم اعتبار أن متغيرات الاستجابة الثلاثة $Y_i = \{Y_{i1}, Y_{i2}, Y_{i3}\}$ التي تمثل نسبة السكر والضغط العالي والواطي تتبع التوزيع الطبيعي متعدد المتغيرات والتي تعتمد على المتغيرات التوضيحية الثلاثة $X_i = \{X_{i1}, X_{i2}, X_{i3}\}$ التي تمثل (الوزن، العمر، الطول). أي أن:

$$Y \sim N_{n\times p}[X\,B + D_z\,A, \Psi]$$

حيث أن A مصفوفة الموقع المزاح وان التوزيع الأولي لـ A هو:

$$A \sim N_{20\times 3}(A_*, P_* \otimes I_{20}) \qquad \ldots\ldots\ldots\ldots(19)$$

وان

$$v(a_i) = P_* = dig(\mathrm{var}(Y)) \qquad , \; i = 1,2,3$$

حيث أن $P_* = dig[60.0167, 3.3053, 0.5368]$.

وان D_z مصفوفة المؤشر للقيم الشاذة وان التوزيع الأولي لـ D_z هو:

$$D_z \sim ber(\varepsilon_i) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(20)$$

حيث تم فرض $\varepsilon_i = 0.1$ قيمة ثابتة لكل القيم للسهولة.

وان التوزيع الأولي لـ B هو:

$$B \sim N_{3\times 3}(B_*, G_* \otimes H_*) \qquad \ldots\ldots\ldots\ldots (21)$$

افرض أن $G_* \otimes H_* = \Gamma_3 \otimes I_3$

حيث أن $\Gamma_3 = dig(\lambda_i) = dig(0.0000225, 0.0006639, 0.0067794)$ وأن λ_i تمثل القيم الذاتية للمصفوفة YY'.والوسط الأولي ببساطة يوضع ليمثل النقطة الوسطى لمتغير القياس والتي تساوي 1.5 بسبب وجود ثلاث متغيرات استجابة. أي أن $B_* = 1.5 * 1_3$.

وان المعلومات الأولية لـ Ψ هي:

$$\Psi^{-1} \sim W_3(\Psi_*, n_*) \qquad .....(22)$$

حيث أن $\Psi_* = diag(\text{var}(y_1), ..., \text{var}(y_3)) = dig(60.0167, 0.5368, 3.3053)$

وان $n_* = 1$.

وبواسطة استخدام البرنامج الجاهز WinBUGS14 إذ تم تكرار هذا النموذج 1000 مرة والحصول على التكرارات لكل من المعلمات A, B, D_z, Ψ .وكذلك الحصول على الاحتمال اللاحق.كذلك تم كتابة برنامج بنظام Matlab6.5 لحساب عامل بيز من المعادلة (15) والمعادلة (18) وكانت النتائج كما موضح في الجدول الأتي:

جدول (2): عامل بيز واحتمالية المشاهدة i بأنها شاذة

التسلسل	عامل بيز BF	$p(outlier)$
.1	0.9754	0.4347
.2	1.4069	0.3478
.3	8.7191	0.0783
.4	2.1536	0.2944
.5	16.2256	0.0427
.6	7.6636	0.1033
.7	1.4854	0.3817
.8	11.2842	0.0724
.9	2.5477	0.2633
.10	1.3677	0.3618
.11	5.8098	0.1254
.12	3.1008	0.2181
.13	10.354	0.0811
.14	12.4165	0.0650
.15	13.5809	0.0600
.16	2.3111	0.2535
.17	1.25	0.4033
.18	1.0542	0.4222
.19	18.7768	0.0468
.20	20.0644	0.0417

نلاحظ من خلال الجدول أعلاه إن المشاهدات ذات التسلسل (3، 5، 8، 13، 14، 15، 19، 20) تملك احتمالية صغيرة جدا مقارنة مع بقية احتمالات المشاهدات الأخرى وعليه يمكن اعتبارها ممكن اعتبار هذه المشاهدات حالات متقدمة من المرض يجب الوقوف عندها من ناحية وصف الدواء واتخاذ الحمية.

الاستنتاجات

1. استخدام طريقة معاينة جبس يعطي نتائج دقيقة في حالة الكشف عن القيم الشاذة.
2. تحديد دالة الإمكان الحدية بشكل كامل باستخدام معاينة جبس.
3. إيجاد دالة الكثافة الحدية البعدية التامة.

التوصيات

1. يمكن استخدام أسلوب معاينة جبس في الكشف عن القيم الشاذة في بيانات تحليل التباين متعدد المتغيرات والتحليل ألعاملي.
2. تطبيق هذا التحليل على بيانات في مجالات العلوم الطبيعية والاقتصادية والنفسية والتربوية والاجتماعية.

المصادر

1) عبد الغني، تميم معاذ (2010): أسلوب بيز في الكشف عن القيم الشاذة في الانحدار الخطي متعدد المتغيرات،رسالة ماجستير غير منشورة، كلية التربية، جامعة الموصل.

2) إسماعيل، يونس حازم (2005): الكشف عن القيم الشاذة باستخدام التوزيع المختلط بالاعتماد على معاينة جبس وأسلوب بيز. رسالة ماجستير غير منشورة، كلية التربية، جامعة الموصل.

3) Barnett, V. and Lewis, T. (1984): Outliers in Statistical Data, Wiley, Chichester.

4) Chib, S.(1995): Marginal Likelihood from the Gibbs Output. JASA, 90, 1313-1321.

5) Casella, G. and George, E. (1992): "Explaining the Gibbs Sampler". The American Statistician, 46, 167-174.

6) Chaloner, K. and Brant, R. (1988): A Bayesian Approach to Outlier Detection and Desidual Analysis, Biometrika 75, 9 – 651.

7) Ekiz, U. (2002): A Bayesian Methods to Detect Outlier in Multivariate Linear Legression. Gazi university Faculty of science, Turkey.

8) Ekiz, U. (2006): Bayesian Approach to Identity Masking and Swamping Problems in Multivariate Analysis. Gazi university Faculty of Science. Turkey.

9) Kitagawa, G. and Akaike, H. (1992): A Quasi Bayesian Approach to Outlier Detection, *Ann. Inst. Stat. Mathematics* 34, 95–104.

10) Johnson, A.R. and Wichern, D.W. (2007): Applied Multivariate Statistical Analysis. Pearson Prentice Hall. United States of America.

11) Mardia, K.V. and Kent, J.T. and Bibby, J.M.(1997): Multivariate Analysis. Academic press London.

12) Pettit, L.I. and Smith, A.F.M. (1985): Outliers and Influential Observations in Linear Models. In Bayesian Statistics, 2, Ed. J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith, 473-494. Amsterdam North Holland.

13) Polasek, W.(1999): Gibbs Sampling in B-VAR Models with Latent Variables. in: Innovations in Multivariate Statistical Analysis –A Festschrift for Heinz Neudecker, R.J. Heijmans, D. S. G. Pollock and A. Satorra, (eds), Kluwer Academic Publishers, Dordrecht, 137-16-56.

14) Polasek, W. and Ren, L (1998): Structureal Breaks and Model Selection with Marginal Likelihoods, in: W. Racugno (ed). Proceedings of the Workshop on Model Selection, Pitagora Ed. Bologna, 223-273, with discussion.

15) Rao, C.R. and Kleffe, J. (1988): Estimation of Variance Components and application North Holland, Amsterdam.

16) Verdinelli, I. and Wasserman, L.(1991): Bayesian Analysis of Outlier Problems Using the Gibbs Sampler, Statistics and Computing 1991– 1, 105 – 117.