ISSN1813-1719

مجلة تكريت للعلوم الإدارية والاقتصادية

تعنى بالبحوث الإدارية والاقتصادية
والمحاسبية والمعلوماتية

دورية فصلية علمية محكمة

المشاهدات غير المألوفة في نموذج الانحدار

د. محمد طه أحمد الغنام
كلية التربية / جامعة تكريت

الخلاصة
تعاني نماذج الانحدار من مشكلة وجود المشاهدات غير المألوفة (الشواذ ونقاط الانعطاف leverage point) وعندما تكون القياسات واقعية فلا يمكن حذفها حينئذ ويجب أن يتم التحليل بوجودها ، وهكذا فقد بحثنا المشكلة في كيفية تشخيص هذه النقاط وبيان مدى تأثيرها على معالم نموذج الأنحدار وصيغته المقدرة . وقد تطرق البحث لخصائص المصفوفة H وفائدتها في تشخيص نقاط الأنعطاف . وفي الجانب التطبيقي قام الباحث بسحب عينة من مجتمع يعتقد الباحث وجود هذه القحياسات فيه ، فتم تقدير نموذج الأنحدار بطريقة المربعات الصغرى ثم بعد ذلك شخصنا المشاهدات موضوع بحثنا وبيان مدى تاثيرها على معالم النموذج .

Unusual observation in regression model

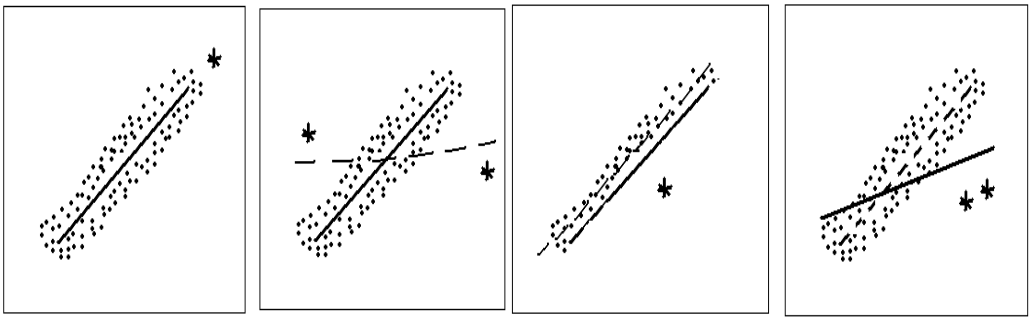
Abstract

The existence of unusual observation (outliers and leverage points) affects regression models and when these points are real they can not be omitted they must be involved in analysis. Thus, I discus this problem to show how to diagnose this kind of points and to display the influence on regression parameter and the estimated model .This paper studies the properties of the important matrix H (hat matrix) that is useful to diagnose the leverage points .In the application part ,I chose a random sample from a population where such points exist as the researcher think. Then the model is estimated by least square method and the leverage and outlier are diagnosed and their influence on the model is shown .

١- مقدمة:

إن احدى الصعوبات في بناء نموذج الانحدار هو وجود المشاهدات غير المألوفة فيه ، والتي تصنف الى نوعين: القيم الشاذة outlier وقيم الانعطاف leverage point وهذه كلها تؤدي الى نتائج مضللة عند تمثيل النموذج للمسألة المدروسة. وعند وجود مثل هذه المشاهدات في بيانات العينة ماذا يكون القرار تجاهها أنبقي عليها أم نحذفها ، وفي حالة الإبقاء عليها فأن هذه البيانات تقلل من كفاءة التحليل بطريقة المربعات الصغرى وستكون بواقي هذه المشاهدات كبيرة مؤثرة في اتجاه خط الأنحدار. أما عند حذفها من البيانات

سوف يكون النموذج غير معبر عن شريحة من مجتمع الدراسة ، عليه يكون الأفضل هو الأبقاء عليها ومعالجتها بطرق معينة.
ان معظم التحليلات التي عالجت مسألة البيانات غير المألوفة (اذا كانت واقعية حقيقية) تعتمد اسلوب تحليل المربعات الصغرى الموزونة التكرارية التي تعمل على تقليل تأثيراتها غير المقبولة.
تنقسم المشاهدات غير المألوفة النوعين:-
- الشواذ (outliers): وهي القياسات غير الطبيعية في قبم المتغير المعتمد y وهذه تجعل مقدراتالمربعات الصغرى تفقد خصائصها الجيدة.
- نقاط الأنعطاف (leverage points) : تمثل القياسات غير الطبيعية في قيم المتغير x والتي تسبب قي سحب خط الأنحدار نحوها بعيدا" عن موقعه الحقيقي .
هذين النوعين قد يتوحدان فيكون تأثيرهما قويا وقد يتعاكسان فيلغي تأثير أحدهما الآخرحزئيا او كليا ، الا أن موقع كل نقطة يختلف تأثيره عن موقع نقطة أخرى ، ولتوضيح هذين التأثيرين على خط الأنحدار لاحظ الأشكال البيانية الآتية :



شكل(1) شكل(2) شكل(3) شكل(4)

في الشكل(١) نقطتين مؤثرتين باتجاه المتغير x لأنهما في جهة واحدة وغير متسقتين مع عموم خط ألأنحدار (الخط المتقطع الممثل للبيانات عدا النقطتين المذكورتين) فسحبته نحوها بموجب (الخط المتصل الذي يمثل كل البيانات)، وفي شكل (٢) هناك نقطة خارج اطار البيانات ولكنها في الوسط فكان التأثير أقل ، وفي شكل (٣) تأثر خط ألأنحدار ومعالمه كثيرا" لأن موقع الشواذ للمتغير y كان في طرفي الخط ، أما حالة الشكل (٤) فلم يحصل تأثير ذلك لأن موقع النقطة الشاذة متسق مع خط الأنحدار ولم تتغير قيم المعالم .

وبالتأكيد فأن هناك حالات أخرى تبين شدة تأثير النقاط غير المألوفة او النقاط الشاذة في قياسات المتغيرين المعتمد والمستقل، إلا اننا نستطيع أن نستنتج من ملاحظة الأشكال السابقة ان مواقع هذا النوع من النقاط يلعب دوراً مهماً في مدى تأثيره على النموذج. وفي دراستنا هذه سوف نتطرق إلى الشواذ في المتغير المستقل × بشكل تفصيلي اكثر من شواذ المتغير y ذلك لان الثانية قد حظيت باهتمام اكثر من قبل الباحثين.

٢- نقاط الأنعطاف (Leverage Points):

وهي على نوعين (الجيدة تلك التي تتفق مع خط ألأنحدار ، والرديئة التي لا تتفق معه ويتطلب معالجتها) وقد سميت بهذا الاسم لأنها تعطف (تسحب) خط الانحدار نحوها، اما شدة تأثيرها فيقاس بواسطة عناصر مصفوفة H (ذات الشفقة) Hat matrix التي اذا ضربت بمتجه قيم y فانها تقبعة فنحصل على القيم التقديرية للنموذج، وكما يأتي :-

$$H = X(X'X)^{-1}X'$$
$$\hat{y} = H'y$$

وتسمى المصفوفة H بأنها مصفوفة التنبوء predictor matrix عليه يمكننا كتابة النموذج للتنبوء بدلالة عناصر هذه المصفوفة كالأتي:-

$$\hat{y} = \sum_{i=1}^{n} h_{ij} y_j \text{.....(1)}$$

حيث تمثل n عدد المشاهدات او حجم العينة .

تتصف المصفوفة H بأنها مربعة ذات أبعاد (nxn) متماثلة وصماء Idempotent تحقق أن (HH=H) ، عليه فإنها مصفوفة إسقاط projection matrix قيمها المميزة اما صفر أو واحد وعدد قيمها الصفرية يساوي رتبتها (tr(H)=1) وان عناصرها h_{ij} تمثل وزن المشاهدة y_j إلى القيمة التقديرية وفق المعادلة (1) أعلاه، كذلك فان عناصرها القطرية تتصف بما يأتي:-

$$h_{ii} = \sum_j h_{ij}^{\ 2}$$
$$\frac{1}{n} \le h_{ij} \le l$$
$$h_{ii} = \frac{l}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^{\ 2}}$$

أما القيمة المتوقعة للعنصر h_{ii} فهي:- $Eh_{ii} = \frac{p}{n}$

حيث p تمثل عدد المعالم . ومن الخصائص المهمة الأخرى لهذه المصفوفة أن مصفوفة التباين والتباين المشترك لبواقي النموذج $\sum$ يمكن أن تكتب بدلالة المصفوفة H وكالاتي:

$$\sum^{2} = \sigma^2(I - H)$$

أي أن كلَّ تباين سوف يساوي :

$$\sigma_{ei}^2 = \sigma^2(I - h_{ii}), \text{.....} i = j$$
$$\sigma_{eiej}^2 = \sigma^2(0 - h_{ij}) = -hij\sigma^2, \text{.....} i \neq j$$

ويمكن كتابة تقدير التباين بالصيغة الآتية:

$$S_{ei}^2 = MSE\ (1 - hii)..... ....(2)$$
$$S_{eiej} = -h_{ij}\ (MSE\ )$$

وحيث انه يفترض على الأكثر أن المتغير X يمتلك التوزيع الطبيعي المتعدد، فان هذا يمكن الاعتماد عليه في تحديد توزيع قيم عناصر المصفوفة H ، فلقد اثبت بعض الباحثين أن توزيع عناصر قطر H عندما يمتلك المتغير X التوزيع الطبيعي المتعدد المركب Complex multivariate Gausian ، هو توزيع بيتا [Chave & Thomson, ٢٠٠٣]:-

$h_{ii} \sim$ Beta (p,n-p)

بأشارة أن برهان هذا يكون وفق السياق الذي بدأ به (Muirheadعام 1982) حيث انه يمكن كتابة التحويل الأتي:-

$$f = \frac{h_{ii}}{1-h_{ii}}(\frac{n-p}{p})..... .....(3)$$

فان المتغير f هذا يتبع توزيع F المركزي كالأتي:- $f \sim F(2p,2(n-p)$

وقد وضع (Johnson وآخرون عام 1995) الفرضية الآتية التي تنفع للتحويل من توزيع F إلى توزيع بيتا:-

$$N = n - p, \frac{N}{N+pf} = l - h_{ii}$$

عليه تكون دالة h_{ii} هي دالة بيتا وبالصيغة الاتية [Chave & Thomson, 2003]-

$$p(h_{ii}) = \frac{1}{B(p,n-p)} h_{ii}^{\ p-l}(1-h_{ii})^{n-p-l}, \ldots 0 \le h_{ij} \le 1 \ldots\ldots\ldots\ldots\ldots(4)$$

اما الدالة التوزيعية إلى h_{ii} فسوف نحصل عليها بموجب تكامل بيتا الناقص.

٣ - تشخيص القيم غير المألوفة:-

(3-1) تشخيص نقاط الأنعطاف :

إن احد أهم فوائد المصفوفة H هو أنها قاعدة لتحديد شواذ المتغير التنبؤي X ، حيث تعتبر عناصر القطر الرئيسي h_{ii} مقياس لبعد المسافة بين قيم X_i ومتوسط هذه القيم، او بينها وبين مركز متوسطات المتغيرات التنبؤية X_{ij} في حالة الانحدار المتعدد .

وعندما تزداد قيم العناصر h_{ii} تقل بالمقابل قيمة تباين البواقي e_i حتى إذ أصبحت $h_{ii}=1$ سيكون تباين البواقي $\sigma_{ei}^2 = 0$. ان قيم الانعطاف h_{ii} تعبر عن حالة وجود الشواذ في المتغيرX

فاذا كان $h_{ii} > \alpha \bar{h}$ ، حيث ان $\bar{h} = \frac{p}{n}$ وبأفتراض أن (α=2 , 3) فان قيمة X_i تكون قيمة غير مألوفة (نقطة انعطاف) ومؤثرة في النموذج [Neter.. etc . 1996] .

ومن جانب أخر اذا كانت ($h_{ii} >= 0.5$) فأنها تمثل نقطة انعطاف عالية أو عالية جداً ، أما إذا كانت ضمن الفترة $0.2< h_{ii} <0.5$ فتعبر عن انعطاف معتدل ، إلا انه ينبغي ملاحظة الفجوات في سلسلة البيانات المدروسة.

ان هذه المعايير التي ذكرت في تشخيص القيم الشاذة تعمل بشكل جيد في العينات ذات الاحجام الكبيرة والمتوسطة نسبة إلى عدد المعالم، أما العينات الصغيرة فأنها تكون ضعيفة الاداء وليست عملية، ذلك لان نسبة عدد المعالم إلى حجم العينة $\frac{p}{n}$ سيكون كبيراً.

(3-2) شواذ المتغير y Ouliers :-

لقد درس هذا الموضوع كثيراً ومدى تأثيرات هذه القيم على نموذج الانحدار وتشخص هذه الشواذ عادة اعتماداً على تحليل البواقي،حيث انه بالإمكان الاستفادة من المصفوفة H في تشخيص شواذ y انطلاقاً من حقيقة أن البواقي e_i قد لا تكون متساوية التباين، فيكون مناسباً ان تكون نسب خاصة لكل قيمة من e_i إلى انحرافها المعياري المقدر لتعطي اهتمام لكل مشاهدة بمفردها وانعكاس ذلك على الأخطاء، وبكلام آخر إننا سوف نستخرج بواقي معيارية جديدة بموجب المعادلة الآتية [Neter وآخرون ١٩٩٦] :-

$$r_i = \frac{e_i}{S_{ei}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}} \dots\dots\dots\dots\dots\dots(5)$$

إلا أن الأسلوب الآخر هو إيجاد البواقي المحذوفة التي تساوي الفرق بين القيم التقديرية في حالة وجود كل القيم وبينها عند حذف القيمة الشاذة i وكما يأتي:-

$$d_i = \hat{y}_i - \hat{y}_{i(j)} = \frac{ei}{1-hii} \dots\dots\dots\dots\dots\dots(6)$$

ومن جانب آخر فان هناك نوع اخر هي بواقي t التي تستخرج كما يأتي :-

$$t_i = \frac{d_i}{S(d_i)} = e_i[\frac{n-p-1}{SSE(1-h_{ii}-e_i^2}]^{\frac{1}{2}} \dots\dots\dots\dots\dots\dots(7)-$$

التي تسمى بواقي t المعيارية Studentized residuls [Neter ..et ..al Fox 2005 & 1996] فاذا كانت قيم t_i كبيرة دلّ ذلك على وجود الشواذ في y_i ، ويتم التحليل بواسطة اختبار Bonferoni إلى عدد n من الاختبارات كل حالة وحدها، وتكون النقطة الحرجة

للاختبار t((1-α)⁄2n ,n-p-1) فاذا كانت قيمة t المستخرجة اكبر من القيمة الجدولية تعتبر القيمة المحذوفة y_i شاذة.

٤- تأثيرات القيم غير المألوفة :

(٤-١) التأثير المنفرد: من الضروري التحري عن النقاط غير المألوفة على نموذج الانحدار بعد تشخيصها فقد تكون قليلة التأثير أو بالعكس، ويتم ذلك بالأسلوب الأتي: نحذف هذه القيم ثم نجري التحليل ونحسب الفرق في القيم التقديرية بين الحالتين وفق الصيغة الآتية [Neter ..et al 1996]:-

$$DF = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} ------(8)$$

حيث تعني الإشارة (i) حالة استبعاد القيمة الشاذة i من البيانات وإجراء التحليل بعد الحذف وهذه الصيغة يمكن كتابتها كالأتي:

$$DF = t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}} ----------(9)$$

$$t_i = e_i \sqrt{\frac{n-p-1}{SSE\,(1-h_{ii}-e_i^2)}}$$

وقاعدة القرار هنا ،إذا كانت القيمة المطلقة إلى DF اكبر من (١) في العينات الصغيرة والمتوسطة واكبر من $\frac{2}{\sqrt{p/n}}$ في العينات الكبيرة فان القيمة (i) تعتبر مؤثرة في النموذج .

ويعد معيار مسافات كوك Cooks distance بالصيغة الآتية مقياس جيد لقياس تأثير النقاط غير المألوفة على النموذج المقدر [Neter ..et al 1996] :-

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE}$$

$$D_i = \frac{e_i^2}{pMSE}[\frac{h_{ii}}{(1-h_{ii})^2}] ----(10)$$

وسوف تعتمد قيمة D_i على ضخامة احد قيم الانعطاف أو الأخطاء e_i أو كليهما.وقد وجد أن من المفيد في تفسير D_i أن ننسبها إلى قيمة توزيع F بدرجات حرية F(p,n-p) ثم يتم التحقق من النسبة فإذا كان %(20 – 10) فإنها ذات تأثير قليل على القيم المقدرة للنموذج وإذا كانت قريبة من 50% أو أكثر فان التأثير كبير . إن مسافات كوك هذه يمكن إعادة كتابتها عند تعويض المعادلة (٥) في(١٠) كمايأتي :

$$D_i = \frac{r_i^2}{p}(\frac{h_{ii}}{1-h_{ii}}) ----(11)$$

ليمثل الحد الأول معيار إلى شواذ Y والحد الثاني معيار إلى شواذ X (نقاط الانعطاف). ولقد اقترح chatterjee &Hadi المتباينة $D_i > \frac{4}{n-p}$ كمقياس لمدى التأثير .

أما تأثير النقاط غير المألوفة على معاملات نموذج الانحدار فيمكن قياسه بموجب المعيار الآتي حيث k=0,1,……..p-1), (المصدر ٤) :

$$Dp = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} C_{kk}}} ------(12),..... ....C_{kk} = diag\,(x'x)$$

ويمكن كتابها بصيغة المصفوفات عندما تكون الأخطاء ذات توزيع طبيعي كالأتي :

$$Dp = \frac{(b_k - b_{k(i)})x'x(b - b_{(i)})}{pMSE} - - - - (13)$$

ان Dp تتوزع توزيع F(p,n-p) ، فإذا تجاوزت القيمة المطلقة لها الواحد للعينات الصغيرة والمتوسطة وقيمة $\frac{2}{\sqrt{n}}$ للعينات الكبيرة فإنها تعتبر مؤثرة في المعالم .

٤-٢ التأثير المشترك:-

قد يتوحد تأثير الشواذ باتجاهي X,Y فتكون النتيجة تأثير قوي على نموذج الانحدار وقد يتعاكس اتجاه التأثير فيحصل توازن بين الاتجاهين حيث يلغي واحد تأثير الأخر نسبياً ، ويقاس التأثير بأحد الأساليب الآتية :

١. عند تحليل الانحدار المتعدد من الممكن إجراء التحليل بشكل منفرد لكل حالة، ونستخدم معيار مسافات كوك وبواقي t المعيارية مع المادة بتقدير النموذج لكل حالة.

٢. باستخدام الرسم البياني : فإذا كانت v_i تمثل بواقي انحدار Y على كل متغيرات X عدا X_1 التي نفرضها غير مألوفة وw_i تمثل انحدار X_1 على المتغيرات التوضيحية الأخرى ، فان هنالك علاقة بين نوعي البواقي w_i، v_i وعند رسم v_i مقابل w_i فانه يمكن معرفة التأثير على المعاملات المقدرة (المصدر Fox 2005).

٣. إذا رسمنا بواقي t المعيارية على المحور العمودي بحدي قطع ± 2 مقابل قيم الانعطاف Hat-values بحدي قطع $2\bar{h},3\bar{h}$ النقاط التي تتجاوز هذه الحدود تعتبر مؤثرة على نموذج الانحدار بشكل كبير .

وعند تشخيص النقاط غير المألوفة في النموذج فيتم معالجتها كما يأتي :-

- اعادة تشخيص النموذج واقتراح إضافة تعديلات على المتغيرات ذات النقاط الغير المألوفة المؤثرة، لان هذه المركبات تكون مهمة في تحليل النموذج.
- إجراء تحويلات على البيانات .
- استخدام الأساليب الحصينة التي تعتبر تحليل مقاوم للنقاط الشاذة بشكل عام .

٥- الأساليب الحصينة: Robust methods

أصبحت الأساليب الحصينة معروفة في مقاومتها للشواذ ومن اهم هذه الاساليب طريقة نوع الإمكان الأعظم type of maximam liklehood وطريقة مربعات الوسيط الصغرى LMS واسلوب التشذيب Trimed method وغيرها ..

ان اشهر المقدرات الحصينة هي مقدرات M التي نحصل عليها باستخدام دالة وزن معينة وباسلوب المربعات الصغرى المتغيرة الوزن التكرارية IRLS والتي تعمل على تقليل تاثير الشواذ باعطائها وزن اقل من غيرها ، الا ان مقدارات – M الناتجة من هذه الطريقة تمتلك نقاط انهيار واطئة ، كما انها لاتقاوم الشواذ باتجاه X أي لاتقاوم نقاط الانهيار العالية ، وعلى هذا اقترح بعض الباحثين مقدارات اخرى مثل المقدارات المحددة التاثير (BI) Bounded influence estimation التي تقاوم وجود الشواذ باتجاه X,Y في آنٍ واحد .

ان مقدارات –M السالفة الذكر نحصل عليها من حل المعادلات الاتية :

$$\sum X_{ij}\psi(\frac{e_i}{\hat{\sigma}}) = 0 - - - - - - - - - (14)$$

حيث $\hat{\sigma}$ يمثل تقدير معلمة القياس التي يمكن تقديرها مسبقاً ، ويتم الحل بموجب الطرائق العددية او التكرارية مثل طريقة المربعات الصغرى المتغيرة الوزن التكرارية وبما ان مقدرات M

تمتلك نقاط انهيار واطئة فهي ليست مقاومة للشواذ في اتجاة المتغير X (أي عند وجود نقاط الانعطاف) وخاصة الرديئة منها ، فكان اقتراح تعديل المقدرات بنوع اخر هي مقدرات GM (Generalized M-estimtor) والتي نحصل عليها من خلال حل المعادلات الاتية وفقاً لما اقترحه Mallows :

$$\sum WiXi\psi\frac{e_i}{\hat{\sigma}}----------(15)$$

حيث تمثل Wi دالة اوزان نكون منها المصوفة W القطرية ، وقد اقترح Simpson الصيغة الاتية لغرض استخراج مقدرات GM (المصدر ٦) :

$$W_i=\min\left[1,\frac{\chi_p}{(X_i-\mu_x)'C_x^{-1}(X_i-\mu_x)}\right]$$

حيث ان μ_x المتوسط ، C_x مصفوفة القياس المشترك لقيم X ، χ_p :هو الجذر التربيعي لقيمة مربع كاي عند درجة حرية p ومستوى 95% وسنحصل علىمقدرات GM ذات الخطوة الواحدة بأستخدام الصيغة الاتية :

$$\hat{\beta}=\hat{\beta}_0+A^{-1}\hat{\sigma}_0\sum A\,\psi(\frac{e_i}{\hat{\sigma}})W_iX_i-------(16)$$

$$A^{-1}=n^{-1}\sum\psi'(\frac{e_i}{\hat{\sigma}_0})X'WX$$

حيث ان :

X_i : صفوف مصفوفة الانحدار.

$\hat{\beta}_\circ$: التقدير الاولي والمقترح ان يكون وفق طريقة LMS (مربعات الوسيط الصغرى التي تمتلك نقاط انهيار عالية) ولقد وضعت تعديلات على دالة الهدف (١٥) لغرض الحصول على تحليل اكثركفاءة منها بموجب الصيغة الاتية :

$$\sum Wi\psi(\frac{e_i}{\tau(x)d})X_{ij}=0----------(17)$$

حيث ان : Wi : اوزان تعتمد على الانعطاف ، $\tau(x)$: دالة في صفوف المتغير x ، d: تقدير حصين لتباين البواقي وأن (i=1,2,.....N, j=1,2,.....m).

فاذا وضعنا T(x)=1 ، $Wi=\sqrt{1-h_{ii}}$ حصلنا على صيغة Mallows السابقة ، وإذا كانت $T(x)=w_i$ حصلنا على صيغة Sehweppe ومن قبله تعود الفكرة الى Handschin وجماعته ١٩٧٥ (Chave et al 2003) . ولقد اقترح Chave &Thomson ان نضع T(x)=1 وبالأعتماد على ماذكره باحثين سابقين أن يكون حل المعادلة (١٧) بطريقة المربعات الصغرىالتكرارية المعادة (المتكررة) الوزن ألتي يمكن كتابتها بصيغة أخرى وكما يأتي :

$$\sum w_i^{(k)}v_i^{(k)}e_i^{(k)}x_{ij}=0-------(18)$$

حيث يمثل k عدد مرات التكرار و $e_i^{(0)}$ بواقي المربعات الصغرى ألأعتيادية، و v_i أوزان تناقصية تأخذ الصيغة الآتية:

$$v_i=\frac{\psi(e_i)^{(k-1)}/d^{(k-1)}}{e_i^{(k-1)}/d^{(k-1)}}$$

كما ان d : تقدير حصين لتباين البواقي المرتبة .

$w_i^{(k)}$:أوزان تناقصية تخص نقاط الانعطاف .

ان صيغة w_i التي تخفض من تأثير نقاط الانعطاف الكبيرة بطريقة مرنة أقترحت من قبل الباحث Thomson 1977 وكما يأتي :

$$w_i^{[k]} = w_i^{(k-1)} \exp\{\exp(-\chi^2)\}\exp\{-\exp[x(t_i - \chi)]\ \} ----(19)$$

$$t_i = \frac{m^{(k)} p_{ii}^{(k)}}{p} \qquad ، \qquad w_i^{(0)} = 1 \qquad \text{وان:}$$

حيث ان اكبر قيمة للحد الاسي في المعادلة (١٩) هو (١) ، مما يعني انها لتخفيض الوزن ، أما p_{ii} فتساوي:

$$p_{ii}^{(k)} = u_i^{(k-1)} x_i (X' U^{(k-1)} X)^{-1} x_i u^{(k-1)} ----(20)$$

حيث ان : $U^{(k)} = W^{(k)} V^{(k)}$

$W^{(k)}$مصفوفة اوزان قطرية للانعطاف ،$V^{(k)}$ مصفوفة اوزان قطرية للبواقي ، كما أن $V^{(0)}, W^{(0)}$ مصفوفات الوحدة وأن $M^{(k)} = tar(U^{(k)})$. اما χ معلمه حرة تمثل نقطة بداية تخفيض الوزن للانعطاف ، ويمكن اختيارها بانها تساوي $\frac{p}{m}$. وبشأن V_i فيمكن ان تكون وفق الصيغة الأخرى الاتية (Chave et al 2003) :

$$V_i^{(k)} = \exp[\exp(-\xi^2)]\exp[-\exp\{\zeta(|e_i^{(k-1)}| / d^{(k-1)} - \zeta)\}] ------20$$

وتمثل ξ النقطة التي يبدأ عندها تخفيض الوزن ، كما ان d مقدر القياس يقدر بطريقة MAD .

ان عدم اختيار دالة وزن متناقصة جيداً يجعل المقدرات الناتجة لا تمتلك صيغة أو خاصية الوثبة not bounded وان مقدرات الوثبة لها نقطة انهيار تساوي $\frac{1}{p}$.

من ناحية اخرى فان مقدرات عائلة BI (Bounded Influence) قليلة الكفاءة احياناً وعند وجود قيم انعطاف عالية ، وعليه فقد اقترحت مقدرات هجينة من عائلة LMS , LTS مع مقدرات BI من قبل Hettmansperger&Coakley لتجاوز هذه المشكلة .

بقي ان نشير الى ان Haw and Olin قد قلل من معيار نقطة الانهيار في المقدرات الحصينة والاتجاه صوب معيار الكفاءة في العينات الكبيرة .

٦- الجانب التطبيقي:-

جمع الباحث بيانات عن الدخل الشهري لعينة من منتسبي جامعة تكريت بناء على اعتقاد وجود فروقات كبيرة في دخولاتهم أي وجود للقيم غير المألوفة بسبب التفاوت الكبير بين مستوياتهم الوظيفية أو العلمية أوالعمرية أو غيرها . لقد أفترض الباحث أن المتغير المعتمد y هو الدخل الشهري(الراتب يضاف له أية دخولات متفرقة أخرى) ومتغيرين مستقلين هما العمر x_1 ومدة الخدمة بالسنوات x_2 ووفقا لبيانات العينة المؤلفة من ٩٣ مفردة كانت معادلة الأنحدار الخطي باستخدام النظام الأحصائي spss كما يأتي:-

$$Y=9.346X_1+27.372X_2+65.311$$

ولغرض ايجاد المصفوفة H بموجب العلاقة المعرفة في الفقرة (٢) وبأستخدام البرنامج MatLab حصلنا على المصفوفة ذات أبعاد ٩٣x 93 وكانت عناصر القطر الرئيسي h_{ii} كما يأتي (من اليمين الى اليسار):-

٠.٠١٠٢ ٠.٠٢٩٤ ٠.٠١١٥ ٠.٠٢٧٦ ٠.٠٠٩٢ ٠.٠١٢٠ ٠.٠٤٨٨ ٠.٠١٠٧ ٠.٠١٤٠ ٠.٠٢٤١
٠.٠١٧٦ ٠.٠٢٥٤ ٠.٠٠٦٤ ٠.٠٢١٩
٠.٠١٨٣ ٠.٠٤١٠ ٠.٠٣١٣ ٠.٠٢٣١ ٠.٠٢٣١ ٠.٠١٩٢ ٠.٠٥٦٠ ٠.٠٠٧٨ ٠.٠٢٦١ ٠.٠١٦٨
٠.٠١٢٦ ٠.٠٢٩١ ٠.٠١٩٤ ٠.٠٢٦٥
٠.٠٠٩٧ ٠.٠٧٤٠ ٠.٠١٥٦ ٠.٠١٩٢ ٠.٠١٨٠ ٠.٠١٧٦ ٠.٠١١٠ ٠.٠١٤٢ ٠.٠١٢١ ٠.٠١٦٧
٠.٠١٤٥ ٠.٠٣٠٨ ٠.٠١٥٣ ٠.٠٢٦١
٠.٠١٥٨ ٠.٠٢٣٧ ٠.٠٨٤٨ ٠.٠٤٤١ ٠.٠٣٤١ ٠.٠١١٦ ٠.٠١٠٦ ٠.٠٠٧٤ ٠.٠٢٣٦ ٠.٠٤٦٤
٠.٠١٤٣ ٠.٠١٧٦ ٠.٠١٢١ ٠.٠١٩٢

٠.٠١٦٥ ٠.٠١٢٧ ٠.٠١٣٢ ٠.٠١٣٩ ٠.٠١٢١ ٠.٠١٢٩ ٠.٠١١٧ ٠.٠١٦٧ ٠.٠٨٤٨ ٠.٠١٠٠
٠.٠٢١٦ ٠.٠٠٧٩ ٠.٠١٥٤ ٠.٠١٨٤
٠.٠٢٢٣ ٠.٠١٥٩ ٠.٠٤٩٣ ٠.٠٠٧٧ ٠.٠١٣٨ ٠.٠١٤٢ ٠.٠٣٢٣ ٠.٠٢٩٤ ٠.٠١١٨ ٠.٠٢٤١
٠.٠١٢٣ ٠.٠١٤٦ ٠.٠٦١٨ ٠.٠١٠٣
٠.٠١٣٢ ٠.٠١٢١ ٠.٠١٥٩ ٠.٠١٣٥ ٠.٠١٥٤ ٠.٠٢٣٠ ٠.٠١٨٢ ٠.٠١٢٩ ٠.٠٠٨٨

وعنـــد اجـــراء المقارنـــة مـــع معيـــار المقارنـــة مضـــاعفات الوســـط الحســـابي :-

$$\bar{h} = 0.03226, 2\bar{h} = 0.06452, 3\bar{h} = 0.09678, ... \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{93}} = 0.2074 \quad)$$ -

تبين أن المشاهدات ذات التسلسلات (٣٧،٥٠،٦٥) تعتبر نقاط انعطاف مؤثرة (التي تحتها خط)، ولكن هل هذه النقاط احدثت تأثيرا قويا في معاملات خط الأنحدار ، فعند حذف كل واحدة من هذه القياسات ثم اجراء تقدير لخط الأنحدار وحساب التأثيرباستخدام المعادلة (١٢) والمقارنة مع القيمة (٠.٢٠٧٤) يتبين أن المشاهدة (٣٧) مؤثرة في معاملات المتغيرين x_1, x_2 وكانت نتائج الحسابات كما يأتي :-

رقم المشاهدة	معامل العمر x_1	معامل مدة الخدمة x_2
٣٧	١.٣٨١١ مؤثر	٠.٦٩٠٥ مؤثر
٥٠	٠.٠٦٩٣ غير مؤثر	٠.١٩٦٨ غير مؤثر
٦٥	٠.٠٨٤٣ غير مؤثر	٠.١٩٠٧ غير مؤثر

وبفحص تأثير المشاهدات المرقمة(٣،٧،١٥) التي يعتقد أنها مؤثرة بأتجاه المتغير y تبين أن المشاهدة رقم ٧ فقط هي المؤثرة في معاملات المتغيرين وقيمة التأثير تساوي (٠.٩٥٥) بأتجاه x_1 و(٠.٣٨٨) باتجاه x_2 اي ان تأثير النقاط الأخرى قد سوي أو تعادلت فيما بينها ولم يعد له تأثير في معادلة الأنحدار ، ويمكن ان نستنتج ما يأتي :-

١- أن معاملات الأنحدار تتأثر بالقياسات غير المألوفة لكافة المتغيرات(الأستجابة والمستقلة) وهذه التأثيرات يختلف تأثيرها حسب موقعها من خط الأنحدار بين مؤثرة وغير مؤثرة .

٢- أن بعض قياسات المشاهدات غير المألوفة أو الشاذة تكون غير مؤثرة سلبيا على معاملات النموذج حيث يلغي تأثير بعض المشاهدات البعض ألآخر .

المصادر :

1. Chave , Alan and Thomson , Daved J. (2003) " Abounded influence regression estimator based on the statistics of the hat matrix" Appl.Statist., 52, part3 , pp(307-322) ……..
2. Fox , Jhon , 2005 ,Unsual and influential data , Lecture Notes, by Internet .
3. Martin , David W. , 2001 ″ Teaching Leverage , outlires and influential observations in introductory ststistic coures , Davidsen college ,by internet .

4. Neter , John Kutner , Michael Nachtsheim , Christopher and Wasserman , wiliam , 1996 ″ Applied Linear statistical Models ″ , WCB McGrawHill , Fourth Edition .
5. Rousseew P.J. and Leroy . AM (1987) ″Robust regression and outlier detection ″ New York , John Wiley and Sons .
6. Simpson , D.G. Ruppert , D. and carroll , R.J (1992) ″ On one – step GM estimates and stability of inference in linear regression ″ JASA , vol. 87, No. 418 , p.439 .

This document was created with Win2PDF available at http://www.win2pdf.com.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.
This page will not be added after purchasing Win2PDF.