المجلة العراقية للعلوم الاحصائية (15) 2009
ص ص [161–192]

المشاهدات غير العادية في الانحدار الخطي المتعدد وبعض طرائق تشخيصها مع التطبيق

مزاحم محمد يحيى*

الملخص

تم في هذا البحث تشخيص ثلاثة أنواع من المشاهدات غير العادية (Unusual Observation) في الانحدار المتعدد (Multiple Regression)،إذ تم الكشف عن المشاهدات الشاردة (Outlier) من خلال استخدام الرسم الصندوقي (Boxplot) ومقياس البواقي القياسية (studentized residuals)،والكشف عن المشاهدات ذات قوة الرفع (leverage observation) من خلال مصفوفة هات (Hat matrix)،فضلا عن الكشف عن المشاهدات المؤثرة (Influence) باستخدام مقياس (dfbeta).

في الجانب التطبيقي تمت مقارنة اثر حذف المشاهدات الشاردة في التوزيع الطبيعي للبواقي في معادلة الانحدار التي تم توفيقها لمرض ثلاسيميا الأطفال،مع معالجة مشكلة تعدد العلاقات الخطية (Multicollinearity) من خلال حذف بعض المتغيرات فضلا عن استخدام انحدار الحرف (Ridge Regression). لقد تم التوصل إلى نموذج يتفق مع وجهة النظر الطبية وذلك من خلال استخدام البرنامج الإحصائي الجاهز (SAS.9).

Types of Unusual Observations in Multiple Regression & Some Methods of it's Diagnostic with Application

ABSTRACT

The work in this paper is a diagnosis of three types of unusual observations in multiple regression, the outlier observation is diagnostic by using Boxplot & studentized

---

* مدرس مساعد/قسم الاحصاء والمعلوماتية/كلية علوم الحاسبات والرياضيات/جامعة الموصل
تاريخ التسلم : 24 /3/ 2008 ـــــــــ تاريخ القبول : 6 /7 /2008

residuals, Diagnostic leverage points by using (Hat matrix) , & diagnostic of the influence observations by using (dfbeta).

In practice the work is comparing the effects of omitting outlier observations in the normal distribution of the residuals to the equation which is building to the Thalassaemia disease and Treating the multicollinearity by omitting some variables and using ridge regression, getting a good model agrees with the viewing of medicine by using (SAS.9) package.

المقدمة:-

إن موضوع المشاهدات غير العادية وتأثير وجود هذه المشاهدات وانعكاساته على دقة التحليلات الإحصائية كان ومازال مثار اهتمام العديد من الباحثين،حيث يعود الاهتمام بالمشاهدات غير العادية في مجموعة البيانات الإحصائية إلى ما يزيد على القرنين،إذ تشير التعليقات المبكرة التي أبداها العالم (Bernoulli) في عام (1774) عن القيم الشاردة وكيفية معالجتها،إلى أن طريقة تشخيص المشاهدات المتطرفة كانت شائعة في ذلك الوقت،كما أن تاريخ معالجة هذه المشاهدات قد أعقبت هذه التعليقات والى يومنا الحاضر.

إن وجود مشاهدة واحدة أو مجموعة صغيرة من المشاهدات تختلف بشكل جوهري عن بقية البيانات يمكن أن يؤدي إلى اختلافات كبيرة في نتائج تحليل الانحدار،وعليه فانه يتطلب الأمر التحقق من هذه المشاهدات والبحث عن طرائق للسيطرة على تأثير هذه المشاهدات.

يمكن تسمية المشاهدات التي تؤدي إلى حدوث تغيرات جوهرية في النتائج بالمشاهدات غير العادية ويمكن تعريفها على أنها نقاط لها إمكانية التأثير في التحليل الإحصائي(Evans, 1999).إن تشخيص المشاهدات غير العادية يعتبر من الأمور المهمة للباحثين الذين يستخدمون تحليل الانحدار بسبب أن هذه المشاهدات يمكن أن تؤثر في النموذج المستخدم وذلك لأنها تعمل على تشويه النتائج الممثلة لبيانات العينة.

تقسم المشاهدات غير العادية إلى أنواع ثلاثة،وهي المشاهدات الشاردة (Outlier) حيث تتصف بان لها بواقي عالية (Bacon,1995)،ومشاهدات ذات قوة الرفع (Leverage) وهي القيم المتطرفة في المتغيرات التوضيحية(Hoagolin & Welsch,1978)،والمشاهدات المؤثرة (Influence) وهي المشاهدات التي إذا ما تم حذفها فان ذلك سيؤدي إلى إحداث تغيرات جوهرية في معلمات النموذج المقدر(Chatterjee,et al.,1986).

هدف البحث:-

يهدف البحث إلى تناول طرائق تشخيص الأنواع الثلاثة من المشاهدات غير العادية،وبيان مدى تأثير المشاهدات الشاردة في التوزيع الطبيعي للبواقي،كما يتناول البحث أيضا اثر حذف المشاهدات الشاردة في الوصول إلى نتائج أكثر دقة وكفاءة للمعادلة التي تم توفيقها لمرض ثلاسيميا الأطفال.

1-الجانب النظري:-

تعد المشاهدة (أو المشاهدات) غير العادية (Unusual Observation) مشكلة ذات تأثير بليغ وخاصة عندما تكون المشاهدة غير عادية في المتغيرات التوضيحية (Explanatory Variables)،فعلى سبيل المثال يميل خط الانحدار عادة إلى تعقب المشاهدة غير العادية، ومن ثم فإنها يمكن أن تحرف خط الانحدار المقدر عن النمط الحقيقي للبيانات.

إن المشاهدة غير العادية يمكن أن يكون لها تأثير قوي في الميل المقدر والحد الثابت،مما يتسبب في ضعف التوفيق للنموذج المقدر للجزء الأعظم من البيانات، ومن ثم إلى زيادة في التباين المقدر.يعود السبب في وجود المشاهدات غير العادية إلى الأخطاء في تسجيل المشاهدات،أو لكون القيم متأتية من مجتمع يختلف في التوزيع عن مجتمع البحث.أو ربما أن قيم Y مأخوذة من مجتمع لا يتوزع التوزيع الطبيعي(Beckman & Cook,1983).

إن عملية الكشف عن المشاهدة غير العادية في الانحدار البسيط (Simple Regression) يعتبر أمرا في غاية السهولة(Sinha,1997)،إذ يمكن الكشف عن المشاهدة غير العادية من خلال الشكل ألانتشاري (Scatter plot) للمتغير X مقابل

قيم Y،أو من خلال رسم البواقي (Residuals).أما في حالة وجود أكثر من متغير توضيحي واحد فان عملية الكشف عن المشاهدات غير العادية تصبح أمرا أكثر تعقيدا،إذ في بعض الأحيان تظهر المشاهدات غير العادية بشكل واضح من خلال الرسوم الفردية للشكل ألانتشاري للمتغيرات التوضيحية مقابل المتغير المعتمد (Dependent Variable)، ولكن في أحيان كثيرة لا يمكن ذلك بسبب أن المشاهدات غير العادية ضمن المتغير الواحد عادية ولكنها عندما تقترن بمشاهدة (أو مشاهدات) أخرى لمتغير آخر(أو لمتغيرات أخرى) فتصبح مشاهدة غير عادية.

إن الطريقة في البحث عن المشاهدة غير العادية،هو في قياس أي من المشاهدات مسؤولة عن التأثير في معلمات الانحدار،إذ في بعض الأحيان يمكن تشخيص القيم غير العادية من خلال الشكل ألانتشاري للبيانات X-Y ،إذ تبتعد المشاهدات غير العادية عن الاتجاه الخطي العام للبيانات.وعليه فانه عند تقدير خط الانحدار،فان الرسم الصندوقي (Boxplot) ورسم الشكل الاحتمالي للتوزيع الطبيعي (normal probability plot) للبواقي يمكن أن يكشف عن المشاهدات غير العادية في البيانات،كذلك يمكن تشخيص المشاهدات غير العادية من خلال فحص البواقي أو المشاهدات ذات قوة الرفع العالية (high-leverag observation).

إن طريقة المربعات الصغرى الاعتيادية (Ordinary Least Square)(OLS)تستلزم تقليل مجموع المربعات الأخطاء،وبسبب هذا فان خط الانحدار المقدر يمكن أن يكون حساسا جدا إلى المشاهدات غير العادية،بمعنى آخر أن طريقة المربعات الصغرى ليست قادرة على التصدي للبيانات غير العادية.إن وجود مشاهدة تختلف بشكل جوهري عن بقية المشاهدات يمكن أن يؤدي إلى تحيز في النتائج (bias the result) وخصوصا عندما يكون حجم العينة صغيرا نسبيا فان المشاهدة غير العادية سيكون لها تأثير اكبر مما يؤدي إلى تحيز خطير في المعلمات المقدرة،لذا يتطلب الكشف نوعية هذه المشاهدة والتحقق منها.

يمكن تمييز ثلاثة أنواع من المشاهدات غير العادية وهي:-

a- المشاهدة الشاردة (Outlier):-

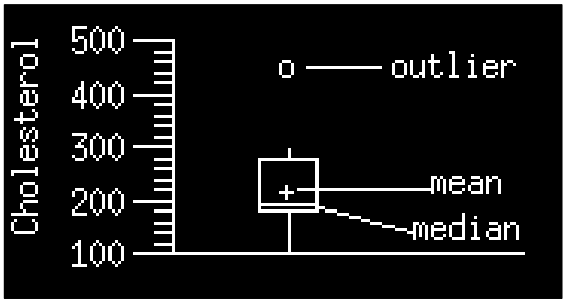
إن المشاهدة الشاردة هي المشاهدة التي لها باقي (Residual)عالٍ،بمعنى آخر أنها المشاهدة التي قيمة متغيرها المعتمد غير عادية. يعود السبب في وجود المشاهدات الشاردة إلى خطا في القياس،وربما إلى الاختلافات العشوائية في البيانات(Barnett & Lewis,1994).يمكن الكشف عن المشاهدات الشاردة من خلال الرسم الصندوقي ومقياس البواقي المعيارية (studentized residuals).

a-1 الرسم الصندوقي (Boxplot) :-

يطلق عليه في بعض الأحيان الرسم الصندوقي والقطع الشعرية (Box and Whisker Plot). وهو مخطط بياني يوضح فيما إذا كان هناك قيم شاردة فضلا عن نمطية (Pattern) البيانات.

الشكل (1) يبين الرسم الصندوقي،ويظهر فيه الخط الوسطي في مركز الصندوق الذي يعبر عن الوسيط،وان المركز (+) يدل على الوسط،،وان القيم الشاردة تظهر كنقاط فردية خارج الصندوق على شكل (o).

الشكل (1): الرسم الصندوقي



يمكن استخلاص نمط البيانات من الرسم الصندوقي،فإذا كان هناك عدد قليل من المشاهدات الشاردة،وان الرسم الصندوقي له وسط (+) قريب من الوسيط،،وان خط الوسيط يقسم الصندوق بالتساوي،فان العينة ربما تحتوي على مشاهدات شاردة.أما إذا كانت هناك مشاهدات شاردة في إحدى الجهتين،وان خط الوسيط لا يقسم الصندوق بالتساوي،عليه فان توزيع المجتمع المأخوذ منه العينة مائل (skewed).أما إذا كانت المشاهدات الشاردة في كلتا الجهتين من الصندوق،فان

توزيع المجتمع المأخوذة منه العينة قد يكون له ذيل ثقيل(heavy-tailed) (Michael,et al(1989)).

2-a البواقي المعيارية (studentized residuals):-

هو نوع من أنواع الـ (Standardized Residuals)* وتستخدم في الكشف عن المشاهدات الشاردة،وهي حاصل قسمة البواقي على الأخطاء المعيارية باستبعاد المشاهدة (i).ويأخذ الصيغة الآتية:-

$$e_i^s = \frac{e_i}{\sqrt{\hat{s}_{(i)}^2(1-h_i)}} \quad \ldots(1)$$

حيث أن:-

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{v^2} \quad \ldots(2)$$

(*) وهي البواقي مقسومة على S_e ، لها وسط حسابي مقداره الصفر وتباين مقداره 1،وتوضح مدى ابتعاد قيمة المشاهدة عن خط انحدار المربعات الصغرى نسبة إلى تشتت المشاهدات حول خط الانحدار.

$$v^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \ldots(3)$$

وأن:-

e_i^s:-البواقي المعيارية.

e_i:-البواقي $(y_i - \hat{y}_i)$.

$\hat{s}_{(i)}$:-الانحراف المعياري باستبعاد المشاهدة (i).

n :-حجم العينة.

x_i:-المشاهدة (i).

$\bar{x}$:-الوسط الحسابي.

يمكن الكشف عن المشاهدة الشاردة عندما $(e_i^s > \pm 2)$ Hair et al.,(1998)) (Iglewicz, B. y Hoaglin, D.C. (1993);.

b- المشاهدة ذات القوة الرافعة(Leverage observation):-

يطلق على المشاهدات التي تعمل كعتلة أو رافعة (Lever) وتغير من الميل لنموذج خط الانحدار بالمشاهدات ذات القوة الرافعة (Leverage).هذه المشاهدة تقع بعيدة عن بقية البيانات وبالتحديد في نهاية احد طرفي خط الانحدار وتؤدي إلى ميلانه باتجاهها.من مقاييس الكشف عن المشاهدات الرافعة مصفوفة هات (Hat Matrix).

1-b مصفوفة هات Hat Matrix :-

يمكن الاستفادة من مصفوفة هات في تحديد المشاهدات الرافعة،إذ يتم الكشف عن المشاهدات الرافعة من خلال العناصر القطرية لـمصفوفة هات (Hat Matrix)، وكالاتي:-

$$\text{Hat Matrix} = X(X'X)^{-1}X' \quad \ldots(4)$$

أي أن

$$H = X(X'X)^{-1}X' \quad \ldots(5)$$

وعليه فان(h_{ii})(العناصر القطرية لمصفوفة هات) تسمى قوة الرفع للمشاهدة i . تعتبر المشاهدة لها قوة رافعة إذا كانت قوة الرفع لها اكبر من ((2k+2)/n)،حيث إن k عدد المتغيرات التوضيحية وان n عدد المشاهدات ((Iglewicz, B. y Hoaglin, D.C. (1993)).

c- المشاهدات المؤثرة (Influence):-

وهي المشاهدة التي تقع في الجهة المقابلة لوسط الخط الممثل لنموذج الانحدار،إذ تكون بعيدة عن الاحداثي الصادي (Y) وكذلك بعيدة عن بقية المشاهدات،وتعمل هذه المشاهدة على تحريك خط الانحدار نحو الأعلى أو نحو الأدنى (بمعنى آخر،باتجاه المشاهدة المؤثرة)،ويمكن أن يكون لها تأثير معنوي في نموذج الانحدار وفي القدرة على التنبؤ.هذه المشاهدة إذا ما تم إدراجها أو حذفها فأنها ستؤدي إلى إحداث تغييرات جوهرية في المعلمات المقدرة (Hair,et al.,1998;Tsay et al.,2000;Hordo,2004).من أهم المقاييس في الكشف عن المشاهدات المؤثرة مقياس (DFBETA).

1-c مقياس (Dfbetas):-

إن Dfbetas هو مقياس للتغيرات في معاملات الانحدار عند حذف مشاهدة من تحليل الانحدار

(Hair et al.,1998).بمعنى آخر انه الفرق بين ميل خطي الانحدار عند وجود المشاهدة (i) وعند استبعادها،

ويأخذ الصيغة الآتية:- $$\mathbf{DFBETAS_{j,i}=\frac{\hat{B}_j-\hat{B}_{j(i)}}{s^2_{(i)}\sqrt{(X'X)^{-1}_{JJ}}}} \quad \ldots(6)$$

حيث أن:-

$\hat{\mathbf{B}}_{\mathbf{J}}$:-معاملات الانحدار باستبعاد المشاهدة (i).

$\hat{\mathbf{B}}_{\mathbf{j(i)}}$:-معاملات الانحدار بوجود المشاهدة (i).

$\mathbf{s^2_{(i)}}$:-التباين باستبعاد المشاهدة (i).

$\mathbf{(X'X)^{-1}_{jj}}$:-العناصر القطرية للمصفوفة $\mathbf{(X'X)^{-1}}$.

يمكن الكشف عن المشاهدة المؤثرة عندما تكون $\mathbf{|Dfbetas_{J,I}|>\frac{2}{\sqrt{n}}}$.

2-الجانب العملي:-

تناول الجانب العملي من هذا البحث العمر من العظم (Y) كمتغير معتمد للمصابين بمرض الثلاسيميا وعلاقته مع المتغيرات التوضيحية أدناه.علما أن توصيف المعادلة والبيانات مأخوذة من المصدر (النعيمي،2005).

$Y=a_0+a_1X_1+a_2X_2+a_3X_3+a_4X_4+a_5X_5+a_6X_6+a_7X_7+a_8X_8+a_9X_9+a_{10}X_{10}+e$ …(7)

حيث أن:- X_1:-العمر الحقيقي مقاسا (بالشهر).

X_2:-عمر المريض عند المرض مقاسا (بالشهر).

X_3:-تضخم الكبد مقاسا (بالسنتمتر). X_4:-هيموكلوبين الدم.

X_5:-مكداس الدم(خلايا الدم المضغوط). X_6:-الخلايا الشبكية.

X_7:-أرومة حمراء. X_8:-الهيموكلوبين الجيني.

X_9:-عدد وحدات الدم. X_{10}:-بداية نقل الدم حسب العمر مقاسا (بالشهر).

ويفترض أن تكون جميع الإشارات الجبرية لمعاملات الانحدار المقدرة للمعادلة (7) موجبة(النعيمي،2005)،بمعنى أن العلاقة طردية بين المتغيرات التوضيحية والمتغير المعتمد.

لقد استخدم الباحث (النعيمي،2005) طريقة انحدار الحرف في معالجة مشكلة تعدد العلاقات الخطية،وتوصل إلى العلاقة الآتية:-

$$Y=-0.00137+0.789X_1-0.126X_4+0.0162X_7 \quad \ldots(8)$$

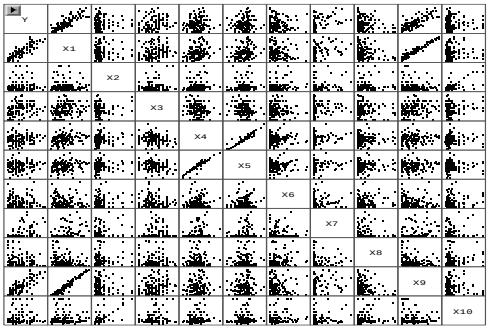
إذ يلاحظ بأن إشارة المتغير (X4) لا تتفق مع وجهة النظر الطبية ،كونها مخالفة للمنطق،وقد وجد أنها غير معنوية.

إن دقة تقدير معلمات النموذج الخطي تتوقف على مدى صحة فرضيات تطبيق طريقة المربعات الصغرى الاعتيادية، فإذا كان هناك من انتهاك لفرضية ما،أو بعض الفرضيات،فان استخدام طريقة (OLS) سيصبح أمرا غير منطقي، مما يؤدي إلى نتائج غير دقيقة،عليه لابد من إجراء الاختبارات للفرضيات المهمة في تطبيق طريقة (OLS) للوصول إلى نتائج قائمة على أسس موضوعية.

إن هذا البحث يمثل محاولة للوصول إلى نتائج منسجمة مع المنطق معززة بالنظرية الإحصائية،وذلك من خلال معالجة البيانات بصورة مختلفة،لذا فقد تم ابتداء رسم البيانات وذلك من اجل التعرف على طبيعة هذه البيانات وتكوين فكرة عما إذا كان هناك من مشاكل محتملة.من الشكل (2) الذي يمثل مصفوفة الشكل ألانتشاري(scatter plot matrix) للبيانات،يمكن ملاحظة أن في كل شكل من الأشكال الانتشارية هناك قيمة واحدة،وفي بعض الرسوم مجموعة من القيم التي تكون بعيدة عن أمكنة تجمع المشاهدات للمتغيرات التوضيحية،ولغرض التعرف على هذه القيم،فقد تم إعطاء رمز لكل مشاهدة بدلا من النقاط (إعطاء المشاهدات (1-150) الرموز (a1-a150)) وتمت إعادة رسم الأشكال الانتشارية للمتغير المعتمد مع كل متغير من المتغيرات التوضيحية،وكما موضح ذلك في الرسوم في الملحق(1) التي تمثل المتغيرات التوضيحية مقابل المتغير المعتمد.يمكن القول أن في كل شكل من الأشكال التوضيحية هناك مشاهدة أو مجموعة من المشاهدات غير العادية والتي يتطلب تشخيصها.

شكل رقم (2)

مصفوفة الشكل ألانتشاري لمتغيرات مرض ثلاسيميا الأطفال



إن وجود مـشاهدة غيـر عاديـة واحـدة (أو مجموعـة صـغيرة مـن المشاهدات)تختلف بشكل جوهري عن المشاهدات الأخرى يمكن أن يكون لها تأثير في تشويه النتائج بدرجة كبيرة،خاصة عندما تستند الطريقة الإحصائية التقليدية في الوصول إلى نتائج مثالية تحت فرضية أن المشاهدات تتوزع توزيعا طبيعيا. يتضح من أعلاه أن البيانات تتطلب اختبارات مهمة لغرض معرفة مدى تطـابق شـروط تطبيق طريقة المربعات الصغرى الاعتيادية.إن الفرضية المطلوب اختبارها هي:-

فرضية العدم: البواقي تتوزع توزيعا طبيعيا.

الفرضية البديلة:البواقي لا تتوزع توزيعا طبيعيا.

لغرض اختبار الفرضية المذكور ة انفا فقد تم إيجاد البواقي واختبارها مـن خـلال المقياس (Shapiro-Wilk)(W).

من الجدول (1) يتضح المعنوية العالية لاختبار (W) عند مستوى المعنوية (5%)،عليه سيتم رفض فرضية العدم وقبول الفرضية البديلـة،أي أن البـواقي لا تتوزع توزيعا طبيعيا،ويمكن ملاحظة ذلك من خلال الشكل (3) الذي يمثـل(رسـم كثافة النواة(اللب)) (Kernal density plot)، حيث يظهر فيه توزيع البواقي قريبا من التوزيع الطبيعي.لذا يتطلب الأمر البحث عن القيم الشاردة والتي غالبا ما تكون السبب في انحراف البواقي عن التوزيع الطبيعي.

الجدول (1)(*)

المعلمات المقدرة وتحليل التباين واختبار التوزيع الطبيعي لبيانات مرض ثلاسيميا الأطفال

Variable	DF	Parameter Estimate	Standard Erorr	t Value	Pr > |t|	Variance Inflation Factor
Intercept	1	5.13662	6.0365	0.85	0.3963	0
X1	1	2.40727	0.21524	11.18	<.0001	41.35083
X2	1	-0.3269	0.24222	-1.35	0.1793	2.72755
X3	1	0.64232	0.59046	1.09	0.2786	1.23351
X4	1	-13.42824	2.56653	-5.23	<.0001	14.96864
X5	1	1.76712	0.70535	2.51	0.0134	11.70736
X6	1	0.16335	0.37435	0.44	0.6633	1.15793
X7	1	0.02645	0.01253	2.11	0.0366	1.33558
X8	1	-0.04441	0.06963	-0.64	0.5247	1.2398
X9	1	-1.28947	0.17118	-7.53	<.0001	37.85468
X10	1	-0.07048	0.2232	-0.32	0.7526	2.98425

Source	DF	Sum of Squares	Analysis of Variance	F Value	Pr > F
Model	10	184752	18475	71.97	<.0001
Error	139	35683	256.71456		
Corrected Total	149	220436			
Root MSE		16.02231	R-square	0.8381	
Dependent Mean		63.72667	Adj R-sq	0.8265	

Tests for Normality		
Test	Statistic---	-----p Value------
Shapiro-Wilk W	0.981634	Pr < W 0.0426

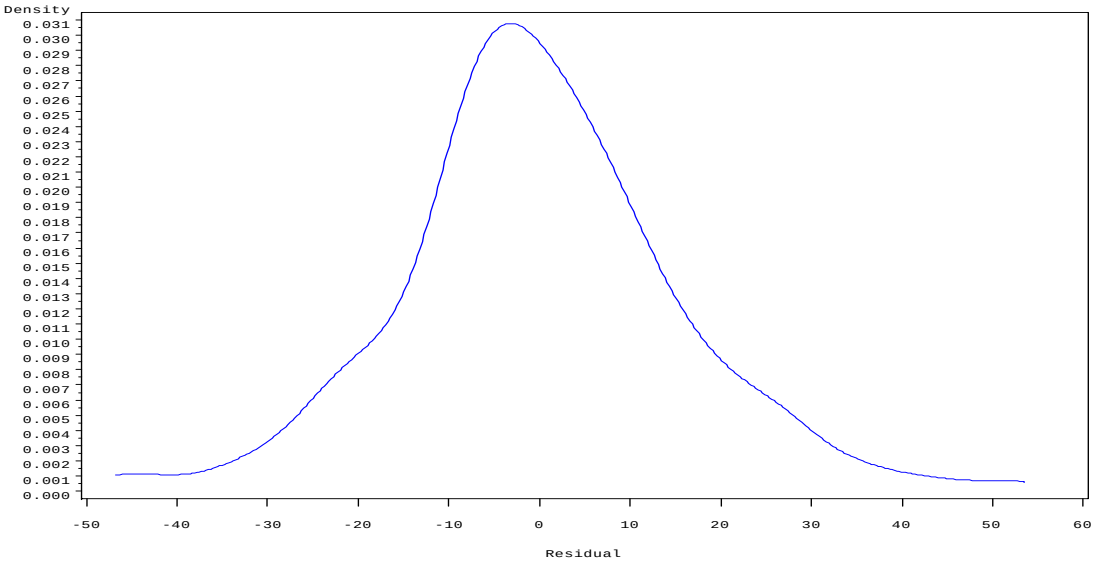
(*) المعلمات المقدرة وعامل تضخم التباين فضلا عن تحليل التباين في الجدول (1) هو استنتاجات الباحث (النعيمي،2005).

الشكل (3)(*)

رسم كثافة النواة (اللاب)

(Kernal density plot)

لبواقي معادلة الانحدار



2-1 تشخيص المشاهدات الشاردة:-

ابتـداء ولغـرض تـشخيص المـشاهدات الـشاردة يمكـن الاسـتعانة بالمقيـاس(studentized residuals).فمـن خـلال اختبـار البـواقي بواسـطة الشكل(stem & leaf)(الشكل 4) يمكن التعرف على هذه المشاهدات التي تتجاوز قيمهـا (± 2) والتـي تعـود إلـى المـشاهدات ((a96)،(a142)،(a127)، (a109)،(a104)،(a75)،(a99)،(a3)) وكما موضح ذلك في الجدول(2)، وتـشير هذه النتيجة إلى أن في مقدمة المشاهدات التي تمتلك اكبر البواقي هـي المـشاهدة (a96).،كما يمكن ملاحظة المشاهدات الشاردة من خلال الرسم الصندوقي والممثلة بالرمز (o) والتي تقع خارج الصندوق.

يلاحظ من الرسم الصندوقي (الشكل 4)أن المشاهدات الشاردة تقع في كلتا الجهتين من الصندوق،وان الوسط (+) يقع فوق خط الوسيط،،وان خط الوسـيط لا يقـسم الصندوق بالتساوي،مما يشير إلى أن توزيع المجتمع المأخوذة منه العينة قد يكـون له ذيل ثقيل(heavy-tailed).

الشكل (4):شكل يوضح قيم (steam & leaf) والرسم الصندوقي

```
                  Stem Leaf
#             Boxplot
                      3 5
1                  0
                      3
                      2 9
1                  0
                      2 34
2                  0
                      1 556777889
9                  |
                      1 1112234
7                  |
                      0 5555566666777778999999
22              +-----+
                      0
00001111112222333333444444444       28              |
+  |
                     -0
444433333333222222111111111000    31
*-----*
                     -0 9877766666666666555555555555
27              +-----+
                     -1 443322110000
12                 |
                     -1 987655
6                  |
                     -2 22
2                  0
                     -2
                     -3 10
2                  0
                    ----+----+----+----+----+----+-
```

(*) تم استخدام البرنامج الإحصائي الجاهز (SAS.9) في رسم كثافة النواة (اللب).

الجدول (2): المشاهدات التي تمتلك اكبر البواقي

المشاهدات الشاردة			
Highest		Lowest	
Obs	Value	Obs	Value
40	1.90524	142	-3.1265
75	2.30187	127	-3.0108
104	2.42619	99	-2.2285
109	2.9294	3	-2.1901
96	3.53914	132	-1.8775

2-2 تشخيص المشاهدات ذات القوة الرافعة (Leverage):-

إن المقياس الثاني الذي تم استخدامه هو مـصفوفة هات،وذلـك لغـرض تشخيص المشاهدات التي يمكن أن يكون لها تأثير محتمل في ميل خط الانحـدار المقدر.تعتبر المشاهدة ذا تأثير في ميل الانحدار المقدر عندما تكون القوة الرافعـة لها اكبر من ((2k+2)/n)،إذ أن (n) عدد المشاهدات،وان (k) عـدد المتغيـرات التوضيحية.وعليه فان المشاهدة التي لها قوة رافعة اكبر (0.1467) تعتبر ذا تـأثير في ميل الانحدار المقدر.

الجدول (3) يبين المشاهدات التي لها قوة رافعة اكبـر مـن (0.1467)،إذ يمكـن ملاحظة أن المشاهدات التي لها قوة رافعة اكبر (0.1467) تختلف عن المشاهدات التي لها بواقي عالية،ويمكن توضيح المشاهدات ذات القوة الرافعة والـشاردة مـن خلال رسم القيم التربيعية للبواقي مقابل القيم الرافعة،إذ يلاحظ من الـشكل (5) أن المشاهدات التي تمتلك اكبر البواقي واقعة في أسفل الزاوية اليمنى،وان المـشاهدات التي لها قوة رفع عالية تقع في أعلى الزاوية اليسرى،أما النقاط الواقعة في الجهـة

اليمنى العليا،فهي المشاهدات التي لها قوة رافعة وبواقي عالية على حد سواء.أمـا بقية المشاهدات فهي متجمعة في الربع الأول والمناطق القريبة نسبيا.

من الشكل (5) يظهر بوضوح المشاهدات (a96,a142,a127,a109) كمـشاهدات تمتلك بواقي كبيرة،مقابل المشاهدات (a71,a73,a137,a54) التي لها قوة رافعـة كبيرة على ميل الانحدار المقدر.

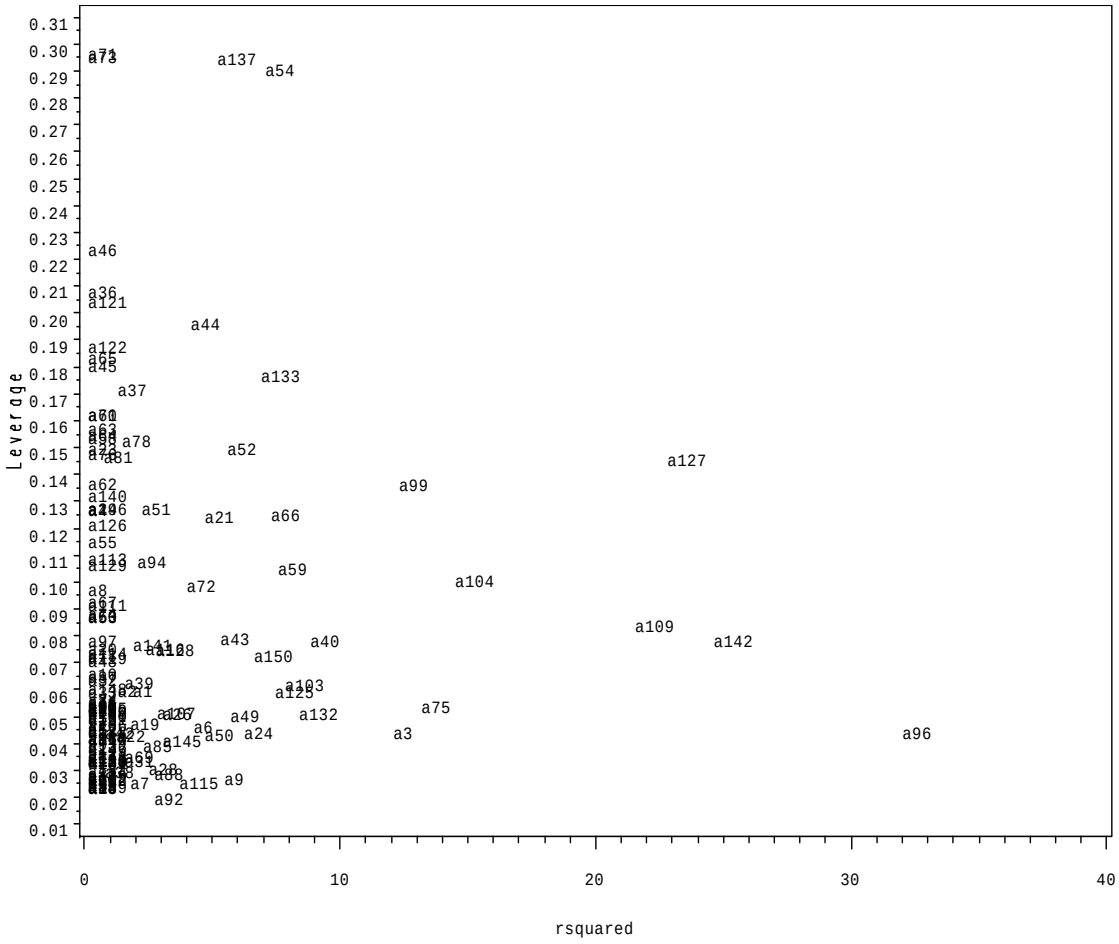
الجدول (3)

المشاهدات ذات القوة الرافعة

Leverage	name
0.30768	a71
0.30598	a137
0.30174	a54
0.28871	a73
0.21695	a46
0.20131	a36
0.19795	a121
0.18931	a44
0.18093	a122
0.17669	a65
0.17384	a45
0.17043	a133
0.16533	a37
0.15583	a70
0.15563	a61
0.1504	a63
0.14813	a64

الشكل (5)

المشاهدات ذات القوة الرافعة في أعلى الزاوية اليسرى والمشاهدات التي تمتلك البواقي العالية في أسفل الزاوية اليمنى

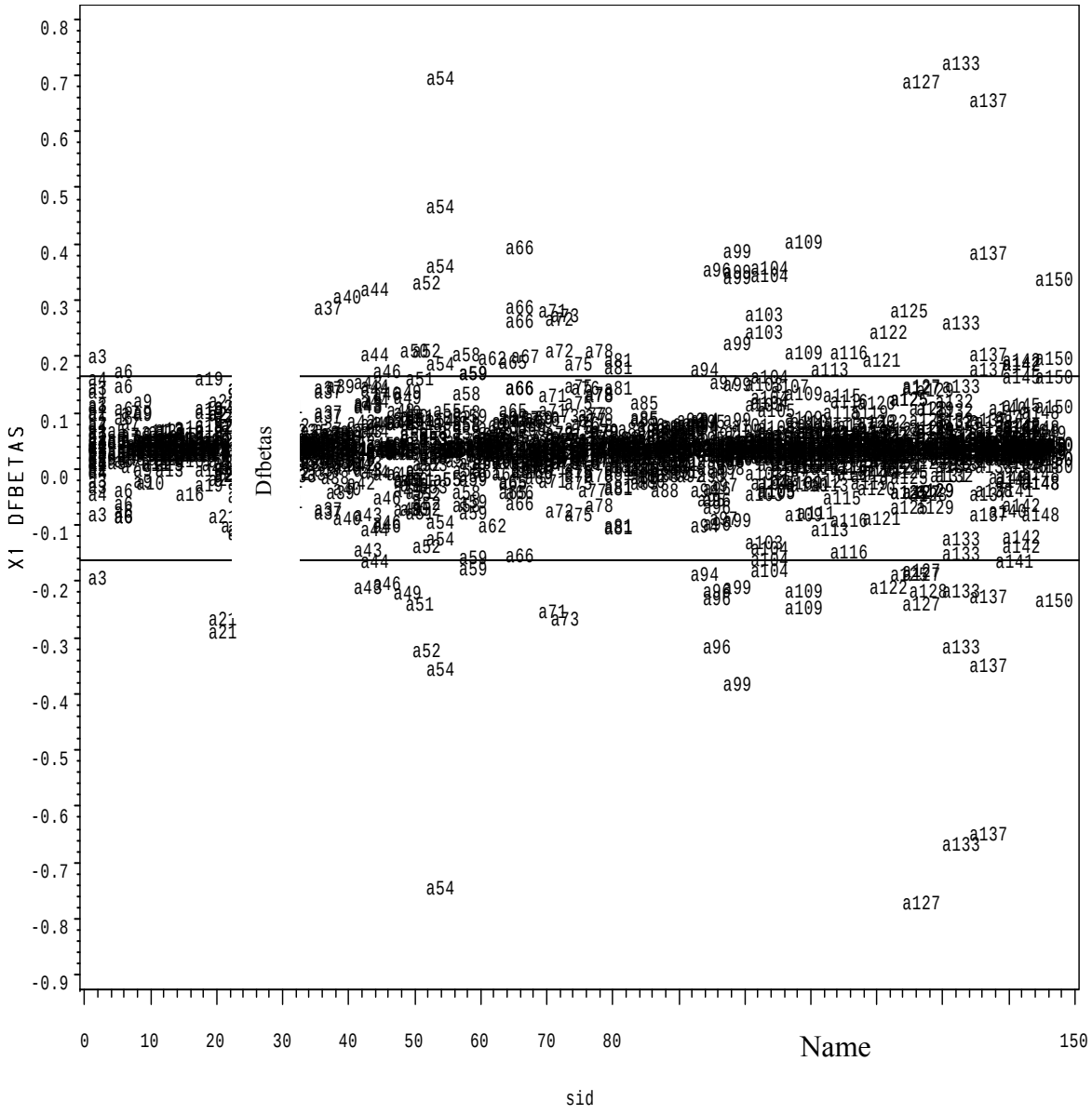


3-2تشخيص المشاهدات المؤثرة (influence):-

إن المقياس الأكثر دقة وتحديدا في قياس السلوك ومقدار التغير الذي يحصل في المعلمات المقدرة عند حذف المشاهدة هو ما يطلق عليه بمقياس (Dfbeta) (إن قيم Dfbeta يمكن أن تكون سالبة أو موجبة،وذلك لأنها تعبر عن مساهمة المشاهدة في تغيير معاملات الانحدار زيادة أو نقصانا).من الجدول (4) ومن الشكل (6) يتبين أن المشاهدات (a127,a54,a137,a133) تمتلك اكبر قيمة للمقياس (Dfbeta) والبالغة (0.701,0.7212,0.7813,-0.808-) على التوالي، حيث أن المشاهدة (a127) تقف في مقدمة المشاهدات الأكثر تأثيرا في المتغير X_{10}،إذ أن إضافة هذه المشاهدة أو حذفها ستعمل على إحداث تغيرات جوهرية في معادلة الانحدار،ويبدو هذا جليا،إذ يلاحظ أن حذف المشاهدة (127) قد أدى إلى تغيير إشارة المتغير X_{10} من الإشارة السالبة إلى الموجبة،كما أن قيمة t ارتفعت من (0.32-) إلى (0.47).

الشكل (6)

قيم (DFbeta) في المتغيرات التوضيحية



4-2 اثر حذف المشاهدات الشاردة في معاملات الانحدار في التوزيع الطبيعي:-

من الجدول (5) يمكن ملاحظة مدى تأثير حذف المشاهدات الشاردة ((a96)،(a142)،(a127)، (a109)) في التوزيع الطبيعي للبواقي، اذ يلاحظ أن البواقي قبل حذف المشاهدات الشاردة كانت تتوزع توزيعا غير طبيعيا،وعند حذف المشاهدة (a96) تحولت البواقي إلى التوزيع الطبيعي استنادا إلى اختبار (W)،وهكذا مع حذف بقية القيم الشاردة يلاحظ أن التوزيع يقترب من التماثل تماما.كما يلاحظ أيضا اثر حذف المشاهدات الشاردة في دقة المعلمات وعلى انخفاض متوسط مربعات الخطأ (MSE) وكذلك في ارتفاع قيمة (f) وارتفاع قيمة معامل التحديد وفي انخفاض (VIF)،كما يلاحظ أن إشارة المعلمة المقدرة للمتغير X_{10} قد تغيرت من السالب إلى الموجب،وهذا التغير يتفق مع المنطق الطبي لموضوع البحث.

2-5 الكشف عن مشكلة تعدد العلاقات الخطية ومعالجتها:-

الآن وبعد أن تم الكشف عن المشاهدات الشاردة واستبعادها من تحليل الانحدار،يمكن معالجة مشكلة تعدد العلاقات الخطية من خلال الكشف عن المتغيرات المسببة لهذه المشكلة ومعالجتها من خلال حذف بعض المتغيرات كأحد وسائل المعالجة للوصول إلى نتائج متفقة مع وجهة النظر الطبية.

يلاحظ من الجدول (6) أن المتغيرات (X_1,X_4,X_5,X_9) لها دليل شرطي (condition index) اكبر من (30) وان جذورها المميزة (eigen value) مساوية أو اقل من المقدار (0.00624) وان نسب تحلل التباين (variance proportion) اكبر من (0.5) ومن خلال (VIF) يتضح أن هذه المتغيرات هي المسببة لمشكلة تعدد العلاقات الخطية،لذا فقد تم حذف بعض من هذه المتغيرات (انظر الجدول (7)) وتم التوصل إلى مجموعة من العلاقات المنطقية والمتفقة مع متطلبات طريقة المربعات الصغرى الاعتيادية،اذ يلاحظ أن العلاقة الثالثة فسرت (0.76%) من التغيرات الكلية،كما أن جميع المعلمات متفقة مع وجهة النظر الطبية،لذا تم اعتمادها،إلا أن هذا لا يمنع من دراسة المتغيرات التوضيحية الأخرى وأثرها في المتغير المعتمد من خلال العلاقات الأخرى،وذلك لبيان اثر هذه المتغيرات على المتغير المعتمد.

لقد تم استخدام انحدار الحرف كوسيلة أخرى من وسـائل معالجـة تعـدد العلاقات الخطية بعد استبعاد المشاهدات الشاردة.

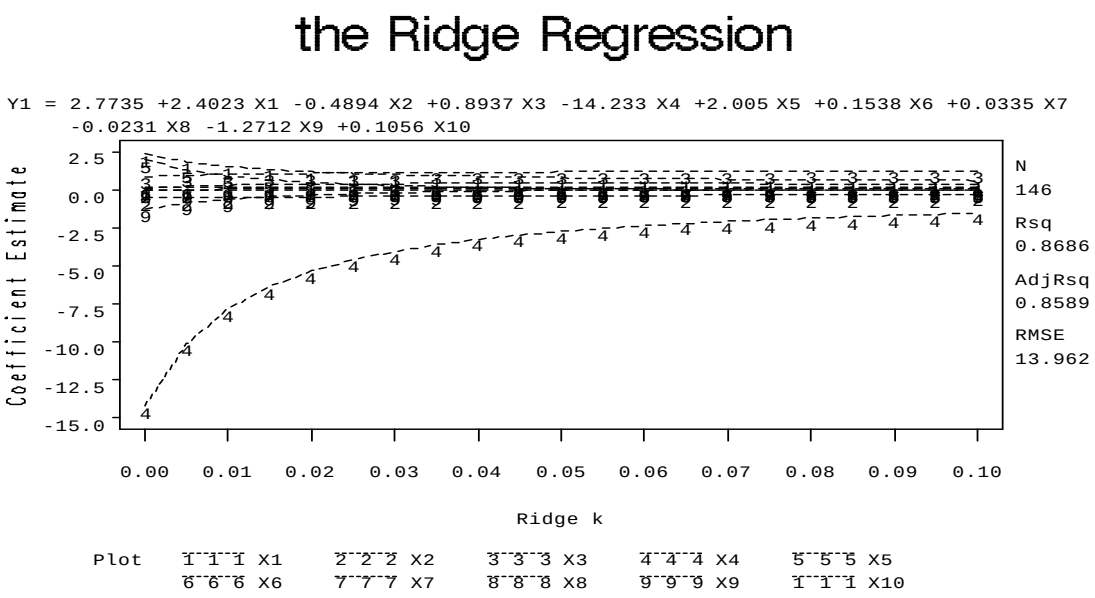
يلاحظ من الشكل (7) أنه عند إضافة المقدار $^{*}(C)$ (تمت إضافة قيم C ابتداء من الصفر إلى 0.1 بزيادة قدرها (0.005) إلى مصفوفة المعلومات ومن ثـم إيجـاد المعلمات المقدرة عند كل قيمة من قيم C،كما تم إيجاد (VIF) فضلا عـن الجـذر ألتربيعي لمتوسط مربعـات الخطـأ (Root Mean Square Error) (RMSE) والانحرافات المعيارية للمعلمات المقـدرة) فـان المعلمـات المقـدرة للمتغيـرين (X_4,X_5,X_9) استقرت عند (C=0.02) في حين أن المعلمة المقدرة للمتغيـر(X_1) استقرت عند (C=0.08) لذا يمكن اعتبار هذه القيمة هي القيمة المثالية للمعلمـات المقدرة .

يلاحظ من الجدول (8) الذي يمثل خلاصة لنتائج انحدار الحرف عنـدما C=0.08 .إن قيم (VIF) اقل من (10) لكل المتغيرات،كما أن المعلمات المقـدرة للمتغيرات (X_3,X_7,X_9,X_{10}) معنوية عند مستوى المعنوية (5%) ومن دون وجود أي ملامح لتعدد العلاقات الخطية،كما أنها متفقة تماما مع وجهة النظر الطبية،عليه يمكن كتابة معادلة الانحدار بالصيغة الآتية:-

$$y=7.97258+1.25253X_3+0.05396X_7+0.07576X_9+0.42331X_{10}$$

الشكل (7)

شكل يمثل المعلمات المقدرة باستخدام انحدار الحرف



*) وهي كمية صغيرة موجبة تقع قيمتها بين الصفر والواحد ($0<C<1$) تضاف إلى العناصر القطرية لمصفوفة المعلومات ($\mathbf{X'X}$) في طريقة المربعات الصغرى الاعتيادية (Ordinary least square) (OLS) للحصول على مقدرات أكثر دقة، حيث تعمل هذه الطريقة على فك الارتباطات بين المتغيرات التوضيحية.

ولغرض المقارنة فقد تم إيجاد مقدرات انحدار الحرف عندما ($C=0.08$) للبيانات الأصلية (بدون حذف المشاهدات الشاردة)،يلاحظ من الجدولين (8و9)،أن ألـ (RMSE) والانحرافات المعيارية لمعلمات انحدار الحرف المقدرة عند استبعاد المشاهدات الشاردة (جدول رقم 8) هي اقل منها عند استخدام البيانات الأصلية (جدول رقم 9)، كما تم الحصول على أربعة متغيرات معنوية عند حذف المشاهدات الشاردة مقابل ثلاثة متغيرات معنوية في حالة البيانات الأصلية.

الاستنتاجات:-

إن من أهم الاستنتاجات التي تم التوصل إليها:-

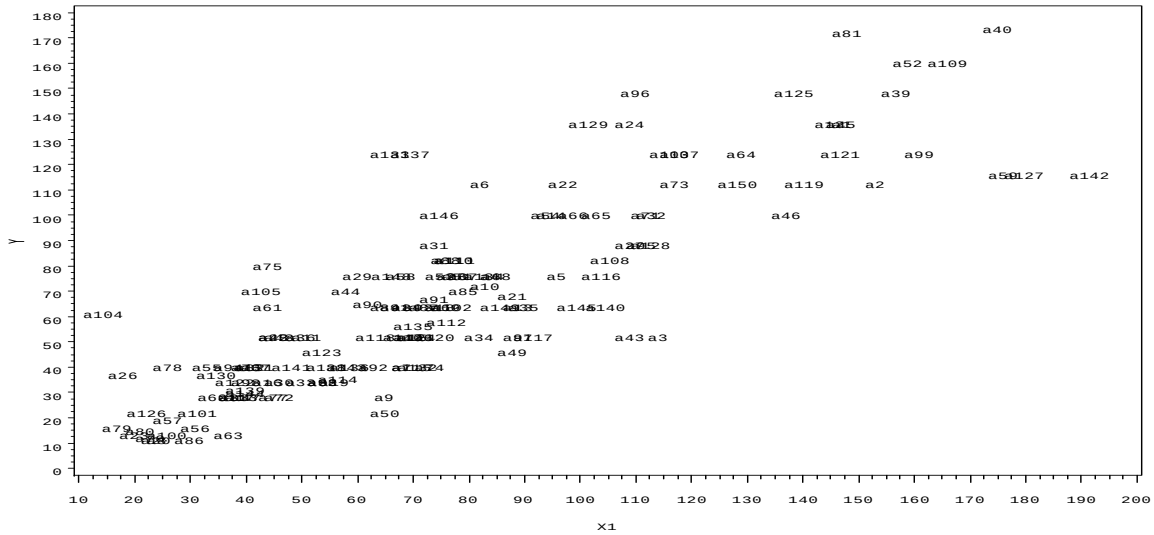
1. اتضح اثر تشخيص المشاهدات غير العادية ومعالجتها في الحصول على تقديرات تتفق مع وجهة النظر الطبية والنظرية الإحصائية.
2. تبين أن المتغيرات التوضيحية المؤثرة في المتغير المعتمد (العمر من العظم) هي تضخم الكبد مقاسا (بالسنتمتر)(X_3) و أرومة حمراء(X_7) و عدد وحدات الدم(X_9) و بداية نقل الدم حسب العمر مقاسا (بالشهر)(X_{10}).وان في مقدمة المتغيرات التي لها تأثير في العمر من العظم هو تضخم الكبد حيث بلغت معلمته المقدرة (1.25243)،وان المتغير الثاني من حيث الأهمية هو بداية نقل الدم حسب العمر حيث بلغت معلمته المقدرة (0.42331)،ومن ثم عدد وحدات الدم (0.07576)،وأخيرا الأرومة الحمراء (0.05396).
3. أتضح اثر حذف المشاهدات الشاردة في دقة النموذج عند استخدام انحدار الحرف حيث بلغت (RMSE) (17.8257) بينما بلغت للبيانات الأصلية (19.5796)،وكذلك على دقة المعلمات المقدرة من خلال الحصول على انحرافات معيارية اقل مما كانت عليه في البيانات الأصلية.
4. إمكانية حذف المتغيرات المسببة لمشكلة تعدد العلاقات الخطية كأحد طرائق المعالجة والوصول إلى أكثر من علاقة منطقية معززة بالنظرية الإحصائية.

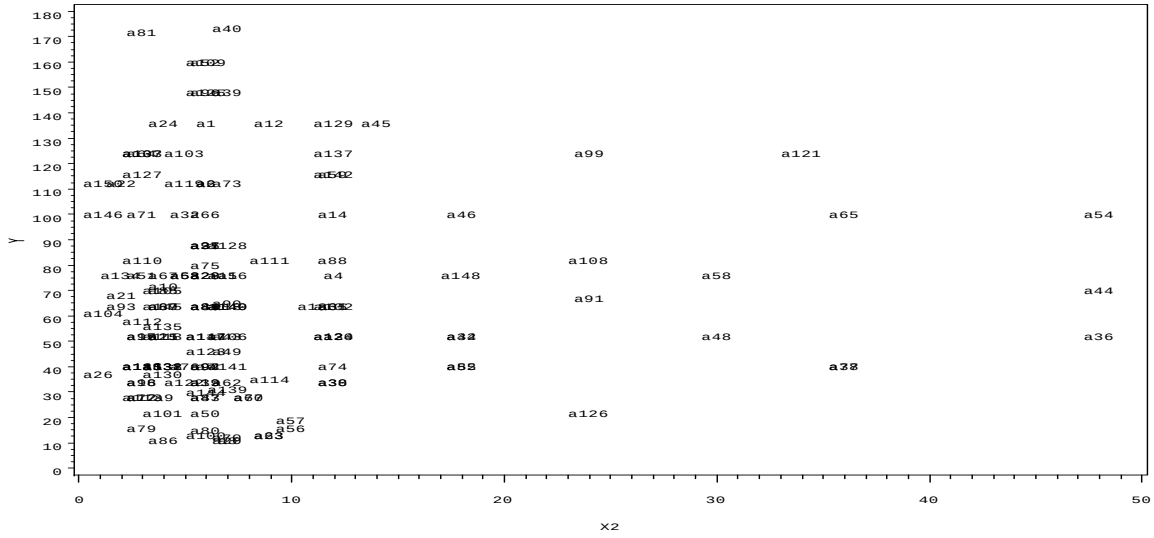
المصادر:-

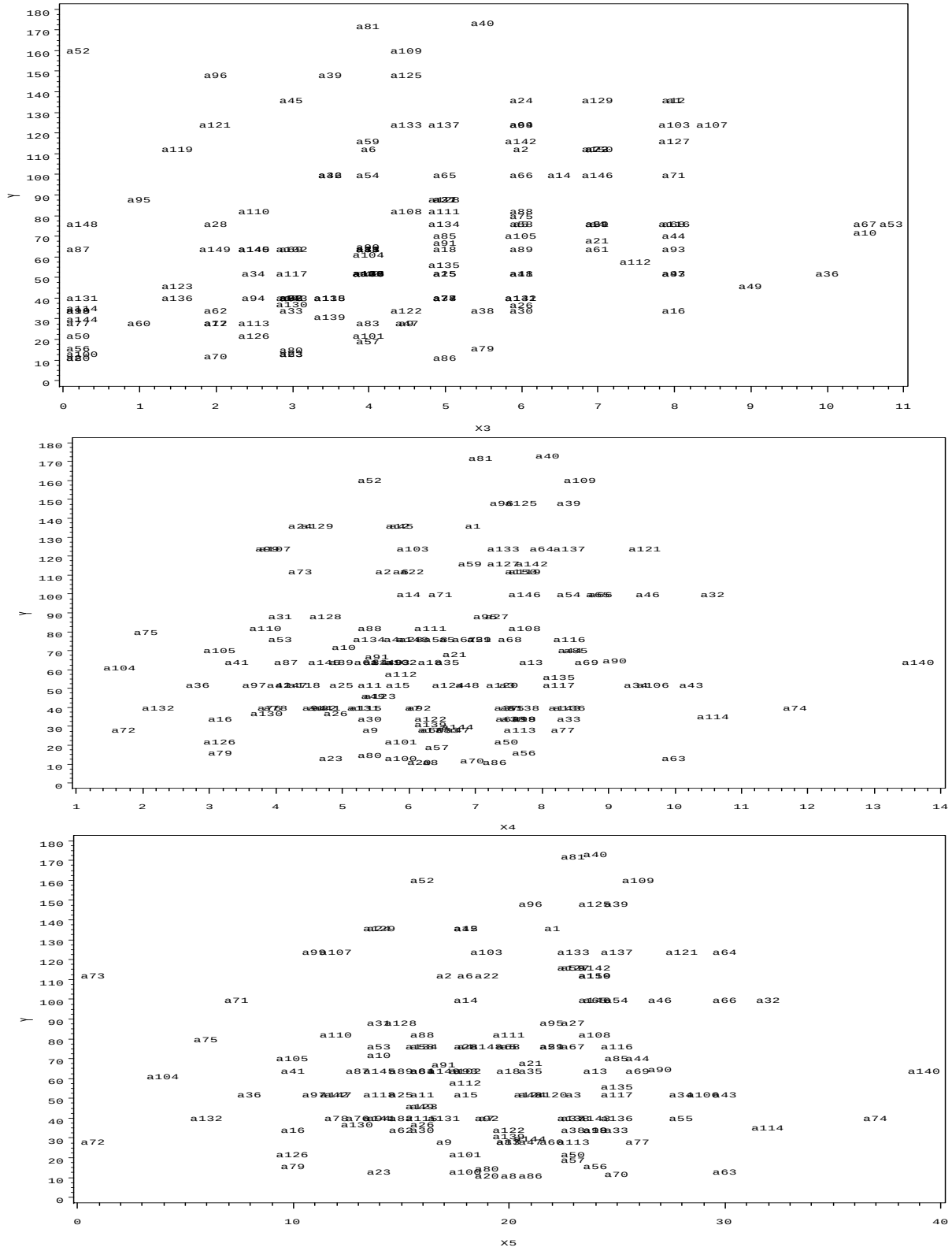
1. النعيمي،أسوان محمد طيب (2005):"اختيار المتغيرات في انحدار الحرف"، رسالة ماجستير غير منشورة،كلية علوم الحاسبات والرياضيات،جامعة الموصل.
2. Bacon,D. (1995).”Amaximum likelihood approach to correlational outlier identification”.Multivariate Behavioral Research,30,125-148.
3. Beckman,R.J. & Cook, R. D. (1983),”Outlier …s”,Technometrics,25,119-163.
4. Chatterjee, S. & H. (1986).”Influential observations,High leverage points,and outlier in linear regression”,Statistical Science,1(3),379-393.
5. Evans,Victoria(1999).”Strategies for Detecting Outlier in Regression Analysis: An Introductory Primer”Paper presented at the annual meeting of the Southwest Educational Research Association ,San Antonio.
6. Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. and Black, W. C. (1998) “Multivariate Data Analysis”, 5th edn. Prentice-Hall, New Jersey.
7. Hoaglin,D.C. & Welsch, R.E. (1978).”The Hat matrix in regression and ANOVA”.The American Statistician,32,17-22.
8. Hordo, M. (2004) ,”Outlier diagnostics on permanent sample plot network in Estonia”. – Research for Rural Development 2004. International Scientific Conference Proceedings. Latvia University of Agriculture, Jelgava, pp. 181-186.
9. Iglewicz, B. y Hoaglin, D.C. (1993). “How to Detect and Handle Outliers”. Quality Press, American Society for Quality Control, Milwaukee, Wisconsin.
10. Michael Frigge and David C. Hoaglin and Boris Iglewicz.(1989) "Some Implementations of the Boxplot". The American Statistician. Vol. 43 (1),pp 50–54.
11. Sinha,Sanjoy Kumar,(1997)”Sequential Application of Multivariate Outlier Test:A Robust Approach”,Master Thesis,Dalhousie University,Halifax,Nova Scotia,pp.1-8 .

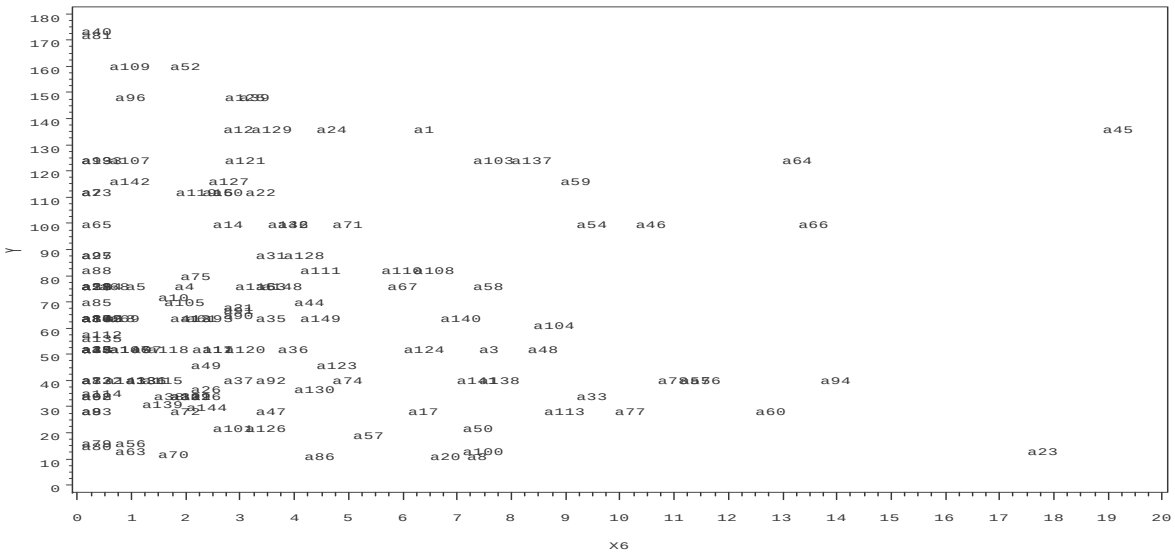
الملحق(1)

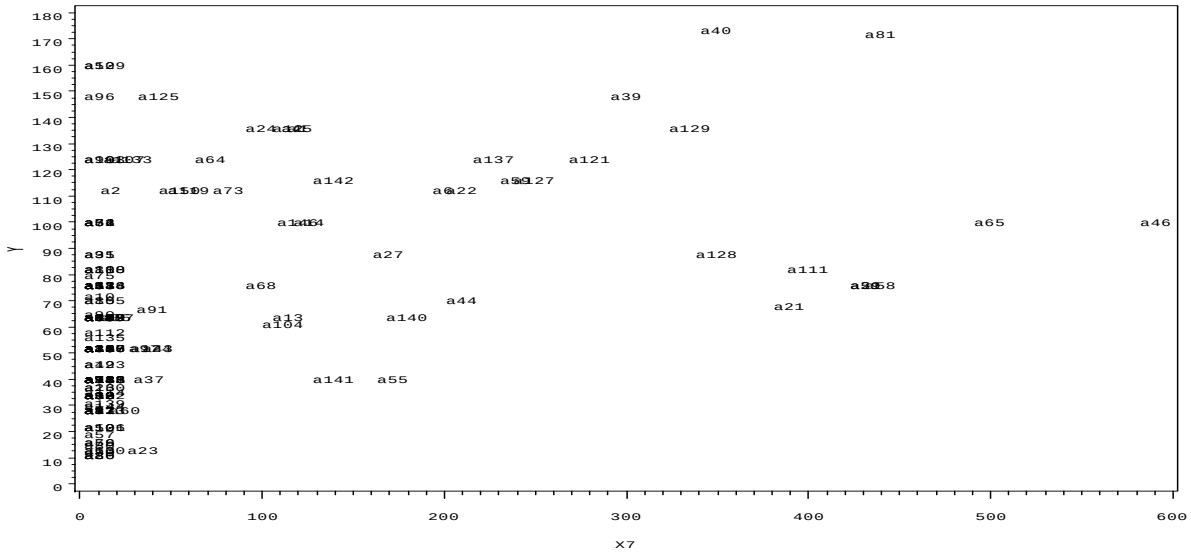
رسوم المتغيرات التوضيحية لمرض ثلاسيميا الأطفال مقابل المتغير المعتمد

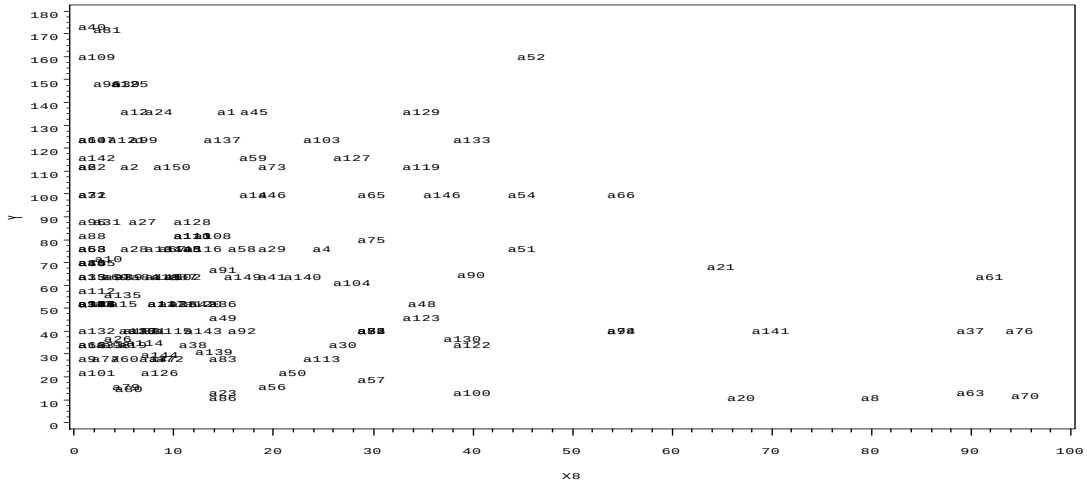


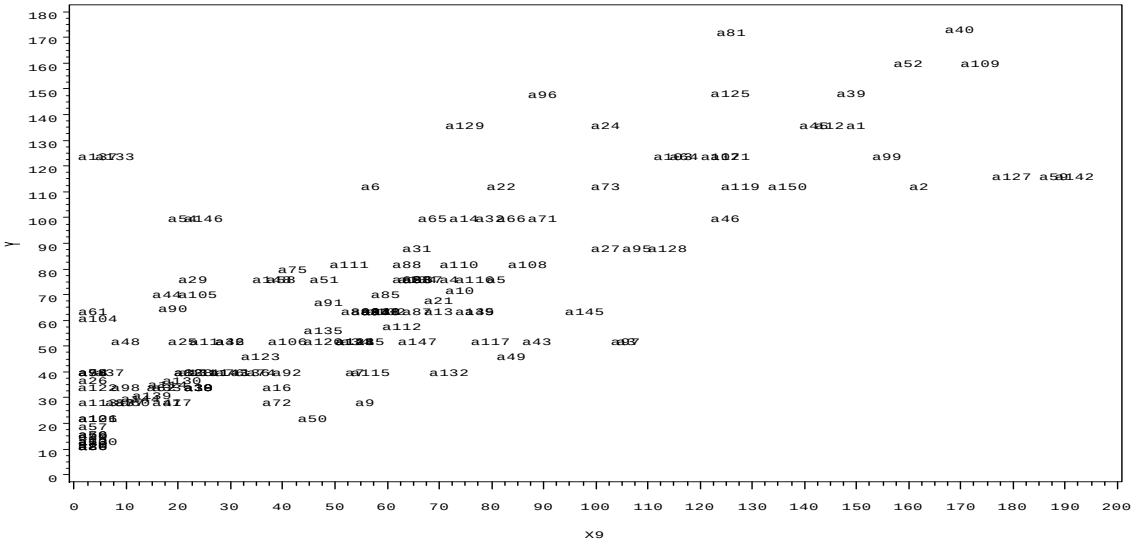


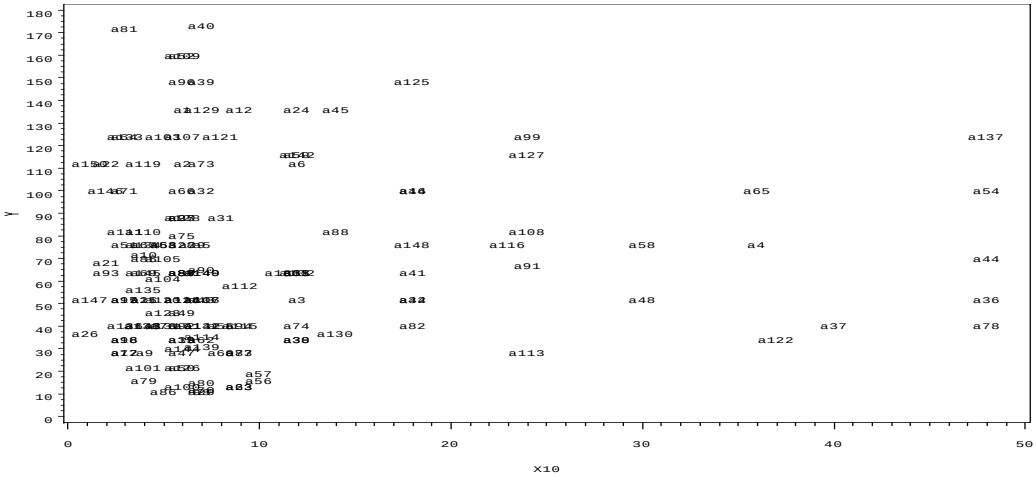

y
x3
y
x4
y
x5


y
X6


y
X7


y
x8


y
x9


y
x10