المجلة العراقية للعلوم الاحصائية (10) 2006
ص ص [85-106]

طرائق مقترحة في انحدار الحرف

مروان عبد العزيز دبدوب* **أسوان محمد طيب النعيمي****

الملخص

طرحت في هذا البحث ثلاثة مقترحات لتعيين قيمة معلمة التحيز في انحدار الحرف، وهي: 1- وضع المعالم المقدرة ذات القيم المطلقة المتقاربة في مجاميع منفصلة، ومن مخطط أثر الحرف لكل مجموعة يحدد مدى لقيمة معلمة التحيز، 2- بتدوير محور أثر الحرف، 3- تطبيق مراحل تحليلية عدة بهدف التوصل إلى قيمة معلمة التحيز المثالية.

كما قُدم أسلوب لاختيار أفضل المتغيرات لتكون في معادلة انحدار الحرف التقديرية. وكان للمصفوفة والمتجه الموسعين أهمية في تبسيط العمليات التحليلية كافة.

Suggested Methods in Ridge Regression

ABSTRACT

Three suggested procedures were adopted to determine the value of biasing parameter (k) in ridge regression: 1-fragmenting the ridge trace to groups each group contains semi-homogeneous absolute values of the estimated parameters, 2-rotating over the axis of ridge trace, 3-the ideal value of k should be determined through many analytical stages.

A procedure was suggested to select the best variables to admit in the ridge regression equation. The augmented matrix

* أستاذ مساعد/ قسم الإحصاء/ كلية علوم الحاسبات والرياضيات.

** مدرس مساعد/ قسم نظم المعلومات الإدارية/ كلية الإدارة والاقتصاد.

تاريخ التسلم : 31/ 5 /2006 ـــــــــــــــــ تاريخ القبول : 21/ 6 /2006

and vector have had a significant importance in facilitating all analyzing processes.

المقدمة

إن تحليل الانحدار من الطرائق الإحصائية الواسعة الاستخدام، الذي يوضح العلاقة بين متغيرات توضيحية ومتغير تابع على هيئة نموذج، بما يفيد التخطيط واتخاذ القرارات.

إن طريقة المربعات الصغرى الاعتيادية تعطي أفضل تقدير خطي غير متحيز وبأقل تباين لمعالم أنموذج الانحدار. ومن المشكلات التي قد تظهر عند اتباع هذه الطريقة هي غياب أحد فروض التحليل الذي يشترط عدم توافر ارتباط خطي تام أو جزئي بين اثنين أو أكثر من المتغيرات التوضيحية، مما قد يؤدي إلى ظهور مشكلة تدعى بتعدد العلاقة الخطية التي تسبب إعطاء تقديرات ضعيفة للمعالم وذات تباينات متضخمة، ومن ثم الحصول على نتائج لاختبار الفرضيات لا يعول عليها.

إن أول من أشار إلى خطورة تعدد العلاقة الخطية وتأثيرها في نتائج تحليل الانحدار هو Fisher وكان ذلك عام 1934 (النعيمي، 2005). وتتابع الكثير من الباحثين الذين أرسوا الجوانب المختلفة للمشكلة وطرائق حلها، حتى أضاف Hoerl and Kennard (1970a) مقدارا موجبا قيمته بين الصفر والواحد إلى عناصر قطر مصفوفة المعلومات **X'X**، وقد أطلقا على المقدار الثابت "معلمة التحيز Biasing parameter"، وعلى الطريقة بانحدار الحرف Ridge regression.

وضع Hoerl et al. (1975) طريقة لاختيار قيمة معلمة التحيز (k)، وحوّرها Hoerl and Kennard (1976) لتصبح طريقة تكرارية Itrative method. كما أشار Montgomery and Peck (1982) إلى طريقة أطلقا عليها أثر الحرف Ridge trace تعتمد على مسار منحنيات المعالم المقدرة مقابل عدد من القيم الثابتة بين الصفر والواحد الموجب. وقد أشار المشهداني (1994) إلى طريقة لاختيار معلمة التحيز محورة عن الطريقة التي وضعها Hoerl et al.

(1975) الغاية منها تقليل قيمة معلمة التحيز اللازمة للتغلب على تعـــدد العلاقـــة الخطية.

هدف البحث:

الكشف عن وجود تعدد العلاقة الخطية، ثُمَّ محاولة تحقيق المقترحات الآتية:

1- طرائق لاختيار قيمة معلمة التحيز (الثابت k)، وهي:

أ- وضع المعالم المقدرة ذات القيم المطلقة المتقاربة في مجاميع، ومـــن كـــل مجموعة يتم الحصول على مخطط لأثر الحرف، ومن المخططات يتم تحديد مدى لقيمة معلمة التحيز.

ب- تدوير محور أثر الحرف، وتحديد مدى تقاطع المنحنيات، مركز التقاطعـــات هي القيمة التقريبية لمعلمة التحيز.

ت- في كلا المقترحين أ وب حُدد مدى لقيمة معلمة التحيـــز، ولتحديـــد القيمـــة المثالية تتبع مراحل تحليلية عدة.

2- استخدام مخططات أثر الحرف للمجاميع المشار إليها في (أ-1) فـــي اختيـــار أفضل المتغيرات لتكون في معادلة انحدار الحرف التقديرية.

لمحة عن تحليل الانحدار:

تحليل الانحدار هو أسلوب البحث عن دالة رياضية تفيد في وصف العلاقة بين متغير تابع ومتغيرات توضيحية بما يفيد التنبؤ والسيطرة (كاظم ومســـلم، 2002)، وصيغة الدالة للحصول على القيمة i من قيم المتغير التابع هي:

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} + \varepsilon_i \qquad \dots \quad (1)$$

حيث أن: عدد المشاهدات $i=1,2,...,n$; عدد المتغيرات $j=1,2,...,m$.

Y_i: متغير تابع.

X_j: m من المتغيرات التوضيحية.

β_0, β_j : ثوابت وهي معالم الانحدار.

ε_i: الخطأ العشوائي.

وتُقَّدر معالم النموذج (1) بإتباع طريقة المربعات الصغرى حسب المساواة الآتية:

$$\hat{\beta}=(\mathbf{X'X})^{-1}\mathbf{X'y} \quad \ldots \quad (2)$$

وتصاغ معادلة الانحدار التقديرية التي تعطي n من قيم متوسط الاستجابة ($\hat{y}$) كالآتي:

$$\hat{\mathbf{y}}=\mathbf{X}\hat{\beta} \quad \ldots \quad (3)$$

قد يواجه المشتغل بهذه الطريقة مشكلة تدعى "تعدد العلاقة الخطية".

تعدد العلاقة الخطية Multicollinearity:

يواجه الباحث مشكلة تعدد العلاقة الخطية عند ارتباط اثنين أو أكثر من المتغيرات التوضيحية بعلاقة خطية، وبذلك لا يمكن فصل تأثيرها في المتغير المعتمد. وكلما زاد عدد المتغيرات التوضيحية في النموذج أصبح الكشف عن المشكلة معقداً (دبدوب، 1998).

قد يكون التعدد الخطي تاما، وبذلك تكون مصفوفة المعلومات **X'X** برتبة غير كاملة، عندئذ لا يمكن تطبيق المعادلة (2). وقد يكون التعدد الخطي غير تام، عندما تكون بعض المتغيرات التوضيحية دالة في التركيبة نفسها لمتغيرات أُخرى مع قيم عشوائية، وهنا يكون محدد مصفوفة المعلومات صغيراً مما يؤدي إلى تضخيم تباين المعالم المقدرة، كما يعجز النموذج عن إظهار أثر المتغيرات التوضيحية منفصلةً في المتغير التابع نظراً لترابطها (النعيمي، 2005).

بعض طرائق الكشف عن التعدد الخطي:

لا يوجد اختبار لفرضية إحصائية للدلالة على وجود تعدد العلاقة الخطية لانها لا معلمة لها، وتتبع طرائق إحصائية مختلفة للدلالة على احتمالية وجود مثل هذه العلاقة، إن فشل طريقة ما ليست دليلا على عدم توافرها. ومن أكثر الطرائق استخداماً هي:

1- اختبار Farrar and Glauber (1967): يعتمد الاختبار على قيمة محسوبة لمربع كآي (Chi square) بالاعتماد على محدد مصفوفة معامل الارتباط $|\mathbf{R}|$

وعدد المتغيرات التوضيحية (m) وعدد الوحدات التجريبية (n) وبدرجات حرية m(m-1)/2.

2- تضخيم تباين العوامل Variance inflation factors (VIF): وضع هذا المقياس Marquardt (1970)، ويمكن التكهن بوجود التعدد الخطي عند زيادة قيمته عن 10، ويعتمد على معامل التحديد (R^2) Coefficient of determination.

3- مقياس العدد الشرطي Conditional number (CN): وهو النسبة بين أكبر وأصغر جذر مميز (Characteristic root) الناتجة من تحليل مصفوفة المعلومات، وقد أشار Montgomery and Peck (1982) الى ضرورة الاستقصاء عن وجود تعدد العلاقة الخطية عند زيادة العدد الشرطي عن 100 .

4- معامل الارتباط البسيط Simple correlation coefficient: تحسب قيمته بين أزواج المتغيرات التوضيحية، فإذا اقتربت من الواحد المطلق أو زادت عن قيمة معامل التحديد وجب التريث بالحكم على عدم توافر تعدد علاقة خطية. وقد يرتبط أكثر من متغيرين بعلاقة خطية في حين يكون معامل الارتباط بين اثنين منهما بعيداً عن الواحد المطلق، ولذا غالباً ما يكون هذا المقياس غير كفٍ في الكشف عن التعدد الخطي.

5- صغر محدد مصفوفة الارتباط بين المتغيرات التوضيحية $|\mathbf{R}|$، وقد أشار Mason et al. (1975) إلى ضعف هذا الأسلوب فقد توجد قيم ذاتية متوسطة الحجم تؤدي إلى صغر المحدد.

6- التغير في إشارات بعض المعالم المقدرة ذات الدلالة الإحصائية بين عينة وأخرى.

تعد الطرائق الثلاث الأولى هي الأسهل استخداما وأكثر شيوعا لكونها تعتمد على قيم ثابتة، خاصة مقياس تضخيم التباين (VIF) الذي يتم الحصول عليه مباشرة بوصفه أحَدَ نتائج تحليل الانحدار عند استخدام بعض البرامج الحاسوبية الجاهزة. وقد ناقش دبدوب (1998) السلبيات في معامل الارتباط ومعامل التحديد

والجذور المميزة وعلاقة الأخيرة بمحدد مصفوفة الارتباط بوصفها مقاييسَ لاتخاذه قراراً باحتمالية وجود تعدد العلاقة الخطية.

وجبت الإشارة إلى فكره شائعة التطبيق ضعيفة النتائج، غايتها التعرف على وجود التعدد الخطي بين المتغيرات التوضيحية، وهي: ظهور أثر ذي دلالة إحصائية عند اختبار الفرضية العامة للنموذج، وعدم ظهور مثل ذلك الأثر عند الاختبار الجزئي لمعالم النموذج، وقد أثبت Geary and Leser (1968) بان هذه الحالة قد تظهر على الرغم من كون المتغيرات متعامدة، وأشار وارطان (1989) إلى أن الاختبار الجزئي لمعالم الأنموذج قد يكون ذا دلالة إحصائية على الرغم من وجود تعدد العلاقة الخطية.

معالجة تعدد العلاقة الخطية:

من الطرائق الشائعة للتغلب على تعدد العلاقة الخطية هي:

1- التحويل المعياري لقيم المتغيرات: ان مثل هذا النمط من التحويل يضعف التداخل الخطي ويخفض قيمة معامل الارتباط بين أزواج المتغيرات.

2-إضافة بيانات جديدة إلى البيانات الأصلية: وهذا مشابه لاسلوب زيادة حجم العينة اذ ترتفع القيم الذاتية لمحدد مصفوفة المعلومات($|\mathbf{X'X}|$)، ولهذه الطريقة بعض النواحي السلبية منها تعود إلى أسباب اقتصادية أو تغيرات في المجتمع المدروس، كما قد يؤدي تضخم البيانات إلى تطرفها.

3- حذف المتغيرات المسببة للتعدد الخطي أو استبدالها: يعاب على هذه الطريقة بحصول مشكله في توصيف البيانات وعدم إستقرارية المعالم المقدرة وظهور أخطاء معيارية فيها.

4- إتباع إحدى طرائق الإحصاء البديلة مثل: المربعات الصغرى المقيدة أو طريقة الدمج في السلاسل الزمنية أو طريقة المكونات الرئيسة أو طريقة انحدار الحرف التي تعد من أكثر الطرائق استخداماً في هذا المجال.

انحدار الحرف Ridge regression:

تتميز طريقة انحدار الحرف بإيجاد قيمة ثابتة (k) تدعى بمعلمة التحيز تضاف إلى عناصر قطر مصفوفة المعلومات **X'X**، وفائدة ذلك هو تقليل قيم عناصر قطر معكوس مصفوفة المعلومات الذي يؤدي إلى خفض قيم تباينات المعالم المقدرة. عند ابتعاد المتغيرات التوضيحية عن الاستقلالية أي عند ارتفاع قوة الارتباط بين أزواج المتغيرات التوضيحية، فإن إضافة الثابت k بقيم صغيره تعمل على تغيير سريع في قيم المعامل المقدرة، ومع زيادة قيمة k تبدأ تلك القيم بالاستقرار تدريجياً إلى أن تصل إلى حد يكون التغير فيها طفيفاً وثابت الإشارة، وكلما كان استقرار المعالم سريعاً دل على أن المتغيرات التوضيحية قريبة من الاستقلالية (النعيمي، 2005).

يتم الحصول على القيم المقدرة لمعالم انحدار الحرف عن طريق المعادلة الآتية:

$$\hat{\beta}_R=(\mathbf{X'X}+k\mathbf{I}_p)^{-1}\mathbf{X'y} \quad \ldots (4)$$

وتتمثل العلاقة بين مقدرات كل من: انحدار الحرف والمربعات الصغرى الاعتيادية، بالآتي:

$$\hat{\beta}_R=[\mathbf{I}_p+k\mathbf{I}_p(\mathbf{X'X})^{-1}]^{-1}\hat{\beta}_{ols}$$

$$=(Z_R)\hat{\beta}_{ols} \quad \ldots (5)$$

من المساواة في (5) يتبين أن مقدرات انحدار الحرف هي تحويل خطي لمقدرات المربعات الصغرى، وعند أخذ التوقع لمقدرات انحدار الحرف يتضح بأنها مقدرات متحيزة لـ β :

$$E(\beta_R)=E(\mathbf{Z}_R\hat{\beta}_{ols})=\mathbf{Z}_R\beta_{ols} \quad \ldots (6)$$

إن قيمة متوسط مربعات الخطأ لمقدرات انحدار الحرف هي:

$$MSE_R=\text{variance}(\hat{\beta}_R)+(\text{bias in }\hat{\beta}_R)^2 \quad \ldots (7)$$

ذكر Montgomery and Peck (1982) أن مقدار التحيز يزداد ومقدار التباين يقل عند زيادة قيمة k، وعلية يجب اختيار قيمة k بحيث أن الانخفاض في قيمة التباين يكون أكثر من الارتفاع في مقدار مربع التحيز، عند ذلك يكون متوسط مربعات الخطأ لانحدار الحرف أقل من التباين لمقدرات المربعات الصغرى الاعتيادية. إن زيادة قيمة k تؤدي إلى انخفاض قيمة معامل التحديد R^2، من هنا

يتضح بأن مقدرات انحدار الحرف ليس من الضروري أن تعطي أفضل مطابقة للبيانات، إذ إننا نبحث عن أفضل معادلة ذات مقدرات ثابتة.

المصفوفة والمتجه الموسعان Augmented matrix and vector:

أشرنا إلى أهمية التحويل المعياري لقيم المتغيرات لإضعاف التداخل الخطي، ومساعدته للتغلب على تعدد العلاقة الخطية بأصغر قيمة لمعلمة التحيز، لذا تُحَّول مصفوفة البيانات $\mathbf{X}$ إلى الصيغة المعيارية $\mathbf{X}^*$ والتي طول كل متجه فيها يساوي واحداً، عندئذ تصبح مصفوفة المعلومات $\mathbf{X}^{*\prime}\mathbf{X}^*$ عبارة عن مصفوفة الارتباط بين المتغيرات التوضيحية. والإجراء نفسه يطبق على متجه قيم المتغير المعتمد لنحصل على المتجه $\mathbf{y}^*$.

لتبسيط العمليات الإحصائية في الحصول على نتائج تحليل انحدار الحرف، تُوسَّع المصفوفة المحولة $\mathbf{X}^*$ والمتجه المحول $\mathbf{y}^*$ كالآتي: توسع لاحقاً صفوف $\mathbf{X}^*$ بمصفوفة قطرية حجمها $m\times m$ عناصرها الجذر التربيعي لقيمة معلمة التحيز (الثابت k)، تسمى المصفوفة الناتجة بالمصفوفة الموسعة $(\mathbf{X_a})$ Augmented matrix حجمها $(n+m)\times m$. كما تضاف لاحقاً قيم صفرية بعدد m إلى $\mathbf{y}^*$ لنحصل على المتجه الموسع $(\mathbf{y_a})$ Augmented vector ذي الحجم $(n+m)\times 1$.

تطبق المعادلتان (2 و3) على المصفوفة الموسعة والمتجه الموسع للحصول على نتائج تحليل انحدار الحرف بمعلمة تحيز k، ويمكن استخدام أي برنامج حاسوبي جاهز، فمثلا في Minitab بإصداراته المختلفة يمكن اتباع المسار الآتي:

Stat→Regression→Regression→Response□+Predictors□ ... (8)

حيث يشار إلى أعمدة المصفوفة الموسعة في موقع Predictors وإلى المتجه الموسع للمتغير المعتمد في موقع Response.

كما تسهل هذه الطريقة الأساليب التحليلية للتوصل إلى أفضل المتغيرات التوضيحية لإدخالها إلى معادلة انحدار الحرف التقديرية، وذلك باتباع المسار الآتي:

Stat→Regression→Stepwise

Or Subsets→ Response□+Predictors□ ... (9)

اختيار قيمة معلمة التحيز:

من طرائق اختيار معلمة التحيز هي:

1- طريقة Hoerl et al. (1975):

$$k=(m\hat{\sigma}^2)/(\hat{\beta}'_{ols}\hat{\beta}_{ols}) \quad \dots (10)$$

وقد أشار المشهداني (1994) إلى تحوير في الطريقة (10) لتصبح كالآتي:

$$k=[(m-2)\hat{\sigma}^2]/(\hat{\beta}'_{ols}\hat{\beta}_{ols}) \quad \dots (11)$$

m: عدد المتغيرات التجريبية.

$\hat{\sigma}^2$: تباين المجتمع المقدر بطريقة المربعات الصغرى من البيانات الأصلية.

$\hat{\beta}_{ols}$: المعالم المقدرة بطريقة المربعات الصغرى من البيانات الأصلية.

كما وضع دبدوب والكاتب (2005) أسلوباً لتعيين أفضل قيمة لمعلمة التحيز بالاعتماد على الطريقتين (10 و11)، ولا يعول على أسلوبيهما عند فشل أيٍ منهما في التغلب على تعدد العلاقة الخطية.

2- وضع Hoerl and Kennard (1976) طريقة تكرارية بالاعتماد على المعادلة (10) التي ستعطي أول قيمة لـِ k (لنرمز لها بـِ k_0) والتي بواسطتها يتم إيجاد مقدرات لمعالم انحدار الحرف بتطبيق المعادلة (4)، ومن ثم إيجاد قيمة جديدة لـِ k_{p+1} (حيث أن: $p=0,1,2,\dots$) وذلك بتطبيق المعادلة (12) أدناه، ثم نعود مرة أخرى لتطبيق المعادلة (4)، وهكذا نستمر بالعملية التكرارية حتى تتحقق المقارنة (13).

$$k_{p+1}=(m\,\hat{\sigma}^2)/[\hat{\beta}'_{R(k)}\,\hat{\beta}_{R(k)}] \quad \dots (12)$$

$$(k_{p+1}-k_P)/k_p \le 20\,T^{-1.3} \quad \dots (13)$$

حيث أن:

$$T=tr(\mathbf{X'X})^{-1}/m$$

3- أدناه طرائق مقترحة، وهي:

أ- أثر الحرف Ridge trace: هو مخطط يحتوي على m من المنحنيات تمثل مسار المعالم المقدرة عند كل قيمة من قيم k_p، المحور العمودي يمثل قيم معلمة

التحيز k_p، والمحور الأفقي يمثل قيم المعالم المقدرة β_{jp}، يتم اختيار قيمة k التي عندها تبدأ المنحنيات بالاستقرار (Montgomery and Peck, 1982).

لوحظ تأثير سلبي في وضوح مسار منحنيات أثر الحرف تسببه زيادة في عاملين، هما: 1- عدد المتغيرات، 2- الاختلاف في قيم المعالم المقدرة المقابلة لقيم معلمة التحيز. لذا اقتُرِحَ وضع المتغيرات ذات قيم مطلقة متقاربة للمعالم المقدرة في مجاميع مختلفة، كل مجموعة تعطي مخططا لأثر الحرف، ومن كل مخطط تعين قيمة لمعلمة التحيز، وبذلك نحصل على مدى لقيم معلمة التحيز.

ب- تدوير محور أثر الحرف: يعطي أثر الحرف عدد m من المنحنيات بمسارات مستقلة عن بعضها. لنفترض أن المنحنيات وضعت في اسطوانة، ووزعت المنحنيات بزوايا متساوية على القاعدة، ويرتفع كل منحنى بمسار تحدده قيم معالمه المقدرة المقابلة لقيم معلمة التحيز k_p. ويحدد ارتفاع الاسطوانة بين أكبر وأصغر قيمة لـِ k_p الواقعة على المركز.

عند النظر إلى المنحنيات من أعلى الاسطوانة يلاحظ اختلاف في مسارها، وقد يتجاوز بعضها مركز الاسطوانة إلى الجانب الآخر. وعند النظر من الجانب، يلاحظ تغير في مسار بعض المنحنيات في مرحلة ما بعد القيم الصغيرة لمعلمة التحيز محدثة بذلك تقاطع المنحنيات في مواقع عدة، وحتى تعين حدود التقاطعات وجب النظر إلى الاسطوانة من الجانب مع تدويرها، إن القيمة المثالية لمعلمة التحيز تكون قريبة من مركز مدى التقاطعات.

ت- إن الطريقتين المقترحتين المشار إليهما انفا تعطيان مدى لقيمة معلمة التحيز، ولذا اقتُرِحَ أسلوب لتعيين قيمة معلمة التحيز المثالية كالآتي: عند كل قيمة من قيم مدى معلمة التحيز يطبق تحليل الانحدار بإتباع المسار الحاسوبي (8)، ويتم ذلك بالاستعانة بالمصفوفة والمتجه الموسعين. وتحدد أصغر قيمة k عندها يتم التغلب على تعدد العلاقة الخطية.

إن قيمة k المثالية هي تلك القيمة التي تبعد مقاييس التعرف على تعدد العلاقة الخطية عن حدودها العليا دون إحداث تأثير يذكر في مقاييس المفاضلة بين نتائج تحليل الانحدار، مثل الدلالة الإحصائية لمعالم النموذج وقيم R^2 وMSE.

اختيار المتغيرات:

توجد طريقتان شائعتان لاختيار أفضل المتغيرات لتكون في معادلة انحدار الحرف التقديرية، هما: الطريقة الأولى: أشار Hoerl and Kennard (1970b) إلى استبعاد المتغير إذا اتصفت قيم معالمه المقدرة المقابلة لقيم k_p بإحدى الصفات الآتية: 1- عدم استقرارها، 2- انحدارها إلى الصفر، 3- ذات قيم قياسية صغيرة مقارنة بالمعالم المقدرة للمتغيرات الأخرى. إن هذا الأسلوب يعتمد على ملاحظة قيم المعالم المقدرة لكل متغير بصورة منفردة، وهذا بدوره يلغي العملية التعويضية لمتغير آخر أو مجموعة من المتغيرات في إظهار تأثيرها بديلةً عن المتغير المدروس، والطريقة الأخرى: أشار إليها Montgomery and Peck (1982) وهي الاستعانة بمخطط أثر الحرف، وهنا تظهر سيطرة متغيرات ذات قيم عالية المعالم المقدرة على المخطط، مما يؤدي إلى عدم إعطاء صورة واضحة عن مسار المنحنيات الأخرى.

للتغلب على القصور في الطريقتين المشار إليهما، نستعين بمقترح إنشاء مجاميع أثر الحرف، ومنها قد يتم اختيار بعض المتغيرات من كل مجموعة حسب الملاحظات التي أشار إليها Hoerl and Kennard (1970b) لاستبعاد المتغير.

ندرج ملاحظة مفادها: عند نجاح انحدار الحرف في التغلب على تعدد العلاقة الخطية معناه أن التحويل الخطي لطريقة المربعات الصغرى الاعتيادية قد أفلح في تقريب البيانات إلى التعامد، وبذلك استوفيت الشروط للسماح بالتعامل بالمثل مع طريقة المربعات الصغرى، من هنا نستنتج بأنه من المعقول جداً أن تتبع طرائق اختيار المتغيرات المستخدمة بالطرائق غير المتحيزة لاختيار متغيرات انحدار الحرف بالاستعانة بالمصفوفة والمتجه الموسعين.

تطبيق

طبقت المقترحات على العديد من الدراسات ذات التخصصات المختلفة. وبالاستعانة بالبرنامج الحاسوبي الجاهزة Minitab 13.12، أتاح أسلوب المحاكاة (Simulation) من تطبيق الطرائق المقترحة ما يقارب المائتي مرة. وقد عول على البرنامجين الحاسوبيين الجاهزين Excel 97 وS^+ 2000 في رسم المنحنيات

وتدوير محور أثر الحرف. وقد ساهمت المصفوفة والمتجه الموسعان في تسهيل العمليات الإحصائية كافة.

أيدت التطبيقات فاعلية الأساليب المقترحة، واختلفت في النتائج الوسطية، إذ ظهر الاختلاف بينها في الآتي: 1- عدد مجاميع القيم المطلقة المتقاربة للمعالم المقدرة، واكتفت معظم التطبيقات بمجموعتين، 2- طول مدى معلمة التحيز، 3- عدد المراحل التحليلية التي اتبعت لاختيار قيمة معلمة التحيز المثالية

ونذكر فيما ياتي مثالاً توضيحياً استخدمت فيه متغيرات تعود لمرضى مصابين بفقر دم البحر الأبيض المتوسط نوع بيتا Beta-Thalassaemia. أخذت البيانات من Awad (1999)، الذي راجع مستشفى أبن الأثير التعليمي للأطفال في مدينة الموصل-العراق، وسجل بيانات عينة عشوائية مكونه من 150 مريضاً تراوحت أعمارهم بين العام والخمسة عشر عاماً من الجنسين. وقد اختيرت متغيرات هذا المرض لسببين: السبب الأول: ملاءمتها لمشكلة بحثنا، إذ إن معظمها ينتمي إلى تكوين الدم (الجوادي، 2000) مما يؤدي إلى ترابطها، والسبب الآخر: الاهتمام الواسع من المؤسسات الصحية محلياً وعالمياً للسيطرة على هذا المرض وتوفر البيانات الموثَّقة عنه (Eleftherion, 2003).

وباستشارة بعض الأطباء ذوي التخصص تم اختيار عشرة متغيرات يعتقد بأن لها تأثيراً في المتغير المعتمد. وقد استخدمت هذه المتغيرات في أسلوب المحاكاة لإعادة تطبيق الطرائق المقترحة. والمتغيرات هي:

المتغير المعتمد (y): العمر من العظم: وهو العمر غير الحقيقي للمريض الذي يكون أقل من عمر أقرانه الأصحاء، ويعتمد على تحديد العمر من خلال صورة شعاعيه لمراكز العظم الأولية والثانوية لتقييم شكل العظم وكثافته ومنها تستنتج قيمة هذا المتغير مقاسة بالأشهر.

المتغيرات التوضيحية (X_j)، وهي:X_1: العمر الحقيقي (شهر)، X_2: العمر عند ظهور أعراض المرض (شهر)، X_3: تضخم الكبد (سنتيمتر)، X_4: هيموكلوبين الدم، X_5: مكدس الدم، X_6: الخلايا الشبكية، X_7: أرومة حمراء، X_8:

الهيموكلوبين الجيني، X_9: عدد وحدات الدم، X_{10}: العمر عند أول عملية نقل للدم (شهر).

الكشف عن تعدد العلاقة الخطية:

1-اختبار Farrar and Glauber: أعطى قيمة لمربع كآي مقدارها 1120.3103.

2-تضخم تباين العوامل VIF: ظهرت زيادة قيمة هذا المقياس عن 10 في المتغيرات الآتية:

VIF: X_1=41.4000, X_4=15.0000, X_5=11.7000, X_9=37.9000

3-العدد الشرطي CN: أكبر جذر مميز أعطته مصفوفة المعلومات (**X'X**) يساوي 3797248 وأصغر جذر مميز يساوي 37 فكان CN=102628.3243.

4-لوحظ ارتفاع قيمة معامل الارتباط بين بعض المتغيرات، فمثلا كان: r_{X1X6}=0.9460 و r_{X4X5}=0.9500، كما أنهما أكبر من قيمة معامل التحديد R^2=0.8380.

5-ابتعدت قيمة محدد مصفوفة الارتباط عن الواحد فكانت: |**R**|=0.000437.

6-ظهور اختلاف في إشارة بعض معامل المتغيرات التوضيحية بين النموذج العام والنموذج الجزئي، وقد كان لها أثر ذو دلالة إحصائية في كلا النموذجين.

إن النتائج كافة أشارت إلى وجود تعدد العلاقة الخطية. وسيُعتَمد في التطبيقات اللاحقة على المقياسين VIF وCN، إذ أن قيمتيهما تقارنان مع عدد ثابت، ويتم الحصول على قيمة VIF مباشرة عند إتباع المسار (12).

تحديد قيمة معلمة التحيز k:

1-طريقة Hoerl et al. (1975)، إن: $\hat{\sigma}^2$ =0.001165، $\hat{\beta}'_{ols}\hat{\beta}_{ols}$ =9.1037، m=10، أعطت المعادلة (10) قيمة k=0.001228، ولم يتم بواسطتها ألتغلب على تعدد العلاقة الخطية، إذ كانت قيم VIF وCN لبعض المتغيرات التوضيحية كالآتي:

VIF>10: X_1=37.4000, X_4=14.2000, X_5=11.3000, X_9=34.3000

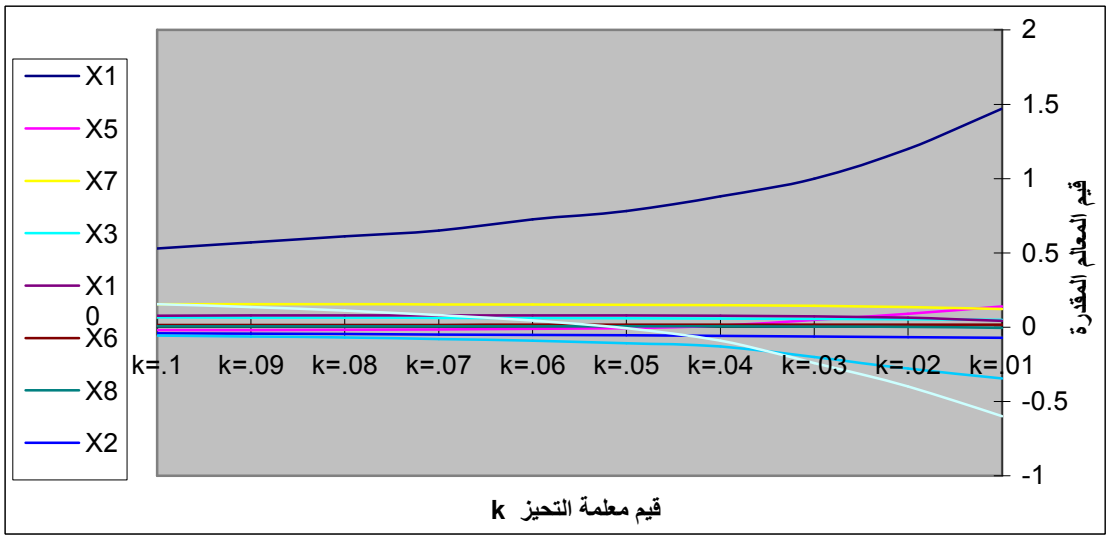
CN=197.0761

كما أعطت المعادلة (11) التي أشار إليها المشهداني (1994) قيمة
k=0.001024، ولم تكن هذه الطريقة أفضل من سابقتها، فقد أعطت النتائج الآتية:
VIF>10: X_1=38.2000, X_4=14.2300, X_5=11.4000, X_9=35.0000
CN=200.9694

2- الطريقة التكرارية لـِ Hoerl and Kennard (1976)، استقرت هذه الطريقة عند الدورة الثانية طبقاً للمقارنة في (13)، حيث كانت:
$(k_1-k_0)/k_0=0.2182<20T^{-1.3}=20(10.7670)^{-1.3}=0.9106$
وقد أعطت الطريقة قيمة لمعلمة التحيز مقدارها: k=0.001559، ولم يتم بواسطتها ألتغلب على تعدد العلاقة الخطية، إذ كانت قيم VIF وCN لبعض المتغيرات التوضيحية كالآتي
VIF>10: X_1=36.7000, X_4=14.0000, X_5=11.2000, X_9=33.6000

CN=193.0369
3- الطرائق المقترحة:

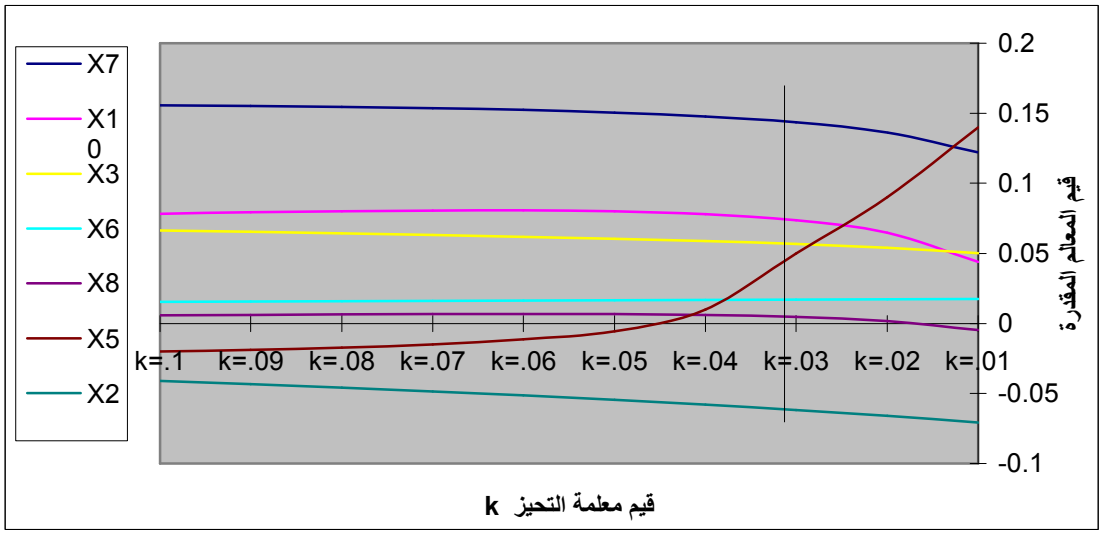
أ- أثر الحرف: يلاحظ من المخطط (1) التأثير السلبي لمنحنيات ذات قيم مطلقة عالية للمعالم المقدرة في وضوح مسار المنحنيات الأخرى، ويشكك ذلك في إمكانية تعيين موقع استقرار المنحنيات ومن ثم يقلل من دقة تعيين قيمة معلمة التحيز واختيار المتغيرات.



المخطط (1): منحنيات أثر الحرف لعشرة متغيرات توضيحية.

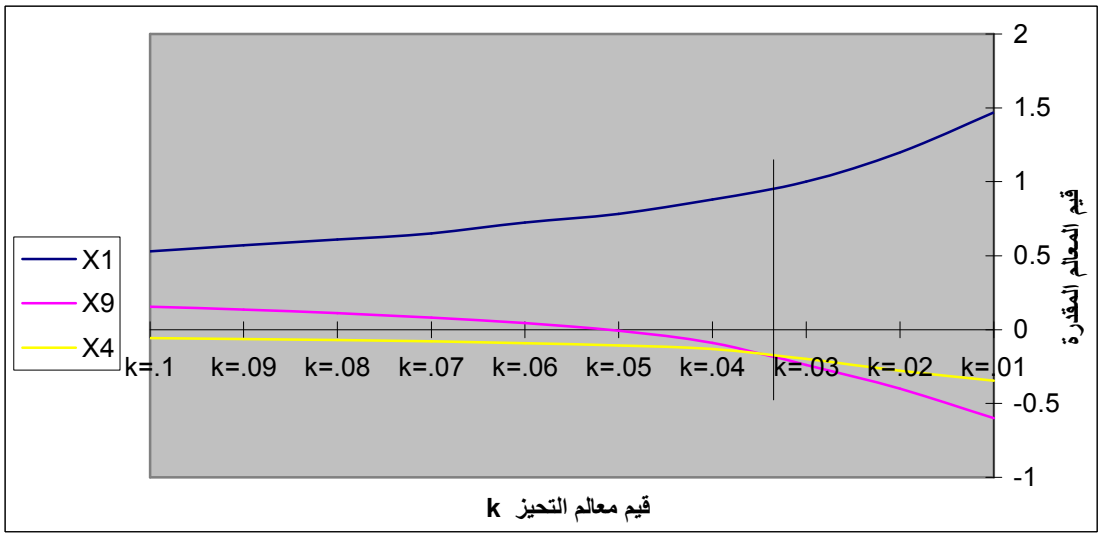
لإعطاء صورة واضحة عن مسار منحنيات أثر الحرف، كُوّنت مجموعتان:

المجموعة الأولى: ضمت سبعة متغيرات، القيم المطلقة لمعالمها المقدرة صغيرة نسبيا، ويوضح المخطط (2) مسار أثر الحرف الذي حدد القيمة 0.045 لمعلمة التحيز.



المخطط (2): أثر الحرف لسبع متغيرات ذات قيم مطلقة صغيرة نسبياً لمعالم مقدرة.

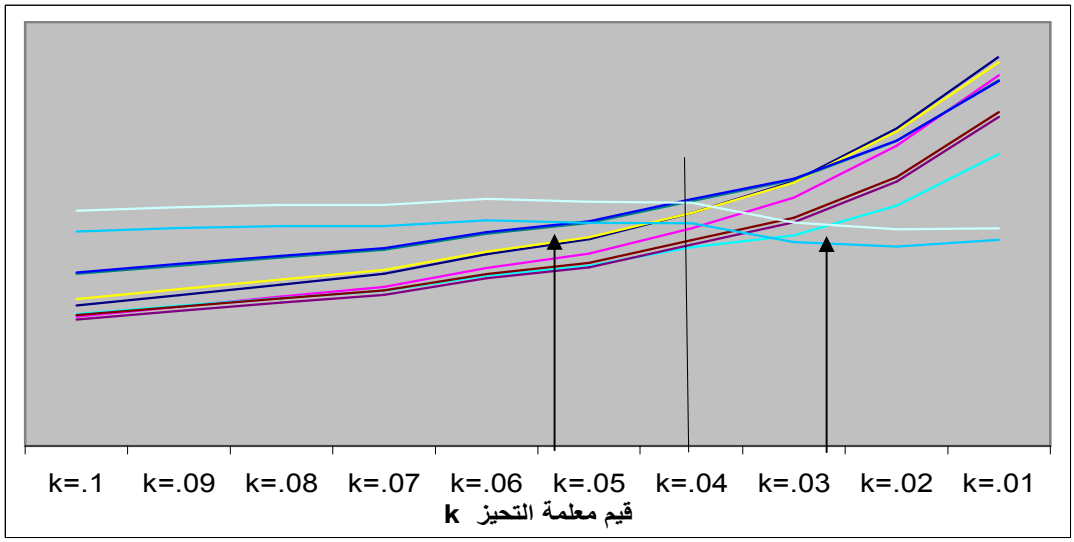
المجموعة الأخرى: احتوت على ثلاثة متغيرات لها قيم مطلقة كبيرة نسبياً لمعالمها المقدرة، ويشير أثر الحرف (المخطط 3) إلى موقع 0.047 لتكون قيمة معلمة التحيز.



المخطط (3): أثر الحرف لثلاثة متغيرات قيم معالمها المقدرة المطلقة كبيرة نسبياً.

من المخططين (2 و3) يعين مدى قيمة معلمة التحيز بين 0.045 و0.047.

ب- تدوير محور أثر الحرف: أثناء عملية التدوير والنظر من الجانب تم الحصول على منظر لتقاطع المنحنيات أعطاه البرنامج الحاسوبي الجاهز Excel 97 (المخطط 4). يشير المخطط (4) إلى مواقع عدة للتقاطع أعطت مدى لقيم معلمة التحيز بين 0.027-0.053، إن مركز التقاطعات يعطي قيمة تقريبية لمعلمة التحيز مقدارها 0.040.



المخطط (4): تدوير محاور أثر الحرف للمتغيرات العشرة.

ت- إن الطريقتين المقترحتين أعطتا مدى متقاربا لمعلمة التحيز، وللتوصل إلى القيمة المثالية، فقد وضع مدى بين 0.040 و0.050، وسيطبق تحليل الانحدار على قيم المدى المحدد مبتدئين بالقيمة 0.040 بزيادة مقدارها 0.001 حتى القيمة 0.050، عند كل قيمة تلاحظ نتائج التحليل لتعيين أصغر قيمة لمعلمة التحيز بواسطتها يتم التغلب على تعدد العلاقة الخطية.

من تحاليل بيانات العينة أمكن التغلب على تعدد العلاقة الخطية عند قيمة k=0.042 وكانت النتائج الآتية:

VIF<10: X_1=9.9000, X_4=6.6000, X_5=6.1000, X_9= 9.2000 .
CN=48.5031<100 .
R^2=0.7430, MSe=0.00173 .
F=43.0600→ P=0.000 .

لوحظ ارتفاع في قيمة MSE وانخفاض في قيمة R^2 كلما زادت قيمة k_p، ويكون التغير طفيفاً عند اقتراب قيمة k_p من القيمة المثالية لها، وعلى هذا الأساس اختيرت قيمة معلمة التحيز لتكون k=0.045 التي أدت إلى ابتعاد قيمة VIF عن الحد الأعلى له مع تغير طفيف في قيم مقاييس المفاضلة، وأعطت النتائج الآتية:

VIF<10: X1=9.5000, X4=6.4000, X5=5.9000, X9=8.8000 .
CN=46.0085<100 .

R2=0.7410, MSe=0.00174 .
F=42.5700→P=0.000 .

اختيار المتغيرات:

يستعان بالمخططين (2 و3) لاختيار المتغيرات لتكون في معادلة انحدار الحرف التقديرية، ففي المخطط (2) يلاحظ استقرار قيم المعالم المقدرة للمتغير X_7 وليستقر عند قيمة تقارب 0.1500، وبالنسبة الى المتغير X_5 فإنه يستبعد لتقاطعه مع الصفر، إن بقية المتغيرات تحقق إحدى ملاحظات Hoerl and Kennard (1970b). من المخطط (3) يتبين أن ميزات منحنى المتغير X_1 تؤهله ليكون في معادلة انحدار الحرف التقديرية، ويستبعد المتغير X_9 لتقاطعه مع الصفر، وليتم اختيار المتغير X_4 مستقراً بقيمة مطلقة تساوي تقريباً 0.3500 .

اتبعت بعض أساليب المربعات الصغرى في اختيار المتغيرات وهي: الاختيار الأمامي والحذف العكسي والخطوات المتسلسلة (الجدول 1)، وقد أشارت جميعها إلى اختيار المتغيرات: X_1 وX_4 وX_7 ، وهذه النتيجة تتفق مع ما تم التوصل إليه عند استخدام طريقة المجاميع المقترحة.

إن للمصفوفة والمتجه الموسعين أهمية كبيرة في تسهيل العمليات الإحصائية والحصول على نتائج تفيد في اتخاذ القرارات، ويوضح الجدولان (1 و2) عدداً من النتائج التحليلية لانحدار الحرف باستخدام المصفوفة والمتجه الموسعين.

الجدول (1): نتائج تطبيق الخطوات المتسلسلة على المصفوفة والمتجه الموسعين.

Stepwise Regression: y versus X_1; X_2; ...;X_{10}
Augmented matrix with k = 0.045

Alpha-to-Enter: 0.01 Alpha-to-Remove: 0.01

Response is y on 10 predictors, with N = 160

Step	1	2	3
Constant	0.0011	-0.0012	-0.0011
X_1	0.8170	0.7630	0.7910
T-Value	19.1000	17.0800	17.7500
P-Value	0.0000	0.0000	0.0000
X_7		0.1480	0.1610
T-Value		3.3000	3.6800
P-Value		0.0010	0.0000
X_4			-0.1260
T-Value			-3.0000
P-Value			0.0030
S	0.0437	0.0424	0.0414
R-Sq	69.7900	71.7500	73.2900
R-Sq(adj)	69.6000	71.3900	72.7800
C-p	17.6000	8.3000	1.5000

الجدول 2: النتائج التحليلية لانحدار الحرف للمتغيرات المختارة.

احتواء معادلة انحدار الحرف التقديرية على نقطة تقاطع β_o.

Predictor	Coef	SE Coef	T	P
Constant	-0.0011	0.0033	-0.3300	0.738
x1	0.7907	0.0445	17.7500	0.000
x4	-0.1265	0.0422	-3.0000	0.003
x7	0.1611	0.0438	3.6800	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	0.7329	0.2443	142.7000	0.000
Residual Error	156	0.2671	0.0017		
Total	159	1.0000			

عدم احتواء معادلة الانحدار على نقطة التقاطع.

Predictor	Coef	SE Coef	T	P
Noconstant				
x1	0.7906	0.0444	17.8000	0.000
x4	-0.1266	0.0420	-3.0100	0.003
x7	0.1609	0.0437	3.6900	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	0.73273	0.24424	143.47	0.000
Residual Error	157	0.26727	0.00170		
Total	160	1.00000			

الاستنتاجات والتوصيات

1- تكوين مجاميع لأثر الحرف حسب التقارب النسبي للقيم المطلقة للمعالم المقدرة أدى لى سهولة اختيار معلمة التحيز ووضوحها والتوصل إلى أفضل المتغيرات لإدخالها إلى معادلة انحدار الحرف التقديرية.

2- بوساطة تدوير محور أثر الحرف تم تحديد مدى تقاطع المنحنيات، مركز المدى يشير إلى قيمة تقريبية لمعلمة التحيز.

3- إعادة تطبيق تحليل الانحدار بقيم مختلفة لمعلمة التحيز قد ساعد في التوصل إلى القيمة المثالية للمعلمة والحصول على أفضل النتائج.

4- نوصي بإتباع إحدى الطريقتين المقترحتين لتحديد مدى لقيمة معلمة التحيز، ومن ثم اتباع الطريقة المقترحة لإيجاد القيمة المثالية للمعلمة. كما نشير إلى أهمية استخدام المصفوفة والمتجه الموسعين لتسهيل العمليات الإحصائية.

المصادر

الجوادي، ولاء عبد الواحد (2000)، " دراسة بعض مكونات وإنزيمات الدم في الأطفال المصابين بالبيتا الثلاسيميا الكبيرة"، رسالة ماجستير، كلية العلوم، جامعة الموصل، العراق.

دبدوب، مروان عبد العزيز (1998)، "تقويم بعض طرق التعرف على تعدد العلاقة الخطية في نماذج الانحدار"، تنمية الرافدين، جامعة الموصل، العراق، **20**، 53، 353-360.

دبدوب، مروان عبد العزيز والكاتب، محمد أسامة (2005)، "تحليل الاتجاه ومشكلة تعدد العلاقة الخطية في تصميم القطع المجزأة"، مؤتة للبحوث والدراسات، جامعة مؤتة، الأردن، **20**، 3، 9-23.

كاظم، أموري هادي ومسلم، باسم شلبية (2002)، القياس الاقتصادي المتقدم النظرية والتطبيق" مطبعة دنيا الأمل، بغداد، العراق.

المشهداني، إيمان محمد (1994)، "استخدام المركبات الرئيسية في تشخيص ومعالجة مشكلة التعدد الخطي مع تطبيق عملي لبعض الظواهر الاقتصادية"، رسالة ماجستير، كلية الإدارة والاقتصاد، جامعة بغداد، العراق.

النعيمي، أسوان محمد (2005)، "اختيار المتغيرات في انحدار الحرف"، رسالة ماجستير، كلية علوم الحاسبات والرياضيات، جامعة الموصل، العراق.

وارطان، هاسميك انترانيك (1989)، "تعدد العلاقة الخطية في أنموذج الانحدار المتعدد"، رسالة ماجستير، كلية الإدارة والاقتصاد، جامعة الموصل، العراق.

Awad, M. H. (1999), Homozygous beta-thalassaemia in Mosul," Ph.D. Thesis, College of Medicine, Univ. of Mosul, Iraq.

Eleftherion, A. (2003), "Thalassaemia international sederation publication 4," World Health Organization, United Nation.

Farrar, D. E. and R. R. Glauber (1967), "Multicollinearity in regression analysis: The problem revisited," Rev. Econ. Statist.,**49**, 92-107.

Geary, R. C. And C. E. Leser (1968), " Significance tests in multiple regression," Amer. Statist., **22**, 1, 20-21.

Hoerl, A. E. and R. W. Kennard (1970a), "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, **12**,55-67.

Hoerl, A. E. and R. W. Kennard (1970b), "Ridge regression: Applications to nonorthogonal problems," Technometrics, **12**, 69-82.

Hoerl, A. E. and R. W. Kennard (1976), " Ridge regression: Iterative estimation of the biasing parameter," Commun. Statist., **A5**, 77-88.

Hoerl, A. E., R. W. Kennard, and K. F. Baldwin (1975), "Ridge regression: Some simulations," Commun. Statist., **4**, 105-123.

Marquardt, D. W. (1970), "Generalized inverses, ridge regression, biased linear estimstion, and nonlinear estimaton, " Technometrics, **12**, 591-
612.

Mason, R. L., R. F. Gunst, and J. T. Webster (1975), "Regression analysis and problems of multicollinearity," Commun. Ststist., **4**, 3, 277-292.

Montgomery, D. C. and E. A. Peck (1982), " Introduction to linear regression analysis," John Wiley and Sons, New York.