مقارنة طريقتي تقدير الدالة اللامعلمية لبيانات عنقودية عن كريات الدم البيضاء لمرضى اللوكيميا

أ.م.د. سجى محمد حسين / كلية الادارة والاقتصاد / جامعة بغداد
م.م. حلا كاظم عبيد /كلية الادارة والاقتصاد / الجامعة العراقية

تاريخ التقديم : 2016/8/16
تاريخ القبول : 2016/11/9

المستخلص:

البيانات العنقودية تظهر في الكثير من العلوم الاجتماعية والصحية والسلوكية. ويتميز هذا النوع من البيانات بوجود الارتباط بين مشاهداتها . وممكن التعبير عن العنقدة من حيث العلاقة بين القياسات على الوحدات ضمن المجموعة نفسها . تم في هذا البحث تقدير الدالة اللامعلمية للبيانات العنقودية باستعمال طريقة المقدرات اللبية غير المرتبطة ظاهريا The Seemingly Unrelated Kernel Estimators , وطريقة المربعات الصغرى المعممة لمقدرات الشريحة التمهيدية The Generalized Least Squares Smoothing Spline Estimators وتم تطبيق الطريقتين المذكورتين على بيانات مرضى اللوكيميا وتمت المقارنة بين الطريقتين عن طريق معيار المقارنة MSE و MAE واوضحت النتائج التطبيقية كفاءة مقدر المربعات الصغرى المعممة لمقدرات الشريحة التمهيدية.

المصطلحات الرئيسية للبحث/ **البيانات العنقودية ، الطريقة المقدرات اللبية غير المرتبطة ظاهريا، وطريقة المربعات الصغرى المعممة لمقدرات الشريحة التمهيدية، MAE،MSE**

مجلة العلوم الاقتصادية والإدارية
العدد 97 المجلد 23
الصفحات 394-419

*البحث مستل من اطروحة دكتوراه

المقدمة:

يطلق على مجموعة من القياسات المشتركة من المعلومات او الاشياء ضمن عنقود واحد (cluster) التي تمت هيكلتها في عناقيد بالبيانات العنقودية .حيث ان مجموعة من الاشياء المشتركة تسمى عنقود والبيانات ضمن العنقود الواحد تكون متماثلة (مرتبطة) اكثر مع بعضها البعض من تلك الموجودة في العناقيد الاخرى[3]. استعملت البيانات العنقودية في الكثير من مجالات الطب الحيوي وعلم الاوبئة والظواهر الزراعية والاقتصادية وغيرها ، ومما تجدر الاشارة اليه ان البيانات العنقودية تشبه البيانات الطولية من حيث اعتمادية البيانات داخل العنقود الواحد (القطاع بالنسبة للبيانات الطولية) على بعضها بعضا [5]. ومن هنا تاتي اهمية الارتباط الموجود داخل المشاهدات لتقدير الدوال حيث ان استخدام الاسلوب المعلمي لايكون دائما مرغوب فية لتقدير بعض الدوال بسبب ان شكل الدالة الملائم للبيانات غير معروف مسبقا او نتيجة لوجود بعض المعوقات لذلك يتم استخدام الاسلوب اللامعلمي لتقدير(تمهيد) الدالة اللامعلمية. لقد تطورت البحوث في الاونة الاخيرة حول استخدام الانحدار اللامعلمي عندما تكون الافتراضات المعلمية غير متحققة، والانحدار اللامعلمي يسمح بمرونة اكبر للدوال المعتمدة على المتغيرات الناتجة من البيانات. لقد تطرقت البحوث السابقة الى حالة تقدير البيانات العنقودية بالطرائق اللامعلمية والطرائق الشبه معلمية واعتمدت حالة اهمال الارتباط داخل بيانات العنقود الواحد الخاصية التي تميز البيانات العنقودية بشكل خاص . حيث ان المقدرات اللبية الموضعية حققت كفاءة اكثر باهمال الارتباط داخل العناقيد (حتى وان كان الارتباط في مصلحة الدراسة).

حيث ان مصفوفة الارتباط داخل العنقود تلعب دورا مهما في كفاءة التقدير للدوال غير المعلومة $\theta(x)$ في نموذج الانحدار اللامعلمي

$$y = \theta(x) + \varepsilon \qquad (1)$$

وتنشا البيانات العنقودية في كثير من المجالات ومنها المجالات الطبية ويعد سرطان الدم احدى المجالات التي يمكن ان تعد البيانات الماخوذة من المرضى المصابين به احدى انواع البيانات العنقودية. حيث اعتبر كل شخص (مريض) كعنقود وتؤخذ له قياسات دورية لنسبة كرات الدم البيضاء ونسبة الهيموغلوبين وتعد القياسات المتكررة في حالات (اوقات) مختلفة وحدات المستوى الاول وتتمثل وحدات المستوى الثاني بالاشخاص.

يتميز سرطان الدم بوجود عدد كبير من الخلايا البيضاء الناضجة نسبيا ولكنها تمتلك شكل غير طبيعي فضلا عن ان خلايا الدم البيضاء في العادة تحتاج الى شهور لتنضج ولكن في هذه الحالة خلايا الدم البيضاء تنتج بسرعة عالية مما يؤدي الى وجود عدد من الخلايا البيضاء غير الطبيعية غير المسيطر عليها التي من غير الممكن التكهن بسلوكها. لهذا النوع من البيانات يمكن تقديرها باستعمال النموذج (1).

هدف البحث:

يهدف البحث الى تقدير الدالة اللامعلمية للبيانات العنقودية باستخدام طريقة المقدرات اللبية غير المرتبطة ظاهريا (seemingly unrelated kernel regression) وطريقة المربعات الصغرى المعممة لمقدر الشريحة التمهيديه (generalized least squares smoothing spline estimator) مع الاخذ بنظر العناية الارتباط الموجود داخل العنقود الواحد حيث ان مصفوفة الارتباط داخل العنقود الواحد تلعب دورا مهما في تحليل البيانات العنقودية مع اخذ حالة الارتباط الموجود داخل العناقيد ومقارنتها مع حالة الاستقلالية لبيانات كريات الدم البيضاء لمرض اللوكيميا من خلال معيار المقارنة MSE و MAE.

الاستعراض المرجعي:

في عام 2000 درس كل من LIN,X.&CARROLL,R.J. طريقة الانحدار اللبي لمتعدد الحدود الموضعي مع متغير واحد لبيانات عنقودية باستخدام المعادلات التقديرية على افتراض ان كل عنقود يشمل عدة مشاهدات ولتكن $m<\infty$ حيث يتم قياس المؤشر في حالة غياب خطا القياس للانحدار العشوائي بتجاهل هيكلية الارتباط داخل العنقود واثبت ان المقدر اللبي الموضعي لمتعدد الحدود يكون اكثر كفاءة في حالة تجاهل الارتباط الموجود داخل العنقود الواحد واستخدموا الطريقة المقترحة في تحليل بيانات نقص المناعة البشرية لخلايا ال CD_4 . واكد ان استعمال الطريقة اللبية المحددة لا تسفر بالضرورة عن مقدر جيد عند حساب الارتباط داخل العناقيد.[12]

وفي عام 2001 استخدم كل من الباحثين LIN, X. and CARROLL R.J. النماذج شبه المعلمية الخطية العامة للبيانات العنقودية باستخدام المعادلات التقديرية المعممة GEE وطبقت النتائج في حالة كون عدد المشاهدات لكل عنقود محدودة وعدد العناقيد كبير ومتوسط النتيجة μ يمثل بالشكل $g(\mu)=X^T\beta+\theta(T)$ حيث g(.) دالة ربط و X و T متغيرات و β متجه المعالم غير المعلومة و $\theta(T)$ دالة ممهدة غير معلومة[11]

وفي عام 2002 قام كل من الباحثين Welsh,A.H.,Lin,X.and Carroll,R.J. بدراسة مقدر الشريحة الذي يمتلك وزنا اكثر وتباين اقل من الطريقة اللبية والتباين المحاذي يكون اقل عند حساب الارتباط بشكل صحيح. حيث ان الطريقة اللبية تسلك سلوك مختلف عن طريقة الشرائح التمهيدية في حالة اخذ هيكلية الارتباط الموجود داخل العناقيد في حين ان كلا الطريقتين الشرائحية واللبية تكونان متكافئتين في حالة اهمال الارتباط داخل العناقيد.[17]

وفي عام 2005 قام الباحث مناف يوسف حمود باستعراض بعض الطرائق اللامعلمية وشبه المعلمية مع بعض الطرائق المقترحة لتقدير دالة الكثافة الاحتمالية كذلك استعراض الطرائق الاهم والخاصة بتقدير المعلمة التمهيدية ومقارنة تلك الطرائق من خلال استخدام اسلوب المحاكاة باستعمال توزيعات وحجوم عينات ومستويات تباين مختلفة للمتغير X [1].

وفي عام 2007 قام الباحث عمر عبد المحسن بدراسة امكانية تقدير النموذج التجميعي (Generalized Additive Model) لتحليل النماذج الشبه معلمية فضلا عن التحليل اللامعلمي واقترح تقدير GAM الحصينة عبر دالة وزن معينة للتحليل المعلمي وشبه المعلمي في حالة وجود تلويث في البيانات واوضح استعمال الشرائح التمهيدية بشكل جزئي (partially) (شبه معلمي) افضل من استعمالها بشكل تام (completely)(لامعلمي) عند تحليل GAM.[2]

وفي عام 2008 قام كل من Wang, Y. G. &Zhao, Y. بتحليل البيانات العنقودية لنماذج الانحدار على اساس الرتبة واقترحوا طريقة رتبة wilixon الموزونة لحساب الارتباط داخل العناقيد ولأحجام مختلفة من العناقيد. واجراء دراسة المحاكاة لمقارنة المقدرات المختلفة لعدد من الطرائق حول هيكلية الارتباط بوجود او عدم وجود القيم المتطرفة وان الطرائق المقترحة تظهر أداء جيد في حالة حساب الارتباط الموزون داخل العناقيد فضلا عن الطرائق الحصينة المتضمنة عدم الحساسية للقيم الشاذة وهيكلية الارتباط.[16]

وفي عام 2010 قام كل من Ibrahim ,N.A.& Suliadi بتحليل البيانات العنقودية /الطولية باستخدام المعادلات التقديرية GEE مع ممهدات الشريحة الطبيعية التكعيبية باستخدام المحاكاة مع اخذ حالة الارتباط الموجود داخل العنقود بعين العناية حيث اظهر ان GEE-Smoothing spline لها خصائص افضل من GEE-Local Polynomial Kernel الذي تناولها Lin &Carroll 2000 الذي اهمل الارتباط الموجود داخل العناقيد للحصول على مقدرات كفوءة في حين برهن Ibrahim, N. A. & Suliadi كفاءة المقدرات الناتجة من حيث التحيز والكفاءة والاتساق في حالة الارتباط الحقيقي للبيانات داخل العناقيد.[13]

وفي عام 2013 قام كل من Ma, S., Song, Q. & Wang. L., بتطوير منهجية عامة لاختيار المتغيرات انيا وتقدير العناصر المجهولة للنماذج التجميعية الخطية جزئيا (additive partially linear models) باختصار (APLMS) للبيانات العنقودية

واقترحوا في الخطوة الاولى المربعات الصغرى للحصول على التقدير المعلمي اما المركبات اللامعلمية فيتم تقديرها على اساس متعدد الحدود للشرائح التمهيدية **polynomial spline smoothing** ويعد هذا الاسلوب مرن وسهل التطبيق على ارض الواقع وركزوا على اخذ الجزء الخطي في النموذج ولتحسين الكفاءة يجب على المرء ان يختار هيكلية الارتباط المناسبة للبيانات المدروسة.[7]

الجانب النظري:

تم في هذا البحث عرض بعض طرائق تقدير الدالة اللامعلمية للبيانات العنقودية وكما يأتي:

(1) **طريقة مقدر النواة غير المتصل ظاهريا**: The Seemingly Unrelated Kernel Estimator

نفرض n من العناقيد ب m_i من المشاهدات داخل كل عنقود حيث $i=1,2,..,n$ وان j^{th} من المشاهدات $j=1,2,..,m_i$ في كل عنقود مؤلف من متغيرات الاستجابة Y_{ij} ومتغير مفرد X_{ij} التي تمثل القياسات المتكررة داخل كل عنقود عبر اوقات مختلفة . بالنسبة للبيانات الطولية حيث ان كل عنقود يمثل قطاع، اما بالنسبة للبيانات العائلية فان كل عائلة تمثل عنقود والمشاهدات داخل العنقود تمثل مختلف افراد الاسر من العائلة نفسها كما موضح في النموذج في المعادلة رقم 1 ،فان لأي مصفوفة تباين وتباين مشترك V فان معدل مقدر النواة غير المتصل ظاهريا ويرمز لها بالرمز $\hat{\theta}_k(x)$ ويساوي لأي نقطة وعندما تكون رتبة مقدر النواة غير المتصل ظاهريا $p=0$ فأن

$$\hat{\theta}_k(x)=k_{wh}^{'}(x)\{I+(\tilde{V}^{-1}-\tilde{V}^{d})k_w\}^{-1}\tilde{V}^{-1}Y \qquad (2)$$

حيث ان: k_{wh}: متجه من الرتبة $nm\times 1$ ويساوي:

$$k_{wh}(x)=\{\sum_{i=1}^{n}\sum_{j=1}^{m}k_h(X_{ij}-x)V^{jj}\}^{-1}\{k_h(X_{11}-x),...,k_h(X_{nm}-x)\}^{'}$$

وان الـ $Y=[\underline{y}_1,\underline{y}_2,.....,\underline{y}_n]^{'}$ مصفوفة من الدرجة $nm\times nm$

و $y_i=(y_{1i}\ y_{2i}\ ....\ y_{im})^{'}$ متجه من الرتبة $nm\times 1$

وان k_w عبارة عن مصفوفة $k_w=\{k_{wh}(X_{11})\ ,\ .\ k_{wh}(X_{nm})\}^{'}\ nm\times nm$

وان $\tilde{V}^{d}=diag(V^{d},V^{d},...,V^{d})\ ;V^{d}=diag(V^{-1})=diag(V^{jj})$

$$\tilde{V}=diag(V_1,.....,V_n)$$

ولمتجه $\hat{\theta}(X)$ من الرتبة $nm\times 1$ يحتوي كل المقدرات $\hat{\theta}_k(x)$ لكل نقاط التصميم X

حيث ان :

$$\hat{\theta}_k(X)=\{\hat{\theta}_k(X_{11})....\ \hat{\theta}_k(X_{nm})\}^{'}$$

فان :

$$\hat{\theta}_k(X)=\{I+K_w(\tilde{V}^{-1}-\tilde{V}^{d})\}^{-1}k_w\tilde{V}^{-1}Y \qquad (3)$$

وحيث ان

$$k_{wh}(x)' = \delta_1'\{\tilde{X}(x)' k_{dh}(x)\tilde{V}^d \tilde{X}(x)\}^{-1}\tilde{X}(x)' k_{dh}(x)$$

$$\delta_1 = (1,0,....,0)'$$

وان $\tilde{X}(x)$ مصفوفة من الرتبة $nm \times p+1$ ب $\{(n-1)^{i+j}\}$ من الصفوف

$$\tilde{X}(x) = \begin{bmatrix} 1 & (X_{11}-x) & (X_{12}-x)^2 & ....(X_{1m}-x)^p \\ : & : : & : & : \quad ..... \\ 1 & (X_{n1}-x)...(X_{n2}-x)^2....(X_{nm}-x)^p \end{bmatrix}$$

$$k_{dh}(t) = diag\{(X_{11}-x)/h,...,(X_{nm}-x)/h\}$$

حيث: $k_h(s) = h^{-1}k_h(s/h)$

$k_h(s)$ تمثل دالة الكيرنل وباستخدام دالة ال Gaussian فان:

$$k(.) = (2\pi)^{-0.5} Exp(-\frac{u^2}{2}) \qquad (4)$$

و h يمثل عرض الحزمة (Bandwidth) التي سيتم حسابها لاحقا
ومصفوفة ال V من الدرجة $nm \times nm$ تحسب كالاتي:

$$V = A^{1/2}R(\alpha)A^{1/2} \qquad (5)$$

A تمثل مصفوفة الانحراف المعياري من الدرجة $nm \times nm$:

$$A = \begin{bmatrix} a_{i1}^{1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{im}^{1/2} \end{bmatrix}$$

حيث ان : $a_{i1} = \sqrt{\text{var } y_{i1}}$

وان $R(\alpha)$: تمثل مصفوفة الارتباطات من الدرجة $nm \times nm$ [11].[7]

$$R(\alpha) = \begin{bmatrix} 1 & \rho_{i1} & \cdots & \rho_{imi} \\ & \vdots & \ddots & \vdots \\ \rho_{imi} & & \cdots & 1 \end{bmatrix}$$

و $R(\alpha) = corr(y_{ij}, y_{ij+1}) \qquad i = 1,...,n-1$

$$\hat{\rho}_{ij} = \frac{y_{ij} - m(y_{ij})}{\sqrt{v(y_{ij})}}$$

(2) **طريقة المربعات الصغرى المعممة لمقدر الشريحة التمهيدية**: The Generalized Least Squares Smoothing Spline Estimator

يراد من دراسة الشرائح التمهيدية مطابقة البيانات بشكل جيد من جهة ويكون الممهد على قدر من التمهيد من جهة اخرى ولتقدير المعلمة الممهدة في النموذج (1) باستخدام ممهدات الشريحة من الدرجة p ولأي مصفوفة تباين عاملة V والذي يقلل المعادلة

$$\frac{1}{n}\sum\{Y_i - \theta(X_i)\}'V^{-1}\{Y_i - \theta(X_i)\} + \lambda\int\{\theta^{(p)}(X)\}^2 dX \qquad (6)$$

حيث ان $\theta^{(p)}(X)$ قابلة للاشتقاق p من المرات لدالة الانحدار،وان الحد الثاني من المعادلة (6) يمثل مقدار عدم التمهيد فان مقدر الشريحة الممهد بالصيغة الاتية :

$$\hat{\theta}_s(X) = (\tilde{V}^{-1} + n\lambda\psi)^{-1}\tilde{V}^{-1}Y \qquad (7)$$

λ : معلمة تمهيد التي تتغير بتغير الدوال وليس بتغير العقد وتكون $\lambda > 0$ [9] وان الـ

$Y = [\underline{y}_1, \underline{y}_2, ..., \underline{y}_n]'$ **مصفوفة من الدرجة** $nm \times nm$

و $y_i = (y_{1i}\ y_{2i}\ ....\ y_{im})'$ **متجه من الرتبة** $nm \times 1$

ومصفوفة ال $\tilde{X}(x)$ **من الدرجة** $nm \times p+1$ **[15],[8],[10]**

ولتقليل المعادلة (6) وباستخدام معلمة تمهيد ثابتة λ لجميع الدوال المستمرة والمختلفة تقود الى استخدام الشرائح التكعيبية مع عقد عند نقاط البيانات (اي تستخدم البيانات كعقد عند نقاط التصميم) حيث ان قيمة الـ $p = 2$ في المعادلة (6) [4]
ولإيجاد مصفوفة ال ψ فان:

$$\psi = CQ^{-1}C' \qquad (7)$$

حيث ان C مصفوفة من الدرجة $n \times n-2$

$C_{n\times n-2} = [C_{ij}] \qquad i = 1,....,n;\ j = 2,...,n-2$

$L_i = x_{(i+1)} - x_i \qquad i = 1,....,n-1$

$C_{ij} = 0 \quad if\ |i-j| \geq 2$

$C_{j-1,j} = L_{j-1}^{-1}$

$C_{jj} = -L_{j-1}^{-1} - L_j^{-1}$

ومصفوفة الـ Q من الدرجة $n-2 \times n-2$

$$Q = [q_{ij}] \qquad i, j = 2,..., N-1 \qquad (8)$$

حيث ان :

$$q_{ij} = 0 \qquad if\ |i-j| \geq 2$$

$$q_{i,i+1} = q_{i+1.i} = L_i / 6 \qquad i = 2,..., N-2$$

$$q_{ii} = (L_{i-1} + L_i)/3 \qquad i = 2,..., N-1$$

$$\tilde{V} = diag(V, V, ..., V) \qquad ;V = D^{1/2} R(\alpha) D^{1/2} \qquad (9)$$

$$D = diag\ [\sqrt{\text{var}(y)}\ ]$$

ولنفرض ان نقاط التصميم $x_1, x_2, ..., x_m$ المرتبة تصاعديا تمثل جميع نقاط المشاهدات في العناقيد $x_{ij}\ , i = 1,2,..., n\ ,\ j = 1,2,..., m$ تستخدم كعقد حيث ان $a < x_1 < x_2 < ... < x_m < b$ حيث ان a, b تحدد مدى فترة الزمن . تستعمل جميع نقاط التصميم ويجب ان تكون العقد اصغر من حجم العينة[14]

معيار العبور الشرعي العام (General Cross Validation):

يتم استخدام معيار العبور الشرعي العام لاختيار معلمة التمهيد من خلال ابدال القيمة الموضحة في المعادلة (6) والتي تساوي $\sum_{i=1}^{n} (I - A_{ii})^{-1'} (I - A_{ii})^{-1}$ بالقيمة $\{1 - (nm)^{-1} trace(A)\}^2$ تكون صيغته كالاتي

$$GCV(\eta) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \{Y_{ij} - \hat{\theta}(X_{ij})\}^2}{\{1 - (nm)^{-1} trace(A)\}^2} \qquad (11)$$

ويتم اختيار معلمة التمهيد ال η التي تقابل اصغر قيمة لـ $GCV(\lambda)$ [17]

معيار المقارنة:

يستخدم معيار المقارنة للتعرف على افضل طريقة للتقدير ، من خلال المقارنة باقل متوسط مربعات الخطا او اقل متوسط القيمة المطلقة للأخطاء واللذان تكون صيغة كل منهما بالشكل الاتي:

$$MSE[\hat{\theta}(X)] = mean[\theta(X) - \hat{\theta}(X)]^2 \qquad (12)$$

$$MAE[\hat{\theta}(X)] = mean\left|\theta(X) - \hat{\theta}(X)\ \right| \qquad (13)$$

الجانب التطبيقي:

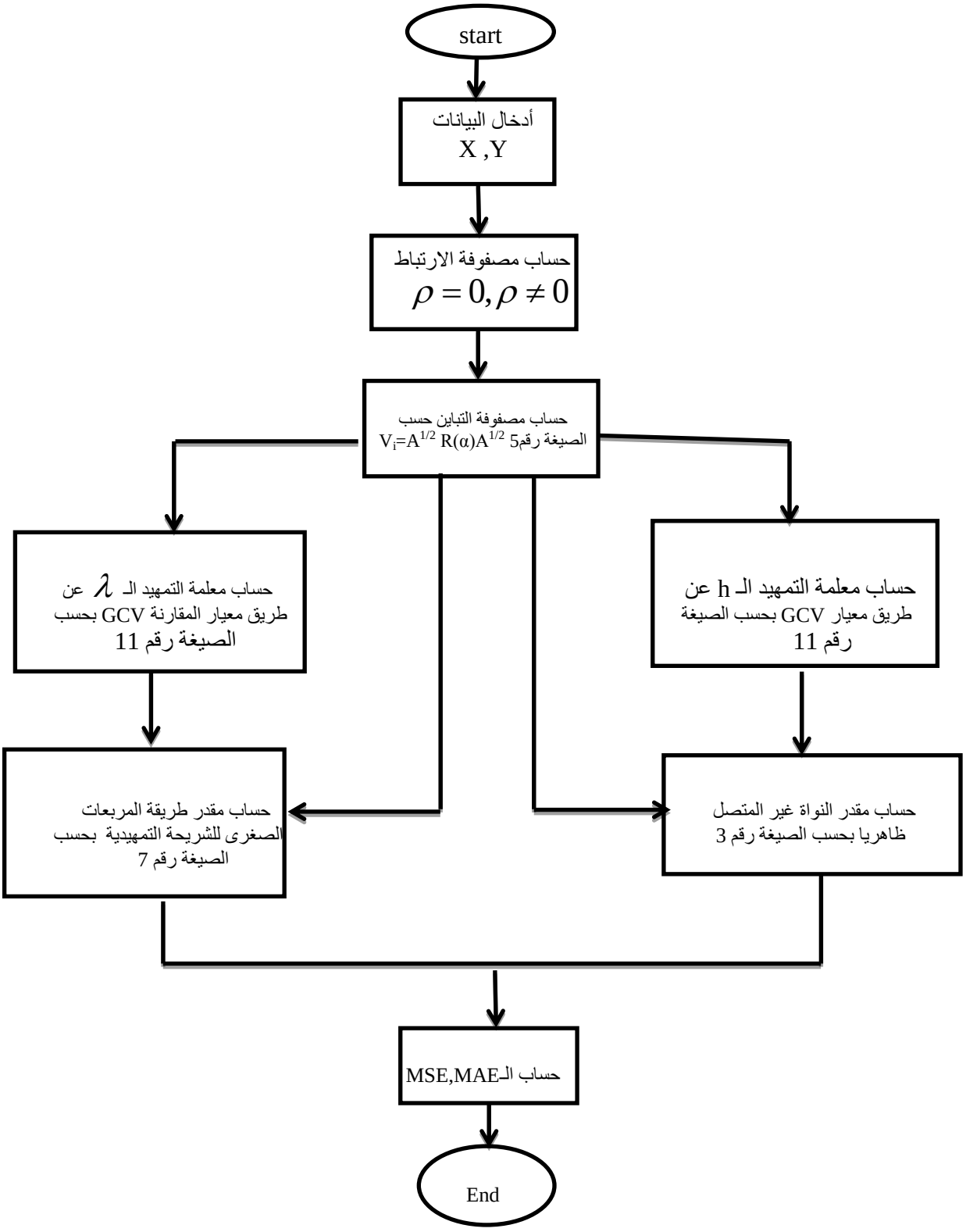
يعرف سرطان الدم (اللوكيميا) بانه عبارة عن تكاثر غير محكوم لخلايا الدم وعادة ما يكون هذا التكاثر لخلايا الدم البيضاء، حيث ان عددها في الدم يزيد عن الحد الطبيعي وهو من 5 الاف الى 10 الاف خلية لكل ملم. ففي سرطان الدم يكون عدد الخلايا من (30-15) الف خلية لكل ملم . وقد يصل هذا التكاثر الى اكثر من 100 الف خلية لكل ملم او حتى يكون اقل من 5 الاف خلية لكل ملم (اي اقل من العدد الطبيعي). ويعود اختلاف هذا العدد الى مرحلة سرطان الدم ومدى تقدمه في الجسم ، فخلايا كرات الدم البيضاء غير طبيعية تنتج في نخاع العظم وهو مصنع ومصدر خلايا كرات الدم جميعها في الجسم فيؤدي تكاثرها في نخاع العظم الى خروجها من الدم . ففي تكاثرها في المراحل الاولى يؤدي ذلك الى تقليل عدد خلايا كرات الدم المصنعة وقبل ان تخرج خلايا كرات الدم البيضاء الى خارج الدم مما يؤدي الى وجود نقص مرحلي في كرات الدم البيضاء في الدم في بداية المرض، ولكن سرعان ما يصدر نخاع العظم هذه الخلايا السرطانية الى الدم ليزيد بذلك عدد كرات الدم البيضاء في الدم. وهذه الزيادة عادة ما تكون مضطردة. ومن اثار هذه الزيادة يصاب المريض بنقص الهيموغلوبين ومن ثم بفقر في الدم فيبدأ يعاني من تعب من اي مجهود يقوم به ودوران وكثرة حب النوم وكسل شديد. ومع استمرار المرض يعاني المريض من نقص في الصفائح الدموية فيصبح دمه يتأخر في التجلط وتكون عدة كدمات على جلده نتيجة عدم قدرة الجسم على وقف التجلطات [6]

تم جمع بيانات عن مرض اللوكيميا لـ 100 شخص واخذت ثلاثة قياسات لكل شخص لثلاث اوقات مختلفة من مستشفى بغداد التعليمي /مدينة الطب، واخذت نسبة التغيرات لعدد كريات الدم البيض كمتغيرات توضيحية ومن ثم اخذت نسبة الهيموغلوبين في الدم كمتغيرات استجابة، ترتبط هذه القياسات فيما بينها مكونه مصفوفة الارتباط التي تلعب دورا مهما في تقدير دالة الانحدار اللامعلمية لنسبة كريات الدم البيضاء حيث تكون مصفوفة الارتباط في حالة الاستقلالية لثلاث أوقات لمتغير الاستجابة بالشكل الاتي:

وفي حالة الارتباط الحقيقي بين العناقيد فان: $R(\alpha)=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$

$$R(\alpha)=\begin{bmatrix}1 & 0.57 & 0.415\\ 0.57 & 1 & 0.584\\ 0.415 & 0.584 & 1\end{bmatrix}$$

مخطط انسيابي يوضح العمل الرياضي



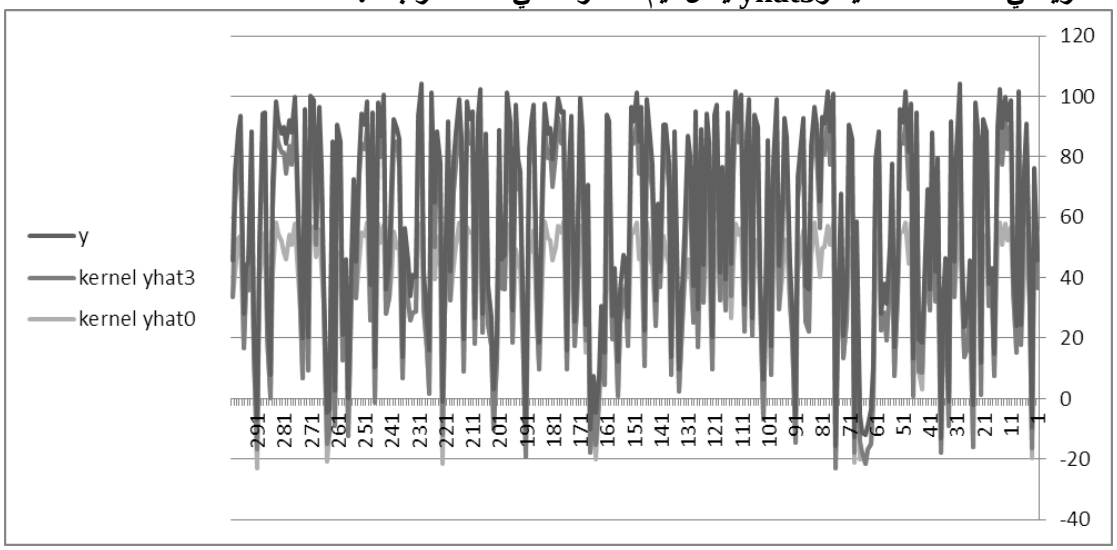
تحليل النتائج

تم تطبيق الطرائق المذكورة في الجانب النظري لبيان افضل طريقة لتقدير الدالة اللامعلمية التي تحقق اقل متوسط مربعات الخطأ واقل قيمة لمتوسط الاخطاء المطلقة باستخدام برنامج بلغة الماتلاب (MATLAB) كما موضح في الملحق رقم (2) والجدول رقم (1) يمثل قيم MAE , MSE لطريقة المقدرات اللبية غير المرتبطة ظاهريا، وطريقة المربعات الصغرى المعممة لمقدرات الشريحة التمهيدية في حالتي الارتباط والاستقلالية ونلاحظ من الجدول ان اقل متوسط مربعات الخطأ واقل قيمه لمتوسط الاخطاء المطلقة كانت لطريقة المربعات الصغرى للشريحة التمهيدية في حالة وجود الارتباط.

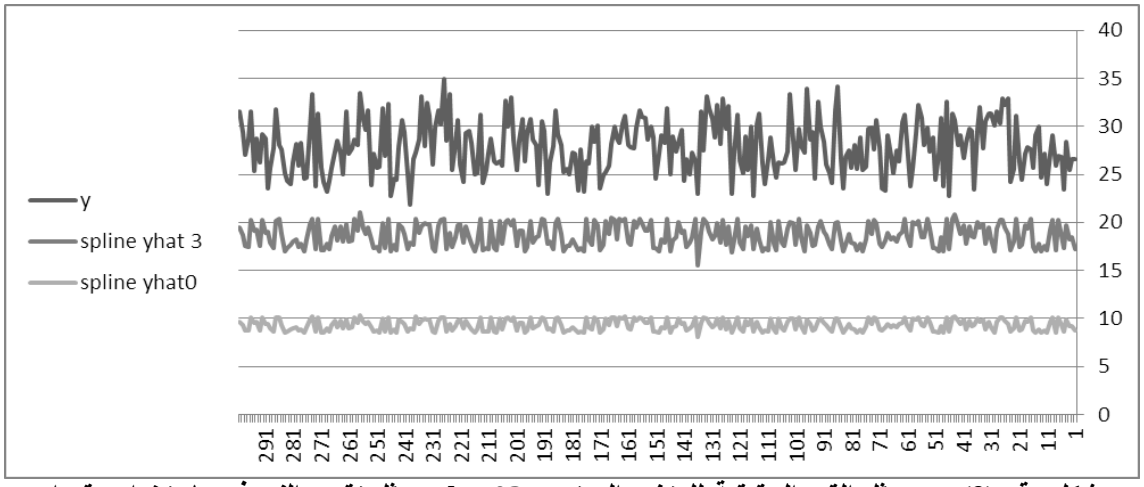
جدول رقم (1) يبين قيم MSE,MAE لحالتي الاستقلالية والارتباط لكل من طريقتي(المقدرات اللبية غير المرتبطة ظاهريا، وطريقة المربعات الصغرى المعممة لمقدرات الشريحة التمهيدية)

Methods	حالة الاستقلالية	حالة الارتباط
Kernel	Mse=(30.60145) Mae=(1156.274)	Mse=(15.21513) Mae=(283.6377)
Spline	Mse=(1.685315) Mae=(4.290639)	Mse=(1.681781) Mae=(4.284745)

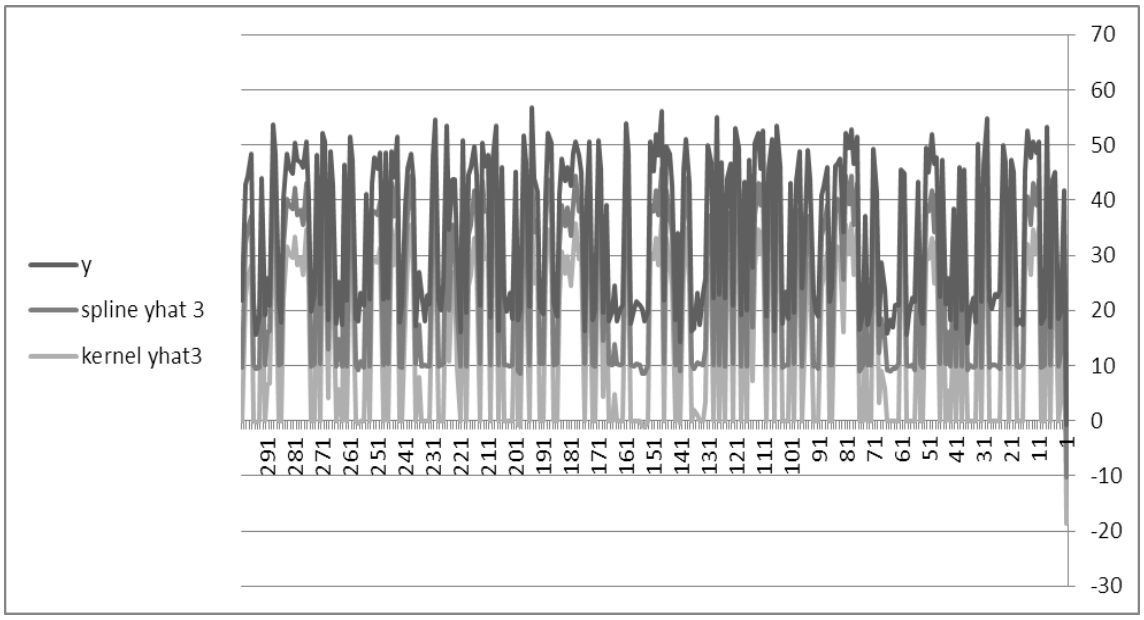
شكل رقم (1): يمثل قيم y الحقيقية وyhat0 يمثل تقدير النموذج باستخدام المقدرات اللبية غير المرتبطة ظاهريا في حالة الاستقلالية وyhat3 يمثل قيم المقدرات في حالة الارتباط .



شكل رقم (2): يمثل قيم y الحقيقية يمثل قيم y الحقيقية وyhat00 يمثل تقدير النموذج باستخدام مقدرات المربعات الصغرى المعممة لممهد الشريحة في حالة الاستقلالية وyhat03 يمثل قيم المقدرات في حالة الارتباط



شكل رقم (3): y يمثل القيم الحقيقية للمتغير المعتمد وyhat03 يمثل تقدير النموذج باستخدام مقدرات المربعات الصغرى المعممة لممهد الشريحة في حالة الارتباط وyhat3 يمثل تقدير النموذج باستخدام قيم المقدرات اللبية غير المرتبطة ظاهريا في حالة الارتباط .



من ملاحظة الاشكال الثلاثة فنرى من الشكل (1) مقدرات طريقة مقدر النواة غير المتصل ظاهريا Seemingly Unrelated Kernel مقارنة مع قيم y الحقيقية حيث تقترب قيم yhat3 في حالة الارتباط من القيم الحقيقية اكثر من قيم المقدر الـ yhat0 في حالة الاستقلالية .

اما في الشكل (2) فنلاحظ بان مقدرات طريقة المربعات الصغرى المعممة لمقدر الشريحة التمهيدية (Generalized least squares smoothing spline) مقارنة مع قيم yالحقيقية حيث تقترب قيم yhat03في حالة الارتباط من القيم الحقيقية اكثر من قيم المقدر الـ yhat00 في حالة الاستقلالية. اما في الشكل (3) فيوضح مقدرات مقدر النواة غير المتصل ظاهريا Seemingly Unrelated Kernel و مقدرات طريقة المربعات الصغرى المعممة لمقدر الشريحة التمهيدية generalized least squares smoothing splineفي حالة ارتباط البيانات مقارنة مع قيم y الاصلية حيث ان مقدرات مقدرات طريقة المربعات الصغرى المعممة لمقدر الشريحة التمهيدية Generalized least squares smoothing spline تقترب اكثر من القيم الحقيقة مقارنة مع قيم مقدرات النواة غير المتصل ظاهريا Seemingly Unrelated Kernel في حالة الارتباط لكلا المقدرين .

الاستنتاجات :

1- ان كل من الطريقتين (طريقة النواة غير المتصل ظاهريا) و(طريقة المربعات الصغرى لممهد الشريحة التمهيدية) حققت اقل MSE واقل MAE عند حساب مصفوفة الارتباط داخل العناقيد بعين العناية بالمقارنة مع حالة استقلال البيانات .

2- حققت طريقة المربعات الصغرى لممهد الشريحة التمهيدية اقل MSE واقل MAE مقارنة مع طريقة ممهد النواة غير المتصل ظاهريا في الحالتين (حالة الارتباط الحقيقي للبيانات)و(حالة الاستقلالية) .

التوصيات :

1- يمكن استخدام الطرائق المذكورة في البحث في حالة الامراض غير المسيطر عليها (النادرة) التي يكون تطورها غير معلوم .

2- يمكن اخذ حالة البيانات العنقودية غير المتزنة وتطبيق الطرائق المذكورة انفا .

المصادر:

المصادر العربيه:

1- حمود د مناف يوسف (2005) ((مقارنة المقدرات اللامعلمية لتقدير دالة الكثافة الاحتمالية)) اطروحة دكتوراه في الاحصاء كلية الادارة والاقتصاد/جامعة بغداد.

2-علي د.عمر عبد المحسن (2007) ((مقارنة مقدرات النماذج التجميعية المعممة باستخدام الشرائح التمهيدية عند تحليل الانحدار اللامعلمي وشبه المعلمي)) اطروحة دكتوراه في الاحصاء كلية الادارة والاقتصاد/ جامعة بغداد.

المصادر الأجنبيه:

3-Galbraith, S. Daniel, J. A .& Vissel ,B.(2010)((A Study of Clustered Data and Approaches to Its Analysis)),The Journal of Neuroscience ,30(32) : 10601-10608.

4-Garrett_Fitzmaurice,_Marie_Davidian,_Geert_ Verbeke & Geert _Molenberghs .(2009) ((Longitudinal Data Analysis)) Chapman& Hall/ CRC.

5-Http://www.jormedic.com/component/mailto/?tmpl=com

6-Https://en.wikipedia.org/wiki/Cluster_analysis

7- Ibrahim, N, A.& Suliadi,2010a ((Analyzing Longitudinal Data Using Gee-Smoothing Spline)) . WSEAS Transactions on Systems and Control

8-Liang ,K.-Y. and Zeger, S.L. (1986) ((Longitudinal data analysis using generalized linear models)).Biometrika,73,13-22.

9- Lin, D. & Ying, Z. (2001) ((semiparametric and nonparametric regression analysis of longitudinal data)). J.Amer. Statist. Assoc . ,96,103-126.
10- Lin, X. Carroll, R. J.(2001) ((Semiparametric Regression For Cluster data) ,Biometrika,88,No.4;1179-1185.
11-Lin,X.&Carroll,R. J.(2000) (Nonparametric function estimation for clustered data when the predictor is measured without/with error) J.Amer. Statist. Assoc.,95,520-534al Data)).
12- Linton, O.B., Mammen, E., Lin, X. and Carroll, R.J.(2003) ((Accounting for correlation in marginal longitudinal nonparametric regression)) 2nd Seattle Symmp. Biostatistics.
13-Ma, S., Song., Q. and Wang,L.,2013,((Simultaneous variable selection and estimation in semi parametric modeling of longitudinal/clustered data))Bernoulli 19,252-274
14-Ruckstuhl,A.,Welsh,A.H.&Carroll,R.J.(2000).((Nonparametric function estimation of the relationship between two repeatedly measured variables)).Statist. Sinica, 10,51-71.
15- Wang, N.(2003)((Marginal nonparametric Kernel regression accounting for within-subject correlation)),Biometrika,90,43-52.
16- Wang, Y. G.& Zhao, Y.(2008)((Weighted rank regression for clustered data analysis)). Biometrics. 64,:39-45.
17- Welsh, A. H., Lin, X.& Carroll, R. J.(2002).((Marginal longitudinal non parametric regression Locality and efficiency of spline and kernel methods)).

ملحق رقم (1)

القيم التالية تمثل قيم المتغيرات التوضيحية x المتمثلة بعدد كريات الدم البيض(wbc) لكل مريض خلال 3 زيارات متتابعة ، اما فيما يخص بيانات متغير الاستجابة y المتمثل بهيموغلوبين الدم (hb)

التسلسل	قيم المتغير التوضيحي x			قيم متغير الاستجابه y		
	الزيارة الاولى	الزيارة الثانية	الزيارة الثالثة	الزيارة الاولى	الزيارة الثانية	الزيارة الثالثة
العنقود الاول	63.7	1.4	1.15	9.38	7.80	7.90
العنقود الثاني	2.9	6.9	10.1	8.1	8.6	8.26
العنقود الثالث	24.3	16.8	18.2	7.27	13.3	13.3
العنقود الرابع	5.97	4.75	8.9	8.69	7.98	17.3
العنقود الخامس	0.8	1.16	7.15	6.1	8.71	12.7
العنقود السادس	3.16	1.82	1.57	8.15	8.25	8.18
العنقود السابع	8.84	8.08	7.8	6.76	6.1	6.3
العنقود الثامن	0.320	0.4	o.277	8.8	7.5	9.06
العنقود التاسع	9.50	1.56	0.754	8.82	8.26	9.07
العنقود العاشر	4.6	7.7	8.68	7.9	8.8	8.49
العنقود الحادي عشر	0.393	0.430	0.644	6.86	9.42	9.99
العنقود الثاني عشر	1.24	0.859	0.7	9.6	6.65	8.2
العنقود الثالث عشر	0.2	0.20	0.41	7.6	9.1	7
العنقود الرابع عشر	1.5	2.4	11.3	12.2	13.1	10.8
العنقود الخامس عشر	0.114	56.9	1.56	12.1	10.8	7.3

العنقود السادس عشر	0.94	0.5	0.15	8.3	5.6	8
العنقود السابع عشر	11.7	6.0	1.5	7.2	10.2	10.6
العنقود الثامن عشر	4.8	3.1	2.9	8.5	6.9	11.0
العنقود التاسع عشر	5.7	5.3	5.3	7.5	9.4	9.8
العنقود العشرون	1.3	0.5	1.7	6.8	7.9	6.4
العنقود الواحد والعشرون	0.9	1.7	6.2	8.9	8.6	6.2
العنقود الثاني والعشرون	13.7	11.1	35.7	10.7	10.8	11.0
العنقود الثالث والعشرون	2.1	2.15	2.93	7.5	9.10	9.8
العنقود الرابع والعشرون	0.44	0.027	0.981	7.1	6.1	8.00
العنقود الخامس والعشرون	21.5	5.47	3.74	18.1	12.5	14.4
العنقود السادس والعشرون	5.15	1.43	0.478	12.7	12.0	11.3
العنقود السابع والعشرون	9.04	9.04	13.6	12.7	12.7	14.5
العنقود الثامن والعشرون	9.9	1.85	9.3	10.0	10.2	9.9
العنقود التاسع والعشرون	4.70	7.01	5.82	12.2	12.2	12.0
العنقود الثلاثون	0.148	3.05	0.147	13.1	10.3	12.9
العنقود الواحد والثلاثون	2.5	2.40	0.818	13.1	12.5	8.69
العنقود الاثنان والثلاثون	5.02	4.00	6.08	11.9	12.9	11.3
العنقود الثلاث والثلاثون	1.56	7.01	6.21	12.8	13.2	12.7
العنقود الاربع والثلاثون	17.69	13.3	6.80	7.9	7.11	8.0
العنقود الخمس والثلاثون	4.96	4.07	5.38	12.6	12.5	13.6
العنقود الستة والثلاثون	7.38	68.6	3.85	9.6	7.4	9.63
العنقود السبعة والثلاثون	22.83	11.6	11.9	4.9	6.0	6.85
العنقود الثمانية والثلاثون	2.96	2.77	1.46	11.0	8.89	8.9
العنقود التسعة والثلاثون	5.39	1.9	1.7	10.2	7.1	4.0
العنقود الاربعون	1.50	0.96	0.653	10.6	9.2	7.7
العنقود الواحد والاربعون	6.01	4.19	3.54	6.89	5.19	10.6
العنقود الاثنان والاربعون	3.55	3.65	5.00	10.2	10.8	11.2
العنقود الثلاث والاربعون	5.81	7.21	6.24	8.3	8.25	7.96
العنقود الاربع والاربعون	49.7	2.85	0.26	9.8	8.9	7.4
العنقود الخمس والاربعون	11.3	2.29	1.34	10.9	10.8	6.90
العنقود الستة والاربعون	0.5	0.15	0.2	5.6	8.0	5.7
العنقود السبعة واربعون	13.8	10.7	14.65	18.2	11.5	12
العنقود الثمانية والاربعون	0.129	1.80	0.81	6.75	10.4	9.6
العنقود التسعة والاربعون	2.74	2.40	6.83	12.5	10.8	12.0
العنقود الخمسون	0.151	0.150	0.09	10.2	10.2	8.8
العنقود الواحد والخمسون	0.693	0.708	0.93	7.17	7.25	8.3
العنقود الاثنين والخمسون	0.7	0.8	0.71	11.6	11.5	9.8
العنقود الثالث والخمسون	4.75	8.9	2.40	7.98	9.7	5.68
العنقود الرابع والخمسون	11.7	4.32	5.1	9.5	9.5	12.2
العنقود الخامس والخمسون	2.9	4.27	3.49	9.6	11.7	10.8
العنقود السادس والخمسون	59.8	7.0	4.71	12.4	11.0	11.3
العنقود السابع والخمسون	8.4	13.82	42.9	12.1	11.3	12.5
العنقود الثامن والخمسون	6.2	5.2	4.18	9.5	10.4	9
العنقود التاسع والخمسون	7.6	8.3	10.2	5.5	7.6	8.3
العنقود الستون	1.2	1.3	2.3	6.2	10.2	9.5
العنقود الواحد والستون	3.0	1.9	2.0	9.9	10.1	9.1
العنقود الاثنان والستون	14.1	12.6	17.8	10.8	10.7	11.8
العنقود الثلاث والستون	20.8	8.1	2.1	11.4	10.2	7
العنقود الاربع والستون	21.9	14.3	18.7	7.7	7.9	7.6

العنقود الخمس والستون	26.1	24.4	25.2	9.4	10.5	10.4
العنقود الستة والستون	22.8	15	5.6	6.6	9.7	7.7
العنقود السبعة والستون	24.5	39.4	3.0	8.7	7.6	7.1
العنقود الثمانية والستون	3.8	3.4	0.5	10.1	7.2	7.2
العنقود التسعة والستون	25.9	53.6	61.9	5.2	5.1	5.4
العنقود السبعون	1.0	1.3	0.4	6.1	7.3	7.1
العنقود الواحد والسبعون	1.6	0.3	0.4	10	6.5	8.5
العنقود الاثنان والسبعون	6.8	5.8	11.2	10.7	10.4	11
العنقود الثلاث والسبعون	10.0	8.9	0.5	7.3	8.2	6.6
العنقود الاربع والسبعون	3.59	1.3	11.9	11.0	12.3	13.0
العنقود الخمس والسبعون	8.4	2.5	6.2	9.6	7.9	9.9
العنقود الستة والسبعون	29.3	11	2.5	7.5	6	6.4
العنقود السبعة وسبعون	0.2	0.06	0.2	8.4	6.2	7.5
العنقود الثمانية والسبعون	1.5	0.7	1.5	11.1	10.3	10.5
العنقود التسعة والسبعون	0.3	0.3	0.9	8.5	6.2	8.5
العنقود الثمانون	1.6	1.6	2.4	10.3	9.4	9.9
العنقود الواحد والثمانون	1.7	2.5	2	7.8	9	8.2
العنقود الاثنان والثمانون	3.6	1.2	1.3	8.7	7.4	6.4
العنقود الثلاث والثمانون	1.3	1.1	0.8	9.2	8.1	7
العنقود الاربع والثمانون	0.1	0.07	0.1	6.6	8.3	8.2
العنقود الخمسة والثمانون	2.3	3.4	3.41	8.3	9.3	8.8
العنقود الستة والثمانون	7.7	11	14.1	14.1	8.8	7.7
العنقود السبعة والثمانون	6.9	6.9	8.2	11.7	11.7	11.6
العنقود الثمانية والثمانون	0.6	0.58	0.7	6.9	10.7	10.4
العنقود التسعة والثمانون	1.4	1.91	1.77	7.4	8.3	7.81
العنقود التسعون	3.3	1.5	3.9	7.3	5.2	4.6
العنقود الواحد والتسعون	20.6	38.3	32.7	9.34	9.2	9.8
العنقود الاثنان والتسعون	8.1	10.7	9.6	9.8	10.2	8.9
العنقود الثلاث والتسعون	5.5	3.1	1.1	13	5.3	8.7
العنقود الاربعة والتسعون	1.35	2.75	4.52	6.9	9.7	9.5
العنقود الخمس والتسعون	1.15	1.75	4.16	11.9	10.8	6.2
العنقود الستة والتسعون	3.7	8.6	9.1	9.7	10.6	18.1
العنقود السبعة والتسعون	5.92	1.0	0.94	18.1	11.8	11.3
العنقود الثمانية والتسعون	0.05	0.17	1.08	10.3	9.4	9.5
العنقود التسعة والتسعون	2.8	4.26	3.38	9.5	11.6	10.7
العنقود المنة	11.6	4.21	5.0	9.4	9.4	12.1

ملحق رقم (2)
لحساب طريقه الـ kernel للبيانات الحقيقية

```
clc
clear
data=xlsread('Rdata.xlsx','data');
n=size(data,1);
subj=data(:,1);x=data(:,2);y=data(:,3);
corry0=[1 0 0;0 1 0;0 0 1];
corry1=[1 0.5700 0.415;0.5700 1 0.584;0.415 0.5841 1];
yy1=[y(1:n/3)];
```

```
yy3=[y(((n/3)+1):2*(n/3))];
yy5=[y((2*(n/3)+1):n)];
usubj=unique(subj);
nsubj=length(usubj);
xmin=min(x);
xmax=max(x);
xfit=linspace(xmin*1.05,xmax*.95,101)';
[ux,flag1,flag2]=unique(x);
nux=length(ux);
dpoly=2;
dy1=std(yy1);
dy2=std(yy3);
dy3=std(yy5);
dy=[dy1 dy2 dy3];
sdy=diag(dy);
dsd01=repmat(dy1,n/3);
dsd1=diag((dsd01));
dsd02=repmat(dy2,n/3);
dsd2=diag((dsd02));
dsd03=repmat(dy3,n/3);
dsd3=diag((dsd03));
dsdy=[dsd1'  dsd2'  dsd3'];
dsdy=diag(dsdy);
eyey=eye(n/3);
output=['h' 'gcv' 'df'];
output1=['vh' 'vgcv' 'vdf'];
cdsdy0=kron(eyey,corry0);
V0=dsdy*cdsdy0*dsdy;
Vd10=inv(diag(diag(V0)));
Vd20=diag(diag(inv(V0)));
[yhat00,hgcv00,vhgcv00,B00,gcf01,gcf02]=glpfit(data,V0,Vd10,Vd20,[0,dpoly,1]);
mse00=mean((y-yhat00).^2);
mae00=mean(abs(y-yhat00));
m00=[mse00 mae00];
xlswrite('D:\KHala.xlsx',m00,'m00','A1:B1');
xlswrite('D:\KHala.xlsx',yhat00,'yhat0');
xlswrite('D:\KHala.xlsx',output,'hgcv0','A1:C1');
xlswrite('D:\KHala.xlsx',hgcv00,'hgcv0','A2:C2');
xlswrite('D:\KHala.xlsx',output1,'vhgcv0','A1:C1');
xlswrite('D:\KHala.xlsx',vhgcv00,'vhgcv0','A2:C10');
saveas(gcf01,'figure1','png')
saveas(gcf02,'figure2','png')
```

```
cdsdy1=kron(eyey,corry1);
V1=dsdy*cdsdy1*dsdy;
Vd11=pinv(diag(diag(V1)));
Vd21=diag(diag(pinv(V1)));
[yhat01,hgcv01,vhgcv01,B01,gcf03,gcf04]=glpfit(data,V1,Vd11,Vd21,[0,dpoly,1]);
mse01=mean((y-yhat01).^2);
mae01=mean(abs(y-yhat01));
m01=[mse01 mae01];
xlswrite('D:\KHala.xlsx',m01,'m01','A1:B1');
xlswrite('D:\KHala.xlsx',yhat01,'yhat1');
xlswrite('D:\KHala.xlsx',output,'hgcv1','A1:C1');
xlswrite('D:\KHala.xlsx',hgcv01,'hgcv1','A2:C2');
xlswrite('D:\KHala.xlsx',output1,'vhgcv1','A1:C1');
xlswrite('D:\KHala.xlsx',vhgcv01,'vhgcv1','A2:C10');
saveas(gcf03,'figure3','png')
saveas(gcf04,'figure4','png')
```

لحساب دالة الـkernel للبيانات الحقيقيه وفق البرنامج الاتي

```
function [yhat,hgcv,vhgcv,B,gcf1,gcf2]=glpfit(data,V,Vd1,Vd2,params,xfit,kstr)
n=size(data,1);
subj=data(:,1);
usubj=unique(subj);
nsubj=length(usubj);
x=data(:,2);
y=data(:,3);
[ux,~,flag2]=unique(x);
nux=length(ux);
xrange=max(ux)-min(ux);
if nargin<2||isempty(V),
   V=dsdy*cdsdy*dsdy;
end
V=V(flag2,flag2);
if nargin<3||isempty(Vd1),
   V=sdy*cdsdy*sdy;
   Vd1=inv(diag(diag(V)));
end
V=V(flag2,flag2);
Vd1=Vd1(flag2,flag2);
if nargin<4||isempty(Vd2),
   V=sdy*cdsdy*sdy;
   Vd2=diag(diag(inv(V)));
end
```

```
V=V(flag2,flag2);
Vd2=Vd2(flag2,flag2);
if nargin<5||isempty(params),
   h=0;
   dpoly=1;
   indfig=0;
elseif length(params)==1,
   h=params;
   dpoly=1;
   indfig=0;
elseif length(params)==2,
   h=params(1);
   dpoly=params(2);
   indfig=0;
elseif length(params)==3,
   h=params(1);
   dpoly=params(2);
   indfig=params(3);
end
if nargin<7||isempty(kstr),
  kstr='(1/sqrt(2*pi))*exp(-.5*t.^2)';
end
if h~=0,
   nh=1;
else
   nh=10;
   hmin=.5*(dpoly+1)*xrange/nux;
   hmax=xrange/8;
   vh = logspace(log10(hmin),log10(hmax),nh);
end
if nh>1,
   for ii=1:nh,
      h=vh(ii);
      for jj=1:nux,
         temp=x-ux(jj);
         t=(temp/h)*h^(-1);
         W=diag(eval(kstr));
         X=ones(n,1);
         for r=1:dpoly,
            X=[X,temp.^r];
         end
         AA=0;
         BB=[];
```

```
        for kk=1:nsubj,
           flagi=(subj==usubj(kk));
           Wi=W(flagi,flagi);
           Xi=X(flagi,:);
           yi=y(flagi);
           ni=length(yi);
           Vi=V(flagi,flagi);
           Vd1i=Vd1(flagi,flagi);
          temp=Wi.*diag(diag(pinv(Vi);
         AA=AA+Xi'*temp*Xi;
           BB=[BB,Xi'*Wi];
        end
        temp=pinv(AA)'*BB;
        iii=eye(size(temp));
        temp1=pinv(iii+temp*(Vd1-Vd2))*temp*Vd1;
        A(jj,:)=temp1(1,:);
      end
      B=A(flag2,:);
      yhat=B*y;
      df(ii)=trace(B);
      gcv(ii)=mean((y-yhat).^2)/(1-df(ii)/n)^2;
   end
   vhgcv=[vh(:),gcv(:),df(:)];
   [gcv,temp]=min(gcv);
   h=vh(temp);
   hgcv=vhgcv(temp,:);
end
for jj=1:nux,
   temp=x-ux(jj);
   t=(temp/h)*h^(-1);
   W=diag(eval(kstr));
   X=ones(n,1);
   for r=1:dpoly,
      X=[X,temp.^r];
   end
   AA=0;
   BB=[];
   for kk=1:nsubj,
      flagi=(subj==usubj(kk));
      Wi=W(flagi,flagi);
      Xi=X(flagi,:);
      yi=y(flagi);
      ni=length(yi);
```

```
 Vi=V(flagi,flagi);
      Vd1i=Vd1(flagi,flagi);
     temp=Wi.*pinv(Vi);
      AA=AA+Xi'*temp*Xi;
      BB=[BB,Xi'*Wi];
   end
   temp=pinv(AA)*BB;
   iii=eye(size(temp));
   temp1=pinv(iii+temp*(Vd1-Vd2))*temp*Vd1;
   A(jj,:)=temp1(1,:);
end
efit=A*y;
B=A(flag2,:);
yhat=B*y;
df=trace(B);
dfa=xrange/h/sqrt(2*pi);
resid=y-yhat;
gcv=mean(resid.^2)/(1-df/n)^2;
hsig2=sum(resid.^2)/(n-df);
hgcv=[h,gcv,df];
uyfit=efit;
uysig=sqrt(hsig2*diag(A*A'));
if nh==1,
   vhgcv=hgcv(1:3);
end
if nargin<6||isempty(xfit),
   xfit=ux;
   yfit=uyfit;
   ysig=uysig;
else
   yfit=spline(ux,uyfit,xfit);
   ysig=spline(ux,uysig,xfit);
end
fits=[xfit,yfit,ysig];
if indfig==1,
   figure(1);
   clf;
   subplot(2,2,1);
   plot(x,y,'o',fits(:,1),fits(:,2),'r-',...
      fits(:,1),[fits(:,2)+1.96*fits(:,3),fits(:,2)-1.96*fits(:,3)],'g-.');
   xlabel('x');
   ylabel('y');
```

```
   title(['(a) ',num2str(dpoly),'-th order GLP fit']);
    subplot(2,2,2);
    plot(x,resid/sqrt(hsig2),'o');
    xlabel('x');
    ylabel('standardized residual');
    title('(b) Standardized residuals vs x');
    subplot(2,2,3);
    plot(yhat,resid/sqrt(hsig2),'o');
    xlabel('yhat');
    ylabel('standardized residual');
    title('(c) Standardized residuals vs fits');
    subplot(2,2,4);
    plot(y,resid/sqrt(hsig2),'o');
    xlabel('y');
    ylabel('standardized residual');
    title('(d) Standardized residuals vs responses');
    gcf1=gcf;
    if nh>1,
       figure(2);
       clf;
       subplot(2,2,1);
       plot(log10(vhgcv(:,1)),vhgcv(:,2),'-o');
       xlabel('log_{10}(h)');
       ylabel('GCV');
       title('(a) GCV  vs  log_{10}(h)');
       subplot(2,2,2);
       plot(log10(vhgcv(:,1)),vhgcv(:,3),'-o');
       xlabel('log_{10}(h)');
       ylabel('df');
       title('(b) df  vs  log_{10}(h)');
    end
    gcf2=gcf;
end
clearvars -except yhat hgcv vhgcv B gcf1 gcf2
end
```

لحساب طريقه الـspline

```
clc
clear
data=xlsread('Rdata.xlsx','data');
n=size(data,1);
subj=data(:,1);x=data(:,2);y=data(:,3);
corry0=[1 0 0;0 1 0;0 0 1];
corry3=[1 0.5700 0.415;0.5700 1 0.584;0.415 0.5841 1];
```

```
yy1=[y(1:n/3)];
yy3=[y((n/3)+1:2*(n/3))];
yy5=[y(2*(n/3)+1:n)];
usubj=unique(subj);
nsubj=length(usubj);
xmin=min(x);
xmax=max(x);
xfit=linspace(xmin*1.05,xmax*.95,101)';
[ux,flag1,flag2]=unique(x);
nux=length(ux);
dy1=std(yy1);
dy2=std(yy3);
dy3=std(yy5);
dy=[dy1 dy2 dy3];
sdy=diag(dy);
dsd01=repmat(dy1,100);
dsd1=diag((dsd01));
dsd02=repmat(dy2,100);
dsd2=diag((dsd02));
dsd03=repmat(dy3,100);
dsd3=diag((dsd03));
dsdy=[dsd1'  dsd2'  dsd3'];
dsdy=diag(dsdy);
eyey=eye(n/3);
output=['spar' 'gcv'];
output1=['vspar' 'vgcv'];
cdsdy0=kron(eyey,corry0);
V0=dsdy*cdsdy0*dsdy;
Vd10=diag(diag(inv(V0)));
[yhat00,spargcv00,vspargcv00,A00,gcf01,gcf02]=ssfit(data,[],1,xfit,Vd10);
mse00=mean((y-yhat00).^2);
mae00=mean(abs(y-yhat00));
m00=[mse00 mae00];
xlswrite('D:\SHala.xlsx',m00,'m00','A1:B1');
xlswrite('D:\SHala.xlsx',yhat00,'yhat0');
xlswrite('D:\SHala.xlsx',output,'spargcv0','A1:B1');
xlswrite('D:\SHala.xlsx',spargcv00,'spargcv0','A2:B2');
xlswrite('D:\SHala.xlsx',output1,'vspargcv0','A1:B1');
xlswrite('D:\SHala.xlsx',vspargcv00,'vspargcv0','A2:B10');
saveas(gcf01,'figure1','png')
saveas(gcf02,'figure2','png')
cdsdy1=kron(eyey,corry1);
V1=dsdy*cdsdy1*dsdy;
```

```
Vd11=diag(diag(inv(V1)));
[yhat01,spargcv01,vspargcv01,A01,gcf03,gcf04]=ssfit(data,[],1,x,Vd11);
mse01=mean((y-yhat01).^2);
mae01=mean(abs(y-yhat01));
m01=[mse01 mae01];
xlswrite('D:\SHala.xlsx',m01,'m01','A1:B1');
xlswrite('D:\SHala.xlsx',yhat01,'yhat1');
xlswrite('D:\SHala.xlsx',output,'spargcv1','A1:B1');
xlswrite('D:\SHala.xlsx',spargcv01,'spargcv1','A2:B2');
xlswrite('D:\SHala.xlsx',output1,'vspargcv1','A1:B1');
xlswrite('D:\SHala.xlsx',vspargcv01,'vspargcv1','A2:B10');
saveas(gcf03,'figure3','png')
saveas(gcf04,'figure4','png')
```

لحساب دالة السبلاين للبيانات الحقيقية

```
function [yhat,spargcv,vspargcv,A,gcf1,gcf2]=ssfit (data, spar, indfig, xfit, V)
subj=data(:,1);
usubj=unique(subj);
nsubj=length(usubj);
x=data(:,2);
y=data(:,3);
n=length(x);
[knots,~,flag2]=unique(x);
nknot=length(knots);
E=zeros(n,nknot);
if nargin<3||isempty(indfig),
   indfig=0;
end
if nargin<5||isempty(V),
   V=eye(n);
end
for i=1:n,
   E(i,flag2(i))=1;
end
S=E'*E;
S=diag(diag(S).^(-1/2));
Gmat=splmat(knots);
[~,D]=eig(S*Gmat*S);
d=diag(real(D));
if nargin<2||isempty(spar),
   nspar=15;
```

```
 d0=sort(d);
   dmin=d0(3);
   dmax=max(d);
   sparmax=((nknot-2)/(nknot-4)-1)/dmin;
   sparmin=((nknot-2)/(4-2)-1)/dmax;
   vspar = logspace(log10 (sparmin),log10 (sparmax), nspar)';
else
   vspar=spar;
   nspar=1;
end
for k=1:nspar,
   A=E*pinv(E'*V*E+(n)*vspar(k)*Gmat)*E'*V;
   yhat=A*y;
   gcv(k)=mean((y-yhat).^2)/(1-trace(A)/n)^2;
end
vspargcv=[vspar(:),gcv(:)];
[gcv,temp]=min(gcv);
spar=vspar(temp(1));
spargcv=[spar,gcv];
A=E*pinv(E'*V*E+(n)*spar*Gmat)*E'*V;
yhat=A*y;
resid=y-yhat;
SSE=sum((y-yhat).^2);
hsig2=SSE/(n-trace(A));
ysig=sqrt(hsig2*diag(A*A'));
if nargin<4||isempty(xfit),
   fits=[x,yhat,ysig];
else
   [xx,xflag]=unique(x);
   yyhat=yhat(xflag);
   yysig=ysig(xflag);
   yfit=spline(xx,yyhat,xfit);
   ysig=spline(xx,yysig,xfit);
   fits=[xfit,yfit,ysig];
end
if indfig==1,
   figure(1);
   clf;
   [~,temp2]=sort(fits(:,1));
   ffits=fits(temp2,:);
   plot(x,y,'o',ffits(:,1),ffits(:,2),'r-',...
      ffits(:,1),[ffits(:,2)-1.96*ffits(:,3),ffits(:,2)+1.96*ffits(:,3)],'g-.')
   gcf1=gcf;
   if nspar>1,
```

```
        figure(2);
        clf;
        yyhat=y-resid;
        resid=resid/std(resid);
        subplot(2,2,1)
        plot(log10(vspargcv(:,1)),vspargcv(:,2),'-o')
        xlabel('log_{10}(\lambda)')
        ylabel('GCV score')
        title('(b) GCV curve')
        subplot(2,2,2)
        plot(yyhat,resid,'o')
        xlabel('SS fit')
        ylabel('Standardized residual')
        title('(a)')
        subplot(2,2,3)
        plot(y,resid,'o')
        xlabel('Response')
        ylabel('Standardized residual')
        title('(c)')
        subplot(2,2,4)
        hist(resid)
        xlabel('Standardized residual')
        ylabel('Frequency')
        title('(d)')
    end
    gcf2=gcf;
    end
clearvars -except yhat spargcv vspargcv A gcf1 gcf2
    end
```

برنامج لحساب مصفوفة الـ بساي في دالة السبلاين

```
function [K,Q,R,h]=splmat(t)
[n,m]=size(t);
if n==1,n=m;
    t=t';
  end
h=t(2:n)-t(1:n-1);
hinv=h.^(-1);
Q=diag([hinv(1:n-2);0],-1)-diag([0;hinv(1:n-2)+hinv(2:n-1);0],0);
Q=Q+diag([0;hinv(2:n-1)],1);
Q=Q(2:n-1,:)';
R=diag(h(2:n-2),-1)/6+diag(h(1:n-2)+h(2:n-1))/3+diag(h(2:n-2),1)/6;
K=Q*inv(R)*Q';
clearvars -except K Q R h
 end
```

Compared of estimating two methods for nonparametric function to cluster data for the white blood cells to leukemia patients

Abstract:

We can notice cluster data in social, health and behavioral sciences, so this type of data have a link between its observations and we can express these clusters through the relationship between measurements on units within the same group.

In this research, I estimate the reliability function of cluster function by using the seemingly unrelated Kernel Estimators method and the Generalized Least Squares Smoothing Spline Estimators method, and I applied these two methods on Leukemia patients and made a comparison between the two methods by using MSE and MAE comparison standard, the empirical results showed the efficiency of the Generalized Least Squares Smoothing Spline Estimators method.

Keywords: **cluster data, the seemingly unrelated Kernel Estimators method, and the Generalized Least Squares Smoothing Spline Estimators method, MSE, MAE.**