مقارنة مقدرات M الحصينة مع مقدرات شرائح التمهيد التكعيبية لأنموذج المعاملات المتغيرة زمنياً للبيانات الطولية المتزنة

أ. د. ظافر حسين رشيد		الباحث علي سيف الدين عبد الحافظ
كلية الادارة والاقتصاد- جامعة بغداد
قسم الاحصاء

المستخلص

في هذا البحث تم مقارنة مقدرات (M) الحصينة لتقنية شرائح التمهيد التكعيبية لتلافي مشكلة الشواذ في البيانات أو تلوث الخطأ مع طريقة التقدير التقليدية لتقنية شرائح التمهيد التكعيبية ، بأستعمال معيارين للمفاضلة بينهما هما (MADE)،(WASE ولمختلف حجوم العينة ومستويات التباين، وذلك لتقدير دوال المعاملات المتغيرة زمنياً للبيانات الطولية المتزنة، والتي تتصف بكون المشاهدات يتم الحصول عليها من n من القطاعات المستقلة كل واحد منها يقاس تكرارياً خلال مجموعة نقاط زمن محددة m ،أذ تكون القياسات المكررة داخل القطاعات مرتبطة على الاغلب ومستقلة بين القطاعات المختلفة.

المصطلحات الرئيسية للبحث/ المعاملات المتغيرة زمنياً، تقدير المرحلتين، تقديرات M الحصينة، تمهيد الشرائح التكعيبية، البيانات الطولية المتزنة.

مجلة العلوم
الاقتصادية والإدارية
المجلد ١٩
العدد 73
الصفحات ٣٩٨-٤١٣

بحث مستل من أطروحة دكتوراه

1. المقدمة :

في الدراسات الطولية المشاهدات غالباً يتم الحصول عليها من n من القطاعات المستقلة كل واحداً منها يقاس تكرارياً خلال مجموعة نقاط زمن محددة ، وغالباً ما يتركز أهتمام هذة الدراسات على تقييم آثار الزمن (t) وكذلك مجموعة المتغيرات المستقلة $r=1,2,...,d$ ، $\chi_r(t)$ على نتيجة المتغير المعتمد $y(t)$ ، نفرض أن (t_{ij}) تمثل الزمن للقياسات j^{th} للقطاع i^{th} ، وأن y_{ij} و χ_{ij} تمثل مشاهدات القطاع i^{th} للمتغير المعتمد والمستقل عندالزمن t_{ij} على التوالي، فأن مجموعة المشاهدات الطولية تعطى كالاتي :-

$$\{(t_{ij}, y_{ij}, \chi_{ij}); i=1,2,....n; j=1,2,...t_i\} .....(1)$$

حيث t_i هي عدد القياسات المتكررة للقطاع i^{th} ،على الرغم من أن القياسات هي مستقلة بين القطاعات المختلفة الاّ أنها على الأغلب تكون مرتبطة داخل كل قطاع.

التحليل الأحصائي مع هكذا نوع من البيانات مهتم بنمذجة منحنى المتوسط لـ $y(t)$ والتأثيرات للمتغيرات المستقلة على $y(t)$ ، وتطوير التقدير وأجراءات الاستدلال ،وتحت أطار النماذج المعلميه مثل النماذج الخطية وغير الخطية ونماذج ذات التأثيرات المختلطة ، درست نظريات وطرائق التقدير للمعالم والاستدلالات بصورة موسعة،وتحت أطار النماذج اللامعلميه مع نقاط زمن تصميم ثابتة ((Hart(1991)) إعتمد طرائق (kernel) لتقدير التوقع $\mathrm{E}(y(t))$ بدون وجود المتغيرات المستقلة ، ولأخذ المتغيرات المستقلة بالحسبان (Zeger and Diggle (1994)) درسا الانموذج شبة المعلمي التالي:

$$Y_{ij} = \mu(t_{ij}) + X'_{ij} B + \varepsilon(t_{ij}).........(2)$$

حيث $B=(B_1, B_2, ..., B_d)'$ هو متجة ثوابت غير معلوم في R^d ، $\mu(t)$ هي دالة ممهدة محددة الى (t) ، $\varepsilon(t)$ عملية عشوائية بمتوسط (0) ، وحققوا أجراءات تكرارية حيث أقترحوا أجراء (backfitting) الذي يقدر في البداية $\mu(t)$ بطريقة (kernel) وثم تكرار التقديرات لـ β و $\mu(t)$ ، وأن الأنموذج في (2) هو أكثر مرونة من النماذج الخطية التقليدية ، ويتطلب لتأثيرات المتغيرات $\beta_r, r=1,2,...,d$ أن تكون ثابتة خلال الزمن ، هذا التقييد يمكن أن لا يكون واقعياً للكثير من الحالات العملية ، بعبارة أخرى حجوم العينة الفعلي في أغلب الدراسات الطولية يمكن أن لايكون حجمة كافي لدعم كامل للأنموذج اللامعلمي العام عندما تكون المتغيرات المستقلة ذات بعد عالي وهو ما يسمى مشكلة البعدية (Curse of Dimensionality) ، لذلك ولأجل تعميم عملي أكثر للأنموذج (2) ،((Hoover et al (1998)) أعتمدوا أنوذج المعاملات المتغيرة التالي:

$$Y(t) = X'(t)\, B(t) + \varepsilon(t).......(3)$$

وأقترحوا صنف متعدد الحدود الموضعي (kernel) لتقدير $B\ (t)$ ، حيث $B(t) = \left(B_1(t), B_2(t), \ldots, B_d(t)\right)'$ هو متجة للدوال الممهدة خلال (t) ، $\varepsilon(t)$ كما معرفة في (2) ولجميع قيم (t) فأن $\varepsilon(t)$ و $X(t)$ مستقلان، وبشكل عام نلاحظ أن الأنموذج (3) هو أنموذج خطي بين $Y(t)$ و $X(t)$ عند كل زمن ثابت (t) ، أجراءات التقدير في ((Hoover et al (1998) طورت الحالة الخاصة لمقدرات (kernel) والتي أعتمدت على عرض حزمة (pandwidth) واحد ومقدرات (spline) والتي أعتمدت على اكثر من معلمة تمهيد، وأعتمدوا على خوارزمية (backfitting) لحل مشكلة تعدد الابعاد وخاصة لتقنية الشرائح التمهيدية ، والتي عانت من بعض المشكلات وهي كثافة الحسابات والجهد البرمجي، وكعلاج للمشكلات السابقة ((Fan and Zhang (2000) أقترحا أجراء المرحلتين (Two Step) كبديل لأجل تقدير $B\ (t)$ في (3) ، أولاً أحتسبوا التقديرات الخام (Raw Estimate) الى دوال المعالم $B\ (t)$ عن طريق مطابقة أنموذج خطي قياسي ، ثانياً مهدوا المقدرات الخام (Smooth Estimate) للحصول على المقدرات التمهيدية لدوال المعاملات بواسطة أستعمال أحدى تقنيات التمهيد المعروفة، واستعملوا شرائح التمهيد كأحدى تلك التقنيات.

أن تقديرات دوال المعاملات $B\ (t)$ بأسلوب المرحلتين يتم الحصول علية بأستعمال طريقة المربعات الصغرى، وكما هو معلوم أن مقدرات الربعات الصغرى تمتلك بعض الخصائص الجيدة، وخاصة عندما الخطأ العشوائي يتبع التوزيع الطبيعي ،ولكن المقدرات المعتمدة على المربعات الصغرى حساسة جداً الى الشواذ في البيانات أو عند تلوث (contamination) الخطأ،لذلك فأن طرائق التقدير الحصينة مطلوبة أكثر، في هذا البحث سيتم الأعتماد على البيانات الطولية المتزنة (عندما تكون عدد القياسات للقطاعات متساوية وهي m) والمصاغة وفق الأنموذج (3) ، لأيجاد تقديرات شرائح التمهيد التكعيبية الحصين لدوال المعاملات بطريقة المرحلتين، وبالأعتماد على أسلوب M الحصين، ومقارنته مع طرائق التقدير التقليدية عن طريق تجارب محاكاة بنسب تلويث مختلفة وحجوم عينة مختلفة ومستويات تباين مختلفة ولأغراض الملاءمة والتعميم، تم عرض كل الصيغ والمعادلات بدلالة d من المتغيرات التوضيحية، على الرغم من إستعمال دراسة المحاكاة لحالة ثنائي المتغيرات (متغيرين فقط).

2. طريقة تقدير المرحلتين (Two-Step Estimation Method) [(1) ، (2) ، (10)]

لنفترض $t_j\ ,\ \ j=1,2,\ \ldots,m$ ، هي نقاط زمن محددة، حيث تم جمع البيانات ، أذ ان m تمثل عدد القياسات المكررة لكل قطاع، لأن هناك عدد من المشاهدات التي جمعت في الزمن t_j ، فمن الممكن لهذا الثابت t_j أستعمال البيانات المجمعة هناك لمطابقة أنموذج (3) والحصول على المقدرات الخام (Raw estimates)

$$b(t_j)\ =\ \left(b_1(t_j)\ ,\ \ldots\ ,\ b_d(t_j)\right)'$$

هذه هي المرحلة الأولى، عادةً المقدرات الخام هي غير ممهدة تحتاج الى تمهيدها للحصول على المقدرات الممهدة الى دوال المعاملات لذلك، في المرحلة الثانية لكل مركبة معطاة $r=1,2,\ \ldots,d$ نطبق تقنية تمهيد الى البيانات $\left\{\left(b_r\ (t_j)\ ,t_j\right)\ ,\ j=1,2,\ \ldots,m\right\}$ ، وإن مرحلة تقديرات التمهيد (Smoothing estimates) هذه حاسمة لأنها تعطي مقدرات تمهيدية لدوال معاملات التمهيد الاساسية، وإضافة الى ذلك فان مرحلة التمهيد ذو بعد واحد (one−dimensional) .

1.2 مرحلة التقديرات الخام (Raw Estimates Step) [8]

ولتوضيح هذه المرحلة نفترض $t_j \ , \ j=1,2,....,m$ هي نقاط زمن محددة لمجموعة البيانات الطولية لكل نقطة زمن t_j ، لنفترض N_j هو مجموعة الفهارس للقطاع الى جميع مشاهدات y_{ij} عند t_j، نجمع كل χ_{ij} و y_{ij} التي تقابل الفهرس للقطاعات في N_j ونشكل مصفوفة التصميم $\widetilde{X}_j$ ومتجه الاستجابة $\widetilde{Y}_j$ بالتتابع، الجدير بالاشارة بان البحث يعتمد على حالة البيانات المتزنة وبذلك فان N_j مجموعة الفهارس للقطاع الى جميع مشاهدات y_{ij} عند t_j ستكون متساوية ولجميع القطاعات. عندئذ فان صيغة أنموذج (3) عندما البيانات تجمع عند الزمن يتبع الأنموذج الخطي الاتي :

$$\widetilde{Y}_i(t_j) = \widetilde{X}_i(t_j)\, B(t_j) + \widetilde{e}_i(t_j) \qquad \left. \begin{aligned} &, i=1,2,...,n \\ &j=1,2,...,m \end{aligned} \right\} \quad .....(4)$$

1.1.2 مقدرات المربعات الصغرى العامة المقبولة مع اخطاء AR(1) [9]

(Feasible GLS Estimation With AR(1) Errors)

لتقدير معالم الأنموذج (4) وتحت افتراض هيكل الارتباطات للأخطاء يتبع AR(1) كالآتي :

$$\widetilde{e}_i(t_j) = \rho\, \widetilde{e}_i(t_{j-1}) + u_{ij} \qquad ....(5)$$

يمكننا تطبيق المربعات الصغرى العامة حيث مقدراتها ستكون كالآتي :

$$b_{GLS}(t_j) = \left(X_i'(t_j)\, \Omega^{-1} X_i(t_j) \right)^{-1} X_i'(t_j)\, \Omega^{-1}\, \widetilde{Y}_i(t_j) \qquad ....(6)$$

ان المشكلة في مقدرات GLS بأنها تفترض مصفوفة التباين المشترك Ω معلومة، بعبارة آخرى إن ρ معلوم وهذا نادراً ماهو معلوم من الناحية العملية.

ولتجنب افتراض GLS علينا ايجاد تقدير متسق لـ $\hat{\Omega}$ أي $\hat{\rho}$ واستعمالة لايجاد تقدير لمعالم الأنموذج (4)،إن هذا المقدر يدعى مقدرات المربعات الصغرى العامة المقبولة (FGLS) والذي يمكن ايجاده بحسب الخطوات الاتية :

a. نجد اولاً مقدرات المربعات الصغرى الاعتيادية الى أنموذج (4) والذي سيكون كالآتي :

$$b_{OLS}(t_j) = \left(\widetilde{X}_i'(t_j)\, \widetilde{X}_i(t_j) \right)^{-1} \widetilde{X}_i'(t_j)\, \widetilde{Y}_i(t_j) \qquad ......(7)$$

ثم نجد الاخطاء باستعمال مقدرات OLS .

إذ إن :

$$\tilde{e}_i(t_j) = \tilde{Y}_i(t_j) - \tilde{X}_i(t_j)\, b_{OLS}(t_j) \qquad \ldots\ldots\ldots(8)$$

b. وتحت افتراض الاخطاء هي عمليات عشوائية مشتركة فان مقدر ρ المشترك يمكن تقديره كالآتي :

$$\hat{\rho} = \frac{\sum_{i=1}^{n}\sum_{j=2}^{m} \tilde{e}_i(t_j)\,\tilde{e}_i(t_{j-1})}{\sum_{i=1}^{n}\sum_{j=1}^{m} \tilde{e}_i^{2}(t_j)} \qquad \ldots\ldots\ldots(9)$$

c. إجراء تحويل للبيانات باستعمال transformation (Prais – Winsten) [3].

$$\overset{*}{Y}_i(t_j) = \begin{bmatrix} \sqrt{1-\hat{\rho}^2}\; \tilde{y}_i(t_1) \\ \tilde{y}_i(t_2) - \hat{\rho}\, y_i(t_1) \\ \tilde{y}_i(t_3) - \hat{\rho}\, y_i(t_2) \\ . \\ . \\ \tilde{y}_i(t_m) - \hat{\rho}\, \tilde{y}_i(t_{m-1}) \end{bmatrix} , \; i = 1,2,\ldots,m \qquad \ldots\ldots\ldots(10)$$

$$\overset{*}{X}_i(t_j) = \begin{bmatrix} \sqrt{1-\hat{\rho}^2}\; \tilde{\chi}_i(t_1) \\ \tilde{\chi}_i(t_2) - \hat{\rho}\, \tilde{\chi}_i(t_1) \\ \tilde{\chi}_i(t_3) - \hat{\rho}\, \tilde{\chi}_i(t_2) \\ . \\ . \\ \tilde{\chi}_i(t_m) - \hat{\rho}\, \tilde{\chi}_i(t_{m-1}) \end{bmatrix} , \; i = 1,2,\ldots,m \qquad \ldots\ldots\ldots(11)$$

d. وبتطبيق المربعات الصغرى الاعتيادية على البيانات المحولة فاننا نحصل على مقدرات FGLS كالآتي :

$$b_{FGLS}(t_j) = \left(\overset{*}{X}'_i(t_j)\, \overset{*}{X}_i(t_j) \right)^{-1} \overset{*}{X}'_i(t_j)\, \overset{*}{Y}_i(t_j) \qquad \ldots\ldots(12)$$

وان مقدرات FGLS هي تقاربياً اكثر كفاءة من مقدرات OLS عندما الاخطاء تتبع هيكل ارتباط AR(1).

2.1.2 مقترح التقديرات الخام الحصينة Robust Raw Estimate

اقترح أسلوب M الحصين اولاً من قبل (Huber) [6]،وتستند الفكرة ببساطة الى تقليل بعض الدوال للأخطاء بدلاً عن مجموع المربعات لها، والمقدر الحصين يحدد عن طريق الاختيار لدالة وزن، وان اسلوب مقدرات M بحاجة الى بعض التوسيع لتطبيقها على البيانات الطولية المتزنة الموصوفة في أنموذج (4) والتي تحتوي على n من القطاعات و m من القياسات المكررة لكل قطاع.

إذ إن اهم ما يميز هذه البيانات هو ارتباطها ضمن القطاع، اي بمعنى ان الأخطاء مرتبطة، وهذا سينافي الافتراض لاسلوب مقدرات M وهو ان تكون الأخطاء غير مرتبطة، ولتلافي هذه المشكلة سيتم الاعتماد على البيانات المحولة في طريقة (Feasible GLS) .

لنفرض ان $\overset{*}{Y}_{ij} = \overset{*}{y}_i(t_j)$ و $\overset{*}{X}_{ij} = \overset{*}{\chi}_i(t_j)$ ومتجه المعالم $\underline{\beta} = \beta(t_j)$ ، ولتوضيح هذا الاسلوب، ان المربعات الصغرى الاعتيادية للبيانات المحولة تخفض مجموع مربعات الخطأ الى اقل ما يمكن كالآتي :

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} \overset{*}{e}_{ij}^{2} = \min \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\overset{*}{y}_{ij} - \overset{*}{\chi}_{ij}\, \beta \right)^2 \qquad \ldots\ldots\ldots(13)$$

إذ إن :

$\overset{*}{y}_{ij}$: المشاهدة j للقطاع i للمتغير المعتمد.

$\overset{*}{\chi}_{ij}$: الصف ij لمصفوفة التصميم $\overset{*}{\chi}$.

وإن :

$$\overset{*}{X}_{ijk} = \left[\overset{*}{\chi}_{ij1}, \ldots, \overset{*}{\chi}_{ijd} \right] \;,\; k = 1,2,\ldots,d$$

β : متجه معالم ذو بعد $dm*1$.

ان تقديرات M المطورة من قبل (Huber) والتي لها خاصية (Scale Invariant) تعتمد على فكرة ابدال مجموع مربعات الأخطاء $\overset{*}{e}_{ij}^{2}$ بدالة اخرى للأخطاء الهدف منها تقليل المقدار الآتي:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} P\left(\frac{\left(\overset{*}{y}_{ij} - \overset{*}{X}_{ij}\, \beta \right)^2}{\sigma_e^2} \right) \qquad \ldots\ldots\ldots(14)$$

وإن P هي دالة محدبة متماثلة (symmetric convex function) ولتقليل المقدار اعلاه تؤخذ المشتقة بالنسبة الى متجه المعالم وجعلها مساوية الى الصفر كالآتي :

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \overset{*}{X}_{ij}\, \Psi\left(\frac{\left(\overset{*}{Y}_{ij} - \overset{*}{X}_{ij}\, \beta \right)}{\sigma_e^2} \right) = 0 \qquad \ldots\ldots\ldots(15)$$

إذ إن :

Ψ : المشتقة الجزئية الى متجه المعالم β للدالة P و $\Psi = P'$ وبهذا يكون هناك $(d\,m)$ من المعادلات غير الخطية والتي يمكن حلها بعدة طرائق منها طريقة المربعات الصغرى الموزونة تكرارياً (IWLS)، ولإيجاد مقدرات M بالاعتماد على طريقة IWLS يتطلب حساب دالة الوزن وفيها يتم اعادة كتابة الصيغة (١٥) كما يلي :

$$\sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}\ \overset{*}{X}_{ij}\left(\frac{\left(\overset{*}{Y}_{ij}-\overset{*}{X}_{ij}\ \beta\right)}{\sigma_e^2}\right)=0 \qquad \dots\dots\dots(16)$$

وبحل المعادلة اعلاه نحصل على تقديرات أسلوب M الحصين b_μ باستعمال IWLS .

إذ إن :

$$b_M=\left(\overset{*}{X}{}'W\overset{*}{X}\right)^{-1}\overset{*}{X}{}'WY \qquad \dots\dots\dots(17)$$

إذ إن :

b_M : متجه ذو بعد $(dm*1)$

W : مصفوفة اوزان قطرية ببعد $nm*nm$ () تحسب عناصرها كما يأتي :

$$w_{ij}=\frac{\Psi\left(\frac{\left(\overset{*}{y}_{ij}-\overset{*}{X}_{ij}\ \beta\right)}{\hat{\sigma}_e}\right)}{\left(\frac{\left(\overset{*}{y}_{ij}-\overset{*}{X}_{ij}\ \beta\right)}{\hat{\sigma}_e}\right)} \qquad \dots\dots\dots(18)$$

إذ إن :

$$\hat{\sigma}_e=1.483\left[Median\left|\overset{*}{e}_{ij}-Median(\overset{*}{e}_{ij})\right|\right]$$

وتم استعمال دالة P ومشتقتها Ψ التالية :

دالة (Andrews)

$$\Psi\left(\overset{*}{e}_{ij}\right)=\begin{cases}A\sin\left(\overset{*}{e}_{ij}/A\right) & \left|\overset{*}{e}_{ij}\right|\le A\Pi\\ 0 & \left|\overset{*}{e}_{ij}\right|>A\Pi\end{cases}$$

$$A=1.339$$

2.2 مرحلة تحسين او تمهيد المقدرات الخام [1]

(Refining Or Smoothing The Raw Estimates)

وقبل البدء بهذه المرحلة فان مركبات دوال المعاملات المقدرة في المرحلة الاولى نحصل عليها كالآتي :
ولـ $r=1,2,\ ....,d$ نفرض أن :
$b_r(t_j)$ يحتوي على r^{th} من المركبات لتقديرات المرحلة الاولى $b(t_j)$ عندئذ

$$b_r(t) = L_r'\left(\overset{*}{X}'\overset{*}{X}\right)^{-1}\overset{*}{X}'\overset{*}{Y} \qquad ....(19)$$

إذ إن :

L_r : يعرف كمتجه وحدة ذو بعد $(dm*1)$ فيه r^{th} من المدخلات (1) والباقي (0).
ومن المعلوم إن

$$E[b_r(t)] = \beta_r(t) \qquad ....(20)$$

وبذلك سيصبح التمهيد عند كل مركبة لـ (r) ذو بعد واحد، حيث سنستعمل تقنية تمهيد شرائح التمهيد التكعيبية CSS، ولكن للبيانات التالية :

$$(b_r(t_j)\ ,\ t)\ \ ,\ j=1,2,...,m \qquad ....(21)$$

إذ إن :

t_j : هي نقاط زمن التصميم.

$b_r(t_j)$: هي الاستجابة عند نقاط زمن التصميم، مع التأكيد بانها اي تقدير مستخرج من المرحلة الاولى.
ان أنموذج الانحدار اللامعلمي البسيط للبيانات السابقة سيكون كالآتي :

$$b_r(t_j) = f(t_j) + \varepsilon_j \quad ,\ j=1,2,...,m \qquad ....(22)$$

ونحن نريد تقدير الدالة الممهدة $f(t_j)$
إذ إن :

ε_j : هي اخطاء القياسات التي لايمكن شرحها بواسطة دالة الانحدار $f(t_j)$.

رياضياً $f(t)$ هي التوقع الشرطي لـ $t_j = t$ \ $b_r(t_j)$ أي

$$f(t) = E(b_r(t_j)\backslash t_j = t)$$

1.2.2 شرائح التمهيد التكعيبية

(Cubic Smoothing Splines) :

لايجاد مطابقة شرائح التمهيد للأنموذج (22) وبدون فقدان التعميم لو أفترض مدى الفترة لـ f في الأنموذج (22) هي فترة منتهية [a، b] ولبعض الاعداد المنتهية a،b.

فان الجزء غير الممهد (Roughness Penalty) لـ f هو عادة يعرف كتكامل لمربع مشتقة (V) من المرات كالآتي :

$$\int_a^b \left\{ f_h^{(V)} \right\}^2 dt$$

ولبعض $V \geq 1$ ، عندما ممهد شرائح التمهيد لـ f في أنموذج (٢٢) يعرف كتقليل $\hat{f}_\lambda(t)$ لصيغة المربعات الصغرى الجزائية PLS التالية :

$$\sum_{j=1}^{m} \left[b_r(t_j) - f(t_j) \right]^2 + \lambda \int_a^b \left\{ f_{(t)}^{(V)} \right\}^2 dt \qquad \ldots\ldots\ldots(23)$$

وخلال V^{th} من فضاء (Sobolev) المرتب $W_2^V[a,b]$

إذ إن :

$\lambda > 0$ هي معلمة التمهيد .

وان اختيارنا لشرائح التمهيد التكعيبية لتقليل الصيغة (٢٣) هو صعوبة الحصول على التكامل الذي يعرف الجزء غير الممهد فضلاً على وجود طريقة لحسابه في شرائح التمهيد التكعيبية.

وباستخدام نقاط الزمن (t_j) كعقد حيث نفترض ان $T_j \ , \ j = 1,2,\ldots,m$ هي نقاط الزمن المحددة وان

$$a = T_1 < T_2 < \ldots . < T_m = b$$

فان جميع العقد لشرائح التمهيد التكعيبية التي تقلل (٢٣) عندما $V = 2$ تكون كالآتي :

$$h_L = T_{L+1} - T_L \qquad , \ L = 1,2,\ldots,m-1$$

نعرف $A = (a_{LS})$ كمصفوفة ذات بعد $(m*(m-2))$ مع جميع مدخلاتها تكون (0) ما عدا عندما $L = 1, 2, \ldots, m-2$ فان

$$a_{L,L} = h_L^{-1}$$

$$a_{L+1,L} = -\left(h_L^{-1} + h_{L+1}^{-1}\right)$$

$$a_{L+2,L} = -h_{L+1}^{-1}$$

ونعرف $C = (c_{LS})$ كمصفوفة ذات بعد $((m-2)*(m-2))$ مع جميع مدخلاتها (0) ماعدا

$$c_{11} = (h_1 + h_2)/3$$
$$c_{21} = h_2 / 6$$

ولـ $L = 1, 2, ..., m-4$ فأن

$$c_{L,L+1} = h_{L+1}/6$$
$$c_{L+1,L+1} = (h_{L+1} + h_{L+2})/3$$
$$c_{L+2,L+1} = h_{L+2}/6$$

$$c_{m-3,m-2} = h_{m-2}/6$$
$$c_{m-2,m-2} = (h_{m-2} + h_{m-1})/3$$

وأخيراً نعرف G كمصفوفة وهي مصفوفة ذات بعد $(m*m)$ غير الممهد التكعيبية كالآتي :

$$G = AC^{-1}A'$$

نفرض أن $f = (f_1, ..., f_m)'$ إذ إن :

$$f_j = f(T_j) \quad , \ j = 1,2,...,m$$

وبذلك الجزاء غير الممهد يمكن ان نعبر عنه كالآتي :

$$\int_a^b [f''(t)]^2 \, dt = f' G f \qquad ....(24)$$

لاجل ذلك : نحن فقط نشير الى G كمصفوفة غير الممهد (Roughness matrix)، وهذا يعني أن صيغة PLS في (٢3) يمكن كتابتها كالآتي :

$$\|b_r - Wf\|^2 + \lambda \, f' G f$$

إذ إن :

$W=(W_{jj})$: هي مصفوفة حدث ذات بعد $(m*m)$ مع $W_{jj}=1$ إذا $t_j=T_j$ و ٠ في الحالات الاخرى.

وأن $\|a\|^2=\sum_{j=1}^{m}a_j^2$ وتعرف عادة L_2-norm لـ a لذلك التعبير لشرائح التمهيد التكعيبية $\hat{f}_\lambda$ ، يتحقق عند العقد $T_j\ ,\ j=1,2,..,m$ وتكون كالآتي :

$$\hat{f}_\lambda=(W'W+\lambda G)^{-1}\,W'b_r$$

وأن متجه التقديرات عند نقاط زمن التصميم هو

$$\overset{*}{b}_{r,\lambda}=A_\lambda\,b_r \qquad \ldots\ldots\ldots(25)$$

إذ إن :

$$A_\lambda=W(W'W+\lambda G)^{-1}\,W'$$

وعندما جميع نقاط زمن التصميم تستعمل كعقد فان التقدير سيصبح كالآتي :

$$\overset{*}{b}_{r,\lambda}=\hat{f}_\lambda=(I_n+\lambda G)^{-1}b_r \qquad \ldots\ldots\ldots(26)$$

إذ إن :

$$W=I_n$$

2.2.2 أختيار معلمة التمهيد

(Smoothing Parameter Selection)

إن احد أفضل طرائق اختيار معلمة التمهيد وأكثرها انتشاراً هو (GCV) (Generalized Cross – Validation) ، ولاختيار معلمة التمهيد (λ) للممهد الخطي $\hat{f}_\lambda(t)$ ولأنموذج اللامعلمي في الصيغة (٢٢) ، عن طريق تصغير المعيار الآتي :

$$GVC(\lambda)=\frac{m^{-1}\sum_{j=1}^{m}\left[b_{r,j}-\overset{*}{b}_{r,j}\right]^2}{\left\{1-{t_r(A_\lambda)}/{m}\right\}^2}=\frac{m^{-1}\,SSE_h}{\left(1-{df}/{m}\right)^2} \qquad \ldots\ldots\ldots(27)$$

ويتم اختيار معلمة التمهيد (λ) التي تقابل اقل GCV .

3.2.2 مقترح التقديرات الممهدة الحصينة [7] (Robust Smoothing Estimates)

ان طريقة تقدير الأنموذج اللامعلمي في (22) وهي طريقة شرائح التمهيد التكعيبية حساسة إتجاه وجود قيم شاذة وجعلها أكثر صرامة بوجود الشواذ يمكن اجراء الاساليب الحصينة في القسم (2.1.2)، إذ إن الاخطاء للأنموذج اللامعلمي في (22) ستكون كالآتي :

$$e_j = b_r(t_j) - \overset{*}{b}_{r,j} \qquad , \; j = 1,2,\ldots,m$$

لذلك سيكون من الواجب تحصين معيار (λ) GCV كالآتي :

$$GCV_{Rob}(\lambda) = \frac{m^{-1}\sum W_j\left(b_{r,j} - \overset{*}{b}_{r,j}\right)^2}{\left\{1 - {t_r(A_\lambda)}/{m}\right\}^2} \qquad \ldots\ldots\ldots(28)$$

إذ إن :

W_j : هي دالة وزن تحتوي على (m) من العناصر ويمكن حسابها دون الحاجة الى التوسيع لاسلوب M في القسم (2.1.2) .

3. المحاكاة (Simulation)

تم تنفيذ تجارب المحاكاة بإستخدام (n=10) ويمثل عدد القطاعات مع (m=5، m=10 ،m=15) وتمثل القياسات المتكررة لكل قطاع. وبذلك سيكون لدينا ثلاث حجوم للعينات (nm=50) و(nm=100) وأخيراً(*nm=150*) ، وللأنموذج التالي:

$$Y_{ij} = X_{1,i}(t_j)\, B_1(t_j) + X_{2,i}(t_j)\, B_2(t_j) + e_i(t_j) \qquad , i = 1,2,\ldots,n \; ; j = 1,2,\ldots,m$$

إذ إن :

$\beta_r(t_j)\;, r = 1\,,2$: هي دوال معاملات ممهدة.

المتغيران التوضيحيان $X_{1,i}(t_j)$ و $X_{2,i}(t_j)$ يتبعان التوزيع الطبيعي بمتوسط μ وتباين σ^2 ويتم توليدهما باستعمال طريقة (Box – Muller) وبصورة مستقلة لكل واحداً منهما،أما الاخطاء العشوائية فيتم توليدها كالاتي:

١. متجه الأخطاء $e_i(t_j)$ يتبع التوزيع الطبيعي بمتوسط (٠) وتباين σ^2 يتم عن طريق استعمال طريقة (Box – Muller) ، وقد تم تناول ثلاث مستويات للتباين:

* تباين عالي (High Noise)

$$\sigma = \left(\frac{1}{2}\right) * Function\ Range$$

* تباين متوسط (Medium Noise)

$$\sigma = \left(\frac{1}{4}\right) * Function\ Range$$

* تباين واطئ (Low Noise)

$$\sigma = \left(\frac{1}{8}\right) * Function\ Range$$

إذ إن :

σ : هو الانحراف المعياري للخطأ e.

٢. أما التوزيع الآخر للخطأ العشوائي $e_i(t_j)$ فهو التوزيع الملوث ويستعمل في حالة تلوث البيانات بقيم شاذه وبنسب 10% و 20% إذ تم توليد بيانات تتبع توزيع طبيعي بمتوسط (٠) وتباين ($=36\sigma^2$).

أما دوال المعاملات فهي كلأتي :

$$\beta_1(t) = \sin(4\pi t)$$
$$\beta_2(t) = \cos(0.5\pi t)$$

أما المتغير المعتمد فيتم توليدة مباشرةً من خلال أستخدام الأنموذج في دراسة المحاكاة.
ولتقييم إداء طرائق التقدير لمرحلة التقدير الخام ومرحلة التقدير الخام الحصينة وكذلك مرحلة التمهيد ومرحلة التمهيد الحصينة تم استعمال المعايير التالية:

١. متوسط الانحرافات المطلقة للأخطاء (Mean Absolute Deviation Error):

$$MADE = (dm)^{-1} \sum_{j=1}^{m} \sum_{r=1}^{2} \frac{|\beta_r(t_j) - \hat{\beta}_r(t_j)|}{range\ (\beta_r)}$$

٢. متوسط مربعات الخطأ الموزون (Weighted Average Squared Error):

$$WASE = (dm)^{-1} \sum_{j=1}^{m} \sum_{r=1}^{2} \frac{\{\beta_r(t_j) - \hat{\beta}_r(t_j)\}^2}{range^2\ (\beta_r)}$$

إذ إن :

$(range(\beta_r))$ هو المدى الى دالة $\beta_r(t_j)$.

وتم تكرار جميع تجارب المحاكاة $(Replicates = 200)$ مرة لكل تجربة وتم وضع جميع النتائج في الجداول من رقم (1) الى (4).

جدول (١)معايير تقدير Two step لحالة عدم استعمال أسلوب الحصانة في المرحلة الاولى والثانية،ولجميع حجوم العينة ولجميع مستويات التباين

ϵ	method	n	m	WASE			MADE		
				$\sigma = \frac{1}{2}$	$\sigma = \frac{1}{4}$	$\sigma = \frac{1}{8}$	$\sigma = \frac{1}{2}$	$\sigma = \frac{1}{4}$	$\sigma = \frac{1}{8}$
10 %	CSS	10	5	214.22	71.06	26.03	13.07	6.52	5.32
		10	10	47.12	11.12	13.52	4.29	2.31	3.63
		10	15	7.19	5.32	5.60	2.37	1.47	1.60
20 %	CSS	10	5	221.43	97.07	138.10	11.24	7.96	9.22
		10	10	10.74	5.67	19.11	2.89	1.92	3.37
		10	15	2.18	1.95	2.03	1.36	1.01	1.14

جدول (2) معايير تقدير Two Step لحالة استعمال أسلوب الحصانة M في المرحلة الاولى فقط ، ولجميع حجوم العينة ولجميع مستويات التباين

ϵ	method	n	m	Rob	WASE			MADE		
					$\sigma = 1/2$	$\sigma = 1/4$	$\sigma = 1/8$	$\sigma = 1/2$	$\sigma = 1/4$	$\sigma = 1/8$
10 %	CSS	10	5	M	163.12	99.05	26.27	11.24	6.68	15.64
		10	10	M	36.27	11.78	12.88	3.87	2.44	3.38
		10	15	M	2.26	7.66	2.57	1.09	2.08	1.23
20 %	CSS	10	5	M	335.80	187.45	213.42	15.59	11.55	11.89
		10	10	M	52.69	19.68	27.06	5.52	3.57	4.08
		10	15	M	2.16	1.90	2.01	1.33	0.99	1.11

4. تحليل النتائج:

لحالة عدم استعمال أسلوب الحصانة في المرحلة الاولى والثانية ، ومن نتائج جدول (1) ، قيم معياري (MADE)، (WASE سجلا أنخفاضاً ترافق مع زيادة حجم العينة (الزمن) ولكلا حالتي التلوث، هذا وأن المعيارين سجلا زيادة عند مستوى التباين العالي σ=(1/2) والواطئ σ=(1/8) مقارنة مع مستوى التباين المتوسط σ=(1/4) على الاغلب .

ولحالة استعمال أسلوب الحصانة M في المرحلة الاولى فقط، ومن نتائج جدول (2) ، قيم معياري (MADE)، (WASE سجلا أنخفاضاً ترافق مع زيادة حجم العينة (الزمن) ولكلا حالتي التلوث ، وبمقارنة النتائج مع نتائج جدول (١) نلاحظ أرتفاع قيم المعيارين عند تلوث 10% ومستوى تباين σ=(1/4) وبنسبة اكبر عند تلوث 20% اذ سجلا ارتفاعاً عند حجم عينة (m=5،n=10) و(m=10،n=10) ،هذا وأن المعيارين سجلا زيادةً عند مستوى التباين العالي σ=(1/2) والواطئ σ=(1/8) مقارنة مع مستوى التباين المتوسط σ=(1/4) على الاغلب.

ولحالة استعمال أسلوب الحصانة M في المرحلة الثانية فقط، ومن نتائج جدول (3) ، قيم معياري (MADE)، (WASE سجلا أنخفاضاً ترافق مع زيادة حجم العينة ولكلا حالتي التلوث، وبمقارنة النتائج مع نتائج جدول (١) نلاحظ أنخفاض قيم المعيارين عند تلوث 10% و 20% ،هذا وأن المعيارين سجلا زيادةً عند مستوى التباين العالي σ=(1/2) والواطئ σ=(1/8) مقارنة مع مستوى التباين المتوسط σ=(1/4) على الاغلب .

ولحالة استعمال أسلوب الحصانة M في المرحلة الاولى و الثانية معاً ، ومن نتائج جدول (4) ، قيم معياري (MADE)، (WASE سجلا أنخفاضاً ترافق مع زيادة حجم العينة ولكلا حالتي التلوث ، وبمقارنة النتائج مع نتائج جدول (١) نلاحظ أنخفاض قيم المعيارين عند استعمال اسلوب M الحصين في المرحلة الاولى والثانية أظهر تقدماً عند تلوث 10% ، وعند 20% سجلا زيادة عند حجم عينة (m=5،n=10) مقارنة مع الطريقة التقليدية،هذا وأن المعيارين سجلا زيادةً عند مستوى التباين العالي σ=(1/2) والواطئ σ=(1/8) مقارنة مع مستوى التباين المتوسط σ=(1/4) على الاغلب .

ومن خلال متابعة الجداول من (١) الى (٤) فأن أفضل النتائج لطريقة تقدير CSS عند تلوث 10 هي أستعمال أسلوب الحصانة M في المرحلة الثانية فقط وعند تلوث 20% هي استعمال اسلوب الحصانة M في المرحلة الاولى والثانية معاً ، اذ أفرزت أفضل النتائج ولكلا المعيارين، وأظهرا المعيارين MADE و WASE تطابق في تفضيل ومقارنة طرائق التقدير.

5. الأستنتاجات و التوصيات :

١- أستعمال اسلوب M الحصين في المرحلة الاولى فقط أظهر تقدماً عند تلوث 10% على الطريقة التقليدية واخفق عند تلوث 20%.
٢- أستعمال اسلوب M الحصين في المرحلة الثانية فقط أظهر تقدماً عند تلوث 10% و 20% على الطريقة التقليدية.
٣- أستعمال اسلوب M الحصين في المرحلة الأولى و الثانية أظهر تقدماً عند تلوث 10% و 20% على الطريقة التقليدية عدا حالة واحدة عند تلوث 20% وحجم عينة (m=5،n=10) .
٤- بصورة مطلقة أفضل النتائج لطريقة تقدير CSS عند تلوث 10% هي أستعمال أسلوب الحصانة M في المرحلة الثانية فقط وعند تلوث 20% هي استعمال اسلوب الحصانة M في المرحلة الاولى والثانية معاً .
٥- أستعمال أحد المعيارين MADE و WASE يكون كافي في تفضيل ومقارنة طرائق التقدير.
٦- وجوب تقصي حجم العينة (عدد نقاط الزمن) ،مع مستوى التباين (Noise) المسموح بة ، لتأثيرهما على قيم المعايير.
٧- وجوب تحري وأختيار معلمة التمهيد وفق معايير أختيار معالم التمهيد مثل معيار GCV لما لها من أثر على قيم المعايير للمفاضلة بين الطرائق.

المصادر

1- Fan، J. and Zhang، J.، (2000)، "Two – Step Estimation of Functional Linear Models with Applications to Longitudinal Data"، Journal of Royal statistical society، vol. 62، no. 2، pp. 303-322.

2- Fan، J. and Zhang، W.، (2008)، "Statistical Methods with Varying Coefficient Models" Statistics and Interface، vol. 1، pp. 179-195.

3- Greene، W. H.، (2003)، "Econometric Analysis" Fifth Edition، Prentice Hall، New Jersey.

4-Hart،T.D.،(1991)،" kernel regression estimation with time series errors "،J. Roy. Stat. Soc. Ser. B،53،173-187.

5- Hoover، D. R.، Rice، J. A.، Wu، C. O. and Yang، L.، (1998)، " Non Parametric Smoothing Estimates of time - Varying Coefficient Models with Longitudinal Data"، Biometrika، vol. 85، no. 4، pp. 809-822.

6- Huber، P. J.، (1981)، "Robust Statistics"، John Wiley & Sons، New York.

7- Kovac، A.، (2002)، "Robust Nonparametric Regression and Modality"، http://Maths.bris.ac.uk/~Maxak/

8- Senturk، D. and Muller، H. G. (2008)، "Generalized varying coefficient models for longitudinal data"، Biometrika، vol. 95، Iss.3، pp.653-666.

9- Wooldridge، J. M.، (2002)، "Introductory econometrics: A modern approach"، Cambridge، MA: MIT press.

10- Wu، C. O.، Tian، X. and Yu، J.، (2010)، "Non parametric estimation for time-varying transformation models with longitudinal data"، Journal of non parametric statistics، vol. 22، no. 2، pp. 133-147.

11-Zeger،S.L. & Diggle،P.J.،(1994)،" semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters" Biometrics،50،689-699.

Comparison Robust M Estimate With Cubic Smoothing Splines For Time-Varying Coefficient Model For Balance Longitudinal Data

Abstract

In this research، a comparison has been made between the robust estimators of (M) for the Cubic Smoothing Splines technique، to avoid the problem of abnormality in data or contamination of error، and the traditional estimation method of Cubic Smoothing Splines technique by using two criteria of differentiation which are (MADE، WASE) for different sample sizes and disparity levels to estimate the chronologically different coefficients functions for the balanced longitudinal data which are characterized by observations obtained through (n) from the independent subjects، each one of them is measured repeatedly by group of specific time points (m)،since the frequent measurements within the subjects are almost connected and independent among the different subjects.

Keywords/ Time varying coefficient- Two step estimation- Robust M estimation- Cubic Splines smoothing- Balance longitudinal data.