تحديد أفضل أسلوب تمهيدي لتقدير إنموذج انحدار لا معلمي مع تطبيق عملي

أ.م.د. فارس طاهر حسن
جامعة بغداد/ كلية الادارة والاقتصاد
م.م. غياث حميد مجيد
الجامعة المستنصرية/ كلية التربية الأساسية

الملخص

تمثل تقنية الانحدار اللامعلمي طريقة مختلفة في تحليل الانحدار عن الانحدار المعلمي، لكنه بدوره لا يعني أن استخدام طريقة محل طريقة يمنع من استخدام الأخرى. فأساليب وتقنيات الانحدار اللامعلمي يمكن أن تستخدم في تقييم شرعية الإنموذج المعلمي المفترض والعكس صحيح، وربما شكل مطابقة منحنى الانحدار والحاصل عليه من تقنيات الانحدار اللامعلمي يقترح الإنموذج المعلمي المناسب ليستخدم في الدراسات المستقبلية.

تناول البحث استخدام بعض أساليب التمهيد، ممهد النواة الموضعي (LPK) وانحدار الشريحة (SR) والشريحة الجزائية (PS)، وذلك لغرض تقدير انموذج الانحدار اللامعلمي، وبهدف الحصول على ذلك تم الاعتماد على نتائج تجربة المحاكاة والموضحة في الجانب التجريبي، وتمت المقارنة بين الدوال التي تم استخدامها في التمهيد باستخدام معياري (MAE) و(MSE) للوصول إلى أفضل مقدر لتمثيل البيانات التي تم توليدها باستخدامها المحاكاة. والجانب التطبيقي تناول تطبيق تلك الطرق على بيانات حقيقية من سوق العراق للأوراق المالية والخاصة بشركة الخليج للتأمين.

مقدمة

تمثل تقنيتا الانحدار المعلمي والانحدار اللامعلمي طريقتين مختلفتين في تحليل الانحدار، لكن هذا بدوره لا يعني أن استخدام طريقة ما يمنع من استخدام الأخرى. فأساليب وتقنيات الانحدار اللامعلمي يمكن أن تستخدم لغرض تقييم شرعية الإنموذج المعلمي المفترض وبالعكس، وربما شكل مطابقة منحنى الانحدار والحاصل عليه من تقنيات الانحدار اللامعلمي يقترح الإنموذج المعلمي المناسب ليستخدم في الدراسات المستقبلية.

لجانب النظري

الانحدار اللامعلمي Nonparametric Regression

يمثل اسلوب الانحدار اللامعلمي المرحلة النهائية في عملية تحليل البيانات، أو خطوة استكشافية في عمليات النمذجة. يمكن تمثيل انموذج الانحدار بالشكل التالي:[3]

$$Y_i = m(X_i) + \varepsilon_i \,, \quad i = 1,\dots,n \qquad \dots (1)$$

حيث أن m تمثل دالة الانحدار المجهولة و ε_i أخطاء المشاهدة العشوائية والتي بدورها تكون غير مرتبطة وتتوزع توزيعاً طبيعياً بمتوسط حسابي صفر وتباين قدره σ^2. والهدف هنا هو تقدير m التي يشار إليها بدالة الانحدار أو منحنى الانحدار عند النقاط $x_1, \dots,$ x_n.

الهدف من تحليل الانحدار هو عمل تحليل سببي معقول لدالة الانحدار المجهولة m . وعن طريق الحد من أخطاء المشاهدة فإنه يتيح لنا التركيز على التفاصيل المهمة للمتوسط بإعتماد Y على X. أسلوب منحنى التقريب هذا يدعى بالتمهيد (Smoothing).

تمهيد النواة (التمهيد اللبّي) Kernel Smoother [6] [8]

يزودنا ممهد النواة Kernel بطريقة تمكننا من ايجاد هيكل مجموعة من البيانات دون الحاجة إلى ضرورة وجود إنموذج معلمي، بالتالي ونظراً لوجود المعوقات في استخدام انموذجاً معلمياً لغرض تقدير الدالة m الموضحة في الصيغة (1) فإن الدالة m تكون مبهمة، وتكوّن قيوداً قد تعيق أن يكون التقدير لدالة الانحدار صحيحاً، وبالتالي فإن عملية استخدام انموذجاً معلمياً هنا قد يؤدي بنا إلى استنتاجات غير صحيحة في تحليل انموذج الانحدار. [4]

1. ممهد النواة المتعدد الحدود الموضعي [5] [6] [8]

Local Polynomial Kernel Smoother [LPK]

إن فكرة تمهيد النواة المتعدد الحدود الموضعي الرئيسية هي التقريب الموضعي للدالة m باستخدام متعدد حدود من درجة معينة. أساس هذه الفكرة هو توسيع تايلر Taylor Expansion ، والذي ينص بدوره على أن أي دالة تمهيدية يمكن أن تتقارب موضعياً من متعدد الحدود من درجة معينة.

فإذا كانت x_o تمثل نقطة اعتباطية زمنية ثابتة بحيث نقوم بتقدير دالة m الموضحة في الصيغة (1). وعلى افتراض أن يكون لدينا (p+1) مشتقة مستمرة لبعض قيم $p \geq 0$

عند النقطة x_o. وباستخدام توسيع تايلر، يمكننا تقريب الدالة $m(x)$ موضعياً باستخدام متعدد الحدود من الدرجة p، وكما موضح بالصيغة التالية:

$$m(x) \approx m(x_o) + (x - x_o) m^{(1)}(x_o) + \ldots + (x - x_o)^p m^{(p)}(x_o) / p!$$

إذ أن $m^{(r)}(x_o)$ تمثل المشتقة r للدالة $m(x)$ عند النقطة x_o.

لنفترض هنا أن تكون $\beta_r = m^{(r)}(x_o)/r!$ حيث أن $r = 0, 1, \ldots, p$ ، وأن $\hat{\beta}_r , r = 0, 1, \ldots, p$ تمثل القيم التي تقلل معيار المربعات الصغرى الموزونة (WLS) التالي:

$$\sum_{i=1}^{n} \left\{ y_i - \beta_o - \beta_1 (x - x_i) - \ldots - \beta_p (x - x_i)^p \right\}^2 K_h (x - x_i) \quad \ldots (2)$$

حيث أن $K_h(.) = K(./h) / h$ ، والتي يتم الحصول عليها بإعادة احتساب دالة النواة $K(.)$ مع الثابت $h > 0$ ، والذي يدعى بدوره بعرض الحزمة أو معلمة التمهيد. أن عرض الحزمة يستخدم بشكل أساسي في تحديد حجم الجوار الموضعي local neighborhood، والموضح بالصيغة التالية:

$$I_h(x_o) = [x_o - h + x_o + h] \qquad \ldots (3)$$

تقوم دالة النواة $K(.)$ بتحديد كيف أن المشاهدات ضمن الفترة $I_h(x_o)$ قد تساهم في الملائمة عند النقطة x_o. ويمكن الاشارة إلى مقدر المشتقة r ، $m^{(r)}(x_o)$ ، على أنه $\hat{m}_h^{(r)}(x_o)$ وكما موضح بالصيغة التالية:

$$\hat{m}_h^{(r)}(x_o) = r! \, \hat{\beta}_r \quad , \quad r = 0, 1, \ldots, p$$

وبشكل خاص، تم الحصول على مقدر LPK من الدرجة p للدالة $m(x_o)$ وهو $\hat{m}_h(x_o) = \beta_o$. كما ويمكن تمثيل $\hat{m}_h(x_o)$ بشكل مصفوفة كما موضح في أدناه:

$$X = \begin{bmatrix} 1 & (x_1 - x_o) & \ldots & (x_1 - x_o)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - x_o) & \ldots & (x_n - x_o)^p \end{bmatrix}$$

وأن

$$W = \text{diag}\,(K_h(x_1 - x_o), \ldots, K_h(x_n - x_o))$$

تمثل مصفوفة التصميم ومصفوفة الوزن لملائمة مقدر النواة الموضعي المحلي حول x_o. وبالتالي يمكن إعادة كتابة معيار WLS كما موضح في أدناه:

$$(y - X\beta)^T W (y - X\beta) \qquad \ldots (4)$$

حيث أن $y = (y_1, \ldots, y_n)^T$ ، وأن $\beta = (\beta_1, \beta_2, \ldots, \beta_n)^T$. وهذا بدوره يؤدي إلى:

$$\hat{m}_h^{(r)}(x_o) = r!\, e_{r+1}^T\, S_n^{-1}\, T_n Y \quad , \quad r = 0, 1, \ldots, p$$

إذ أن e_{r+1} تشير إلى متجه وحدة من الدرجة (p+1) والتي تكون عناصره في المجال (r+1) مساوية للواحد وأما باقي عناصره الأخرى فتكون أصفاراً، وأن:

$$S_n = X^T W X \quad , \quad T_n = X^T W$$

وهنا سيتم التركيز هنا على ممهد المنحنى، وكما موضح في أدناه:

$$\hat{m}_h(x_o) = e_1^T\, S_n^{-1}\, T_n\, y \qquad \ldots (5)$$

هذا في حالة ما لم يتم مناقشة تقدير المشتقة. فإذا افترضنا أن $\hat{y}_i = \hat{m}_h(x_i)$ تمثل القيمة الملائمة fitted value للدالة $m(x_i)$. وباستخدام الصيغة (5) يتضح أن:

$$\hat{m}_h(x_i) = a(x_i)^T y$$

حيث أن $a(x_i)$ تمثل $e_1^T\, S^{-1}\, T_n$ بعد أن يتم استبدال x_o محل x_i . وبافتراض أن تمثل $\hat{y}_h = [\hat{y}_1, \ldots, \hat{y}_n]$ القيم الملائمة عند جميع نقاط وقت التصميم design time points. وبالتالي يمكننا توضيح قيمة $\hat{y}_h$ كما في الشكل التالي:

$$\hat{y}_h = A_h y \qquad \ldots (6)$$

حيث أن:

$$A_h = \left(a(x_1), \ldots, a(x_n)\right)^T \qquad \ldots (7)$$

وهي ما تعرف بمصفوفة التمهيد لممهد النواة المتعدد الحدود الموضعي LPK. وحيث أن A_h لا تعتمد على متجه الاستجابة y، فأن ممهد النواة المتعدد الحدود الموضعي $\hat{m}_h$ يعرف على أنه ممهداً خطياً.

2. الممهدات الثابتة الموضعية والممهدات الخطية الموضعية [6] [8]

Local Constant and Linear Smoothers

تعد الممهدات الثابتة الموضعية والخطية الموضعية من أبسط أنواع ممهدات LPK وأكثرها شيوعاً واستخداماً. إذ يُعرف الممهد الثابت الموضعي بإسم ممهد Nadaraya-Watson (N-W) والذي تم اقتراحه من قبل كل من (Nadaraya-Watson) في العام (1964). إذ تم استحصال هذا الممهد من ممهد النواة الموضعي LPK وذلك بجعل p=0:

$$\hat{m}_h(x_o) = \frac{\sum_{i=1}^{n} K_h(x_i - x_o)\, y_i}{\sum_{i=1}^{n} K_h(x_i - x_o)} \qquad \dots (8)$$

ومع وجود جوار موضعي local neighborhood بالشكل $I_h(x_o) = [x_o - h\,,\, x_o + h]$ والتي تناسب البيانات مع وجود ثابت. وهذا هو مقلل $\hat{\beta}_o$ لمعيار WLS التالي:

$$\sum_{i=1}^{n} \left(y_i - \beta_o\right)^2 K_h(x_i - x_o)$$

لذا فإن مقدّر N-W يعتبر مقدّر سهل الفهم وبسيط الحساب في نفس الوقت. ولنفترض أن $I_A(t)$ يشير إلى مؤشر الدالة لمجموعة معينة A. فعندما تكون دالة النواة K من النوع المنتظم Uniform Kernel الموضحة في أدناه:

$$K(t) = I_{[-1,1]}(x)/2 = \begin{cases} \frac{1}{2} & , \quad x \in [-1,1] \\ 0 & , \quad \text{otherwise} \end{cases} \qquad \dots (9)$$

فإن مقدّر N-W الموضح بالصيغة (8) يمثل المتوسط الموضعي لقيم y_i التي تكون داخل الجوار الموضعي $I_h(x_o)$ الموضح في الصيغة (3):

$$\hat{m}_h(x_o) = \frac{\sum_{i=1}^{n} \mathbf{I}_{[x_o-h+x_o+h]}(x_i)\, y_i}{\sum_{i=1}^{n} \mathbf{I}_{[x_o-h+x_o+h]}(x_i)} = \left\{ \sum_{x_i \in \mathbf{I}_h(x_o)} y_i \right\} \Big/ m_h(x_o)$$

حيث أن تمثل $m_h(x_o)$ عدد المشاهدات الواقعة ضمن الجوار الموضعي $I_h(x_o)$.

أما بالنسبة للممهد الخطي الموضعي الذي أُقترح من قبل كل من Stone (1984) و(1993) Fan فإنه يتم استحصاله عن طريق ملائمة مجموعة بيانات موضعياً مع دالة خطية. هنا نفترض أن يكون لدينا $(\hat{\beta}_o, \hat{\beta}_1)$ التي تقلل معيار WLS التالي:

$$\sum_{i=1}^{n} \left[y_i - \beta_o - (x_i - x_o)\beta_1\right]^2 K_h(x_i - x_o)$$

وبالتالي فإن الممهد الخطي الموضعي هو $\hat{m}_h(x_o)=\hat{\beta}_o$. إذ يمكن الحصول عليه بسهولة وذلك باستخدام ممهد LPK $\hat{m}_h(x_o)$ والموضح في الصيغة (5) وذلك بجعل p=1. لذا يمكن التعبير بشكل مبسط عن الممهد الخطي الموضعي كما موضح بالصيغة في أدناه:

$$\hat{m}_h(x_o)=\frac{\sum_{i=1}^{n}\left[s_2(x_o)-s_1(x_o)(x_i-x_o)\right]K_h(x_i-x_o)y_i}{s_2(x_o)\,s_o(x_o)-s_1^2(x_o)} \quad \ldots (10)$$

حيث أن:

$$s_r(x_o)=\sum_{i=1}^{n}K_h(x_i-x_o)(x_i-x_o)^r \quad , \quad r=0,1,2$$

عادة فإن اختيار درجة ممهد LPK المناسبة، p، لا تكون مهمة بقدر أهمية اختيار عرض الحزمة، h. إذ أن كل من ممهد الثابت الموضعي (عندما p=0) والممهد الخطي الموضعي (عندما p=1) غالباً ما تكون جيدة بما فيه الكفاية لمعظم مشاكل التطبيق إذا تم تحديد دالة النواة K وعرض الحزمة h بشكل كاف.

لذا فإن دالة النواة K(.) المستخدمة في ممهد LPK المتمثل بالصيغة (5) هي عادة ما تكون دالة كثافة احتمالية متماثلة. وبما أن عرض الحزمة h يحدد حجم الجوار الموضعي $I_h(t_o)$ ، فإن دالة النواة K(.) تحدد الكيفية التي تساهم بها المشاهدات لملائمة ممهد LPK عند النقطة x_o.

3. تحديد عرض الحزمة Bandwidth Selection [2] [6] [8]

يُعد الممهد جيداً اذا نتج عنه خطأً تنبوئياً صغيراً، وهذا عادة ما يُقاس عن طريق متوسط الخطأ المطلق (MAE) Mean Absolute Error أو متوسط مربع الخطأ (MSE) Mean Square Error الخاص بالممهد. فبالنسبة لممهد النواة المتعدد الحدود الموضعي LPK $\hat{m}_h(x_o)$ ، فإنه يمكن تعريف كل من MAE و MSE كما موضح بالشكل التالي:

$$\begin{aligned} \text{MSE}\left(\hat{m}_h(x_o)\right) &= E\left(\hat{m}_h(x_o)-m(x_o)\right)^2 \\ &= \text{Bias}^2\left(\hat{m}_h(x_o)\right)+\text{Var}\left(\hat{m}_h(x_o)\right) \end{aligned} \quad \ldots (11)$$

$$\text{MAE}\left(\hat{m}_h\right)=\frac{1}{n}\sum_{i=1}^{n}\left|m_{(i)h}-\hat{m}_{(i)h}\right|$$

تم اعتماد عدة صيغ لدوال النواة التي من الممكن اعتمادها في عملية تمهيد دالة الانحدار الامعلمي. وإن أفضل دالة من دوال النواة المقترحة هي التي تقلل من قيمة معياري متوسط الخطأ المطلق Mean Absolute Error (MAE) ومتوسط مربعات الخطأ Mean Squared Error (MSE). الجدول التالي يوضح بعض أهم دوال النواة والمستخدمة في تقدير نماذج الانحدار اللامعلمي مع ثلاث دوال مقترحة:

جدول رقم (1) يوضح دوال النواة المعتمدة والمقترحة

Function Name	Mathematical Formula	Interval
Gaussian	$K(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$	$I\{\lvert x\rvert \le \infty\}$
Epanechnikov	$K(x)=\frac{3}{4}(1-x^2)$	$I\{\lvert x\rvert \le 1\}$
Quartic	$K(x)=\frac{15}{16}(1-x^2)^2$	$I\{\lvert x\rvert \le 1\}$
Triweight	$K(x)=\frac{35}{32}(1-x^2)^3$	$I\{\lvert x\rvert \le 1\}$
SKernelG1	$K(x)=\frac{25}{48}(1-\frac{x^4}{5})$	$I\{\lvert x\rvert \le 1\}$
SKernelG2	$K(x)=\frac{15}{28}(1-\frac{x^4}{3})$	$I\{\lvert x\rvert \le 1\}$
SKernelG3	$K(x)=\frac{405}{712}(1-\frac{x^4}{3})^2$	$I\{\lvert x\rvert \le 1\}$

إنحدار الشريحة Spline Regression (SR) [2] [4] [8]

سيتم في هذا الجزء من البحث مناقشة تمهيد انحدار الشريحة، إذ سنقوم بتحديد الجوار الموضعية من قبل مجموعة من المواقع وكما يلي:

$$\tau_0, \tau_1, \tau_2, ..., \tau_K, \tau_{K+1}$$

وذلك ضمن مدى معين من مجال اهتمام الباحث، كأن تكون فترة [a , b] حيث أن $a=\tau_0<\tau_1<...<\tau_K<\tau_{K+1}=b$. هذه المواقع تُعرف بإسم العقد knots، وأن τ_r

حيث $r=1,2,\ldots,K$ تدعى بالعقد الداخلية interior knots أو العقد knots. هذه العقد تقوم بتقسيم الفترة [a , b] إلى K من الفترات الفرعية (الجوار الموضعية):

$$[\tau_r\,,\tau_{r+1}),\; r=0,1,\ldots,K$$

بعبارة أخرى، فإن انحدار الشريحة يمثل متعدد حدود مقطعي piecwise polynomial وهو متعدد حدود من درجة معينة داخل أي عقدتين متجاورتين τ_r و τ_{r+1} لكل $r = 0, 1, \ldots, K$ والتي قد تم دمجها في عقد قد تسمح بوجود مشتقات متقطعة على تلك العقد.

1- أساس قوة القطع Truncated Power Basis [5] [8]

يمكن بناء انحدار الشريحة باستخدام ما يدعى بأساس قوة القطع من الدرجة k مع K من العقد $\tau_1,\tau_2,\ldots,\tau_K$:

$$1,x,\ldots,x^k,(x-\tau_1)_+^k,\ldots,(x-\tau_K)_+^k \qquad \ldots(12)$$

حيث أن $(x-\tau_r)_+^k$ تشير إلى القوة k من الجزء الموجب للمقدار، أي أن $(x-\tau_1)_+=\max(0,(x-\tau_1))$. مع ملاحظة أن أول (k+1) من الدوال الأساس بالنسبة لأساس قوة القطع المتمثلة بالصيغة (12) تمثل متعدد حدود من الدرجة قد تصل إلى k، أما البقية فتمثل جميع دوال أساس القطع من الدرجة k. بالتالي فإذا كانت درجة أساس قوة القطع k=0,1,2,3 تدعى بأساس قوة القطع "الثابتة، الخطية، التربيعية والتكيعيبية" على التوالي.

هنا يمكن توضيح انحدار الشريحة باستخدام أساس قوة القطع المتمثلة بالصيغة (12) كما موضح في أدناه:

$$m(x)=\sum_{s=0}^{k}\beta_s x^s+\sum_{r=1}^{K}\beta_{k+r}(x-\tau_r)_+^k \qquad \ldots(13)$$

حيث أن $\beta_0,\beta_1,\ldots,\beta_{k+K}$ تمثل المعاملات المرتبطة. وهذا يمكن أن نطلق عليه انحدار الشريحة من الدرجة k مع العقد $\tau_1,\ldots,\tau_K$. بالتالي فإن انحدار الشريحة المتمثل بالصيغة (13) عند k=1,2,3 عادة ما يدعى بانحدار الشريحة الخطية والتربيعية والتكيعيبية وعلى التوالي.

ويمكننا أن نرى هنا أنه في داخل أي فترة فرعية أو جوار موضعي $[\tau_r\,,\tau_{r+1})$ ، يكون لدينا التالي:

$$m(x)=\sum_{s=0}^{k}\beta_s x^s+\sum_{l=1}^{r}\beta_{k+1}(x-\tau_1)^k$$

وهذا المقدار يمثل بدوره متعدد حدود من الدرجة k. مع ذلك، إذا كان لدينا $r = 1,2,\ldots, K$ فإنه ينتج عن ذلك:

$$m^{(k)}(\tau_r-)=k!\,(\beta_k+\sum_{l=1}^{r-1}\beta_{k+1})$$

$$m^{(k)}(\tau_r+)=k!\,(\beta_k+\sum_{l=1}^{r}\beta_{k+1})$$

وبالتالي يكون لدينا:

$$m^{(k)}(\tau_r+)-m^{(k)}(\tau_r-)=k!\,\beta_{k+r} \qquad \ldots (14)$$

علماً أن $m^{(k)}(x)$ تمثل قفزات عند النقاط τ_r مع المقدار $k!\beta_{k+r}$ لكل $r = 1, 2, \ldots, K$. لذا فإن انحدار الشريحة من الدرجة k مع العقد $\tau_1,\ldots,\tau_K$ يمتلك ما يصل إلى k-1 من المشتقات المستمرة continuous derivatives، و k من المشتقات المتقطعة discontinuous derivatives، وأن المعامل β_{k+r} لدالة أساس قوة القطع من الدرجة r تقيس مقدار القفزة الواحدة.

2- ممهد انحدار الشريحة Regression Spline Smoother [5] [8]

غالباً ما يتم الاشارة إلى أساس قوة القطع الموضحة بالصيغة (12) بالشكل التالي:

$$\Phi_p(x)=\left[1, x,\ldots, x^k, (x-\tau_1)_+^k,\ldots,(x-\tau_K)_+^k\right]^T \qquad \ldots (15)$$

حيث أن p=K+k+1 والتي تمثل بدورها عدد الدوال الأساس المشاركة. وبشكل مماثل يمكن التعبير عن المعاملات المرتبطة يمكن توضحها كما يلي:

$$\beta=\left[\beta_0,\beta_1,\ldots,\beta_k,\beta_{k+1},\ldots,\beta_{k+K}\right]^T$$

وبالتالي فإنه يمكن إعادة صياغة انحدار الشريحة الموضح بالصيغة (2-12) كما في الشكل التالي: $m(x)=\Phi_p(x)^T\,\beta$

وأن الأنموذج الموضح بالصيغة (1) يمكن إعادة صايغته كما يلي:

$$y = X\beta+\varepsilon \qquad \ldots (16)$$

حيث أن:

$$\begin{aligned} y &= (y_1,\dots,y_n)^T \\ X &= \left(\Phi_p(x_1),\dots,\Phi_p(x_n)\right)^T \qquad \dots (17) \\ \varepsilon &= (\varepsilon_1,\dots,\varepsilon_n)^T \end{aligned}$$

وبما أن تمثل $\Phi_p(x)$ الأساس، وأن X هي من رتبة كاملة، لذا فإن X^TX تكون معكوسة invertible عندما تكون $n \ge p$. إذ يمكن توضيح المقدر طبيعي للمعلمة β، والذي يحل الانموذج الخطي التقاربي approximation linear model الموضح بالصيغة (17) وذلك باستخدام طريقة المربعات الصغرى (OLS)، كما في الشكل التالي:

$$\hat{\beta} = \left(X^TX\right)^{-1} X^T y \qquad \dots (18)$$

وهذا يؤدي إلى أن ملائمة انحدار الشريحة للدالة m(x) الموضحة بالصيغة (1) كما يلي:

$$\hat{m}_p(x) = \Phi_p(x)^T (X^TX)^{-1} X^T y \qquad \dots (19)$$

وهذا عادة ما يدعى بممهد انحدار الشريحة للدالة *m*. وعلى وجه الخصوص، فإن قيم $\hat{m}_k(x)$ المحتسبة عند نقاط وقت التصميم x_i لكل $i=1,2,\dots,n$ تكون مرتبطة كما هو موضح في متجه الاستجابة التالي:

$$\hat{y}_p = X\left(X^TX\right)^{-1} X^T y = A_p y \qquad \dots (20)$$

حيث أن:

$$\begin{aligned} \hat{y}_p &= \left(\hat{y}_1,\dots,\hat{y}_n\right)^T \\ \hat{y}_i &= \hat{m}_p(x_i)\ ,\ i = 1,2,\dots,n \end{aligned}$$

وأن:

$$A_p = X\left(X^TX\right)^{-1} X^T \qquad \dots (21)$$

وهو ما يسمى بمصفوفة ممهد انحدار الشريحة. إذ من السهل ملاحظة أن المصفوفة A_p هي مصفوفة التوقعات التي تحقق $A_p^T = A_p$ ، $A_p^2 = A_p$ وأن $tr(A_p) = p$. هنا يجدر الاشارة إلى أن أثر المصفوفة A_p يمثل درجة حرية ممهد انحدار الشريحة، والذي بدوره يقيس مدى التعقيد في أنموذج انحدار الشريحة المستخدم.

3- تحديد عدد ومواقع العقد Selection Number and Location of Knots [8] [9]

إن التحديد الأمثل لعدد العقد K ومواقعها $\tau_i\ ,\ i = 1,2,\dots,K$ في انحدار الشريحة يضمن للباحث أن يكون أداء انحدار الشريحة جيداً. إذ يوجد طرق عدة يمكن اعتمادها في

تحديد عدد ومواقع العقد، ومن أبرز تلك الطرق طريقة المسافات المتساوية Equally Spaced Method ، وطريقة ترتيب العينة بشكل مجزءات متساوية المسافة Equally Spaced Sample Quintiles as Knots Method.

تمهيد الشريحة Smoothing Splines [8]

كما بيّنا سابقاً، فيما يخص انحدار الشريحة عندما يتم تعيين الطريقة المتبعة في تحديد مواقع العقد، فإن الذي يلي ذلك هو اختيار عدد العقد K. وبصورة عامة فإن عدد العقد K يكون أصغر من حجم العينة. وأن يكون عدد العقد K عدد صحيح الامر الذي يؤدي إلى أن تكون المقترحات المناسبة لايجاده محدودة، ومنها الطرق آنفة الذكر. إضافة إلى ذلك يمكن إعتماد أن تكون العقد على كل نقطة من نقاط بيانات العينة وهذا يؤدي إلى إعتماد عدد كبير جداً من العقد. المشكلة هنا أن هذا الاجراء يؤدي إلى الحصول على تمهيد لمنحنى دالة الانحدار يكون أدنى من مستوى التمهيد المطلوب (Under Smoothing) ولكن هذا المنحنى بدوره يكون متذبذباً (Rough) وهذا ما يؤدي إلى تغيرات كثيرة في المنحني ولفترة قصيرة. لذا فإن أسلوب تمهيد الشرائح سيتجاوز تلك المشكلة من خلال تقديم قدي جزائي penalty على الجزء غير الممهد من ذلك المنحنى.

لنفترض هنا أن يكون لدينا فترة محددة [a, b]، فإنه عادة يمكن توضيح الجزء غير الممهد للدالة m على أنه تكامل مربع المشتقة k وكما يلي:

$$\int_a^b \left\{ m^{(k)}(x) \right\}^2 dx \qquad \dots (22)$$

حيث أن $k \geq 1$. لذا فإن ممهد الشريحة الممهدة m الموضح بالصيغة (1) يمكن أن يعرّف على أنه مقلل $\hat{m}_\lambda(x)$ لمعيار المربعات الصغرى الجزائية Penalized Least Squares (PLS) التالي:

$$\sum_{i=1}^{n} \left[y_i - m(x_i) \right]^2 + \lambda \int_a^b \left\{ m^{(k)}(x) \right\}^2 dx \qquad \dots (23)$$

حيث أن $\lambda > 0$ وتكون محددة مسبقاً. الجزء الأول من الصيغة أعلاه تمثل مجموع مربعات البواقي residual sum of squares ، والجزء الثاني منه والذي يكون مرجحاً بالقيمة λ يمثل جزاء الخشونة roughness penalty والذي يفرض بدوره قيداً جزائياً على الجزء غير الممهد من المنحني.[9]

التمهيد الحصين Robust Smoothing [1] [4]

يفترض عادة، في تحليل الانحدار، أن تتبع البواقي residuals التوزيع الطبيعي بمتوسط يساوي صفر وتباين ثابت عادة مايكون مجهول. فإذا كان لدينا انموذج الانحدار الموضح بالصيغة (1) لكل i=1,2,…,n، فإن ε_i تمثل أخطاءً عشوائيةً متوزعة بشكل مستقل ومتماثل i.i.d. random errors ، وإننا كالعادة نحتاج إلى ايجاد مقدر $\hat{m}$ بالنسبة لدالة الانحدار m. ولكننا في التمهيد الحصين فأن الافتراضات التقليدية حول الأخطاء العشوائية أنها تمتلك متوسط يساوي صفر لن تكون مفيدة هنا. ويجدر الاشارة هنا إلى إن العيب الرئيسي لطريقة المربعات الصغرى الجزائية هو حساسيتها تجاه القيم المتطرفة. لذا فإن النقاط البعيدة عن شكل انتشار البيانات قد تؤثر سلباً على نتائج نماذج الانحدار. وسنتناول في هذا الجزء من الفصل بعض أساليب التمهيد الحصينة.

1- الشريحة الممهدة الحصينة Robust Smoothing Splines [1] [4]

يمكن تعريف الشرائح الممهدة الحصينة من خلال ايجاد مقدر $\hat{m}$ الذي يقلل المقدار التالي:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left[y_i-\hat{m}(x_i)\right]^2+\lambda\int_a^b\left\{m^{(k)}(x)\right\}^2dx \qquad \ldots (24)$$

وكما هو واضح من هذه الصيغة أنها نفس صيغة المربعات الصغرى الجزائية ولكن مع وجود الدالة ρ مضروباً بمقلوب حجم العينة n، علماً أن λ تمثل معلمة التمهيد وتكون قيمتها أكبر من الصفر، وبشكل عام فأنه يتطلب أن تكون الدالة ρ متماثلة ومحدبة ومتزايدة بشكل بطئ كلما ازدادت قيمة $|x|$. ويتم الحصول على قيمة ρ وفق الصيغة التالية:

$$\rho_c(x)=\begin{cases} x^2 & \text{if} \quad |x|\le c \\ 2c|x|-c^2 & \text{if} \quad |x|>c \end{cases} \qquad \ldots (25)$$

ومن الواضح أن دالة $\rho(x)$ الموضحة في (25) تحقق الشروط أعلاه. حيث أن C تمثل نقطة التوقف وهي عامل قياس scale factor عادة ما يتم تحديدها من الصيغة الآتية:

$$C=10^{-5}\sqrt{\text{var }(x)}$$

2- انحدار الشريحة الحصين Robust Spline Regression

يمكن تعريف انحدار الشريحة الحصين باستخدام ما يدعى بأساس قوة القطع من الدرجة k مع K من العقد $\tau_1, \tau_2, ..., \tau_K$ ، والموضح بالصيغة (2-11)، كما في الصيغة التالية:

$$m(x) = \rho \sum_{s=0}^{k} \beta_s x^s + \rho \sum_{r=1}^{K} \beta_{k+r} (x - \tau_r)_+^k \qquad \dots (26)$$

حيث أن $\beta_0, \beta_1, ..., \beta_{k+K}$ تمثل المعاملات المرتبطة مضروباً بالدالة ρ. وهذا يمكن أن نطلق عليه انحدار الشريحة الحصين من الدرجة k مع العقد $\tau_1, ..., \tau_K$.

3- ممهد النواة المتعدد الحدود الموضعي الحصين Robust LPK Smoother [1] [3] [5]

يمكن تعريف ممهد النواة المتعدد الحدود الموضعي الحصين عند كل من نقاط x من خلال ايجاد قيم $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ الذي يقلل المقدار الآتي:

$$\sum_{i=1}^{n} \rho \left\{ y_i - \beta_o - \beta_1 (x - x_i) - ... - \beta_p (x - x_i)^p \right\}^2 K_h (x - x_i) \dots (27)$$

حيث أن القيمة المقدرة الناتجة هي $\hat{m}(x) = \sum_{i=1}^{p} \hat{\beta}_i x^i$. وسنفترض هنا أن درجة p هنا p=2، وهي أقصى حد يسمح به هنا. وأن K(.) هي إحدى دوال النواة المعتمدة. وأن h تمثل عرض الحزمة. علماً أن قيمة ρ تحدد بنفس الطريقة المذكورة في أعلاه.

الجانب التجريبي

وصف تجربة المحاكاة

تم تنفيذ تجارب المحاكاة بأعتماد ثلاث حجوم للعينات هي (n_3=150, n_2=100, n_1=50) وذلك لتوليد البيانات الخاصة بالمتغيرات العشوائية لغرض محاكاة البيانات الحقيقية بالنسبة لأسلوب تمهيد النواة الموضعي الحصين وذلك بكتابة برنامج في لغة MATLAB، إذ تم تكرار كل تجربة على حدة 1000 مرة لغرض الحصول على نتائج متسقة. تم اعتماد ثلاث دوال رياضية مختلفة لتمثيل المتغير التوضيحي مع الأخذ بنظر الاعتبار الصيغة الخطية واللاخطية في تلك الدوال لتتلائم وطبيعة البيانات المتعلقة بالمجال الاقتصادي، والدوال التي تم استخدامها هي:

دالة خطية [3] $\qquad f(x) = 2x - 1$

دالة لاخطية [2,3] $f(x) = x + 2\exp(-16x^2)$

دالة من الدرجة الثانية [2] $m(x) = 8(x - 0.5)^2$

كما تم تلويث البيانات لكل من حجوم العينات أعلاه بنسب تلويث (30%,20%,10%). وكانت نتائج تجربة المحاكاة كما يأتي:

جدول (2) يبين المعدل لقيم المعيارين (MAE) و (MSE) * ولجميع حجوم العينات ومستويات التلويث لمقارنة دوال النواة المستخدمة في طريقة النواة الحصينة (RLPK) بالنسبة لأنموذج المقارنة الأول

Sample Size حجوم العينات	Kernel Fu. دوال النواة	مستويات التلويث الثلاث		
		10%	20%	30%
n_1=50	Epanechnikov	0.550606321592160 (0.49590803363779)	0.514424301311390 (0.45106615682822)	0.476091562556496 (0.405022215724994)
	Triweight	0.550605296901387 (0.49590570875978)	0.514422928582408 (0.45106330267745)	0.476090227303251 (0.405020287211766)
	SKernelG1	0.550605296848528 (0.49590570886646)	0.514422928779634 (0.45106330314209)	0.476090227530885 (0.405020287343311)
	SKernelG2	0.550605296809004 (0.49590570893225)	0.514422928910512 (0.45106330345091)	0.476090227683186 (0.405020287427701)
	SKernelG3	0.550605296710188 (0.49590570909674)	0.514422929237701 (0.45106330422297)	0.476090228063935 (0.405020287638677)
	SKernelG4	0.550605980226647 (0.49590725830931)	0.514423842974052 (0.45106520483136)	0.476091117370399 (0.405021572510934)
	SKernelG5	0.550605638421178 (0.49590648340706)	0.514423385015681 (0.45106425370487)	0.476090672495203 (0.405020929744979)
n_2=100	Epanechnikov	0.553927246686775 (0.50285475742417)	0.517165922793710 (0.45728848718117)	0.474983364410770 (0.403104607955904)
	Triweight	0.553926529847818 (0.50285352205501)	0.517165333129738 (0.45728754615361)	0.474982842204403 (0.403103902873135)
	SKernelG1	0.553926529792949 (0.50285352196710)	0.517165333044438 (0.45728754597906)	0.474982842128314 (0.403103902642130)
	SKernelG2	0.553926529753343 (0.50285352190254)	0.517165332983717 (0.45728754585668)	0.474982842074445 (0.403103902481666)
	SKernelG3	0.553926529654330 (0.50285352174115)	0.517165332831918 (0.45728754555074)	0.474982841939775 (0.403103902080507)
	SKernelG4	0.553927007792511 (0.50285434547026)	0.517165726442919 (0.45728817349265)	0.474983190238358 (0.403104372960334)
	SKernelG5	0.553926768908955 (0.50285393362980)	0.517165529859340 (0.45728785971766)	0.474983016031761 (0.403104137802051)
n_3=150	Epanechnikov	0.556793936451551 (0.50803314806272)	0.517509418443547 (0.45987838445150)	0.477831112032046 (0.408009825643928)
	Triweight	0.556793575304523 (0.50803275781841)	0.517509059792099 (0.45987783326517)	0.477830729567047 (0.408009497420028)

* القيم الموضحة في الجدول والمحصورة بين قوسين تمثل قيمة المعيار MSE وباقي القيم هي للمعيار MAE.

Sample Size حجوم العينات	Kernel Fu. دوال النواة	مستويات التلويث الثلاث		
		10%	20%	30%
	SKernelG1	0.556793575152542 (0.50803275740033)	0.517509059659196 (0.45987783305749)	0.477830729459045 (0.408009497083550)
	SKernelG2	0.556793575046734 (0.50803275711197)	0.517509059567280 (0.45987783291372)	0.477830729383135 (0.408009496851294)
	SKernelG3	0.556793574782214 (0.50803275639107)	0.517509059337490 (0.45987783255430)	0.477830729193356 (0.408009496270658)
	SKernelG4	0.556793815927109 (0.50803301816134)	0.517509298753405 (0.45987820074631)	0.477830984630372 (0.408009716383576)
	SKernelG5	0.556793695521890 (0.50803288784467)	0.517509179162448 (0.45987801690082)	0.477830856923972 (0.408009606784607)

جدول (3) يبين المعدل لقيم المعيارين (MAE) و (MSE) ولجميع حجوم العينات ومستويات التلويث لمقارنة دوال النواة المستخدمة في طريقة النواة الحصينة (RLPK) بالنسبة لأنموذج المقارنة الثاني

Sample Size حجوم العينات	Kernel Fu. دوال النواة	مستويات التلويث الثلاث		
		10%	20%	30%
$n_1=50$	Epanechnikov	0.551502479932774 (0.49935227187767)	0.512416659462867 (0.44669809064794)	0.473276147579454 (0.397882895701467)
	Triweight	0.551501022624010 (0.49934933171390)	0.512414753138089 (0.44669559336090)	0.473275079675835 (0.397880748276475)
	SKernelG1	0.551501022760987 (0.49934933213880)	0.512414753406436 (0.44669559360949)	0.473275079793959 (0.397880748724547)
	SKernelG2	0.551501022850743 (0.49934933242099)	0.512414753582971 (0.44669559377077)	0.473275079872537 (0.397880749025056)
	SKernelG3	0.551501023075127 (0.49934933312646)	0.512414754024302 (0.44669559417399)	0.473275080068977 (0.397880749776324)
	SKernelG4	0.551501993053879 (0.49935129116463)	0.512416023447624 (0.44669725773506)	0.473275791550836 (0.397882179257255)
	SKernelG5	0.551501507703879 (0.49935031134555)	0.512415387485790 (0.44669642543700)	0.473275435806884 (0.397881463698522)
$n_2=100$	Epanechnikov	0.553977323859071 (0.50294910799549)	0.517223960740023 (0.45741914766144)	0.474979503865473 (0.403186011770163)
	Triweight	0.553976587158453 (0.50294783174070)	0.517223355045575 (0.45741819796470)	0.474979021445756 (0.403185281149681)
	SKernelG1	0.553976587150698 (0.50294783171998)	0.517223354963371 (0.45741819780676)	0.474979021354372 (0.403185280970320)
	SKernelG2	0.553976587143498 (0.50294783170147)	0.517223354904601 (0.45741819769567)	0.474979021290227 (0.403185280845185)
	SKernelG3	0.553976587125497 (0.50294783165521)	0.517223354757675 (0.45741819741796)	0.474979021129864 (0.403185280532355)
	SKernelG4	0.553977078068155 (0.50294868235212)	0.517223758720659 (0.45741883106267)	0.474979342943126 (0.403185768213849)
	SKernelG5	0.553976832312633 (0.50294825691900)	0.517223556746300 (0.45741851440778)	0.474979182149505 (0.403185524572059)
$n_3=150$	Epanechnikov	0.556906954643633 (0.50820897991896)	0.517504797059077 (0.45832992624582)	0.478202951218034 (0.410574510134826)
	Triweight	0.556906566303909 (0.50820854265084)	0.517504385579340 (0.45832954000413)	0.478202582687192 (0.410573946085847)

	SKernelG1	0.556906566190528 (0.50820854228813)	0.517504385504843 (0.45832953967452)	0.478202582615163 (0.410573945969155)
	SKernelG2	0.556906566111273 (0.50820854203777)	0.517504385452119 (0.458329539446968)	0.478202582564884 (0.410573945887555)
	SKernelG3	0.556906565913134 (0.50820854141187)	0.517504385320308 (0.458329538878086)	0.478202582439185 (0.410573945683556)
	SKernelG4	0.556906825191009 (0.508208834281367)	0.517504659967735 (0.458329797622832)	0.478202828538742 (0.410574322079610)
	SKernelG5	0.556906695809801 (0.508208688322471)	0.517504522685284 (0.458329668689781)	0.478202705641694 (0.410574133997291)

جدول (4) يبين المعدل لقيم المعيارين (MAE) و (MSE) ولجميع حجوم العينات ومستويات التلويث لمقارنة دوال النواة المستخدمة في طريقة النواة الحصينة (RLPK) بالنسبة لأنموذج المقارنة الثالث

Sample Size حجوم العينات	Kernel Fu. دوال النواة	مستويات التلويث الثلاث		
		10%	20%	30%
$n_1=50$	Epanechnikov	0.553737552996597 (0.500966476271125)	0.521532025187929 (0.46030516387083)	0.477698834100755 (0.403110114879408)
	Triweight	0.553735318190301 (0.500962195073756)	0.521529259391818 (0.460300408959781)	0.477696415709360 (0.403106765222452)
	SKernelG1	0.553735317869667 (0.500962195115246)	0.521529259152775 (0.460300409282482)	0.477696415650856 (0.403106765186616)
	SKernelG2	0.553735317642689 (0.500962195129218)	0.521529258975893 (0.460300409484977)	0.477696415602838 (0.403106765150910)
	SKernelG3	0.553735317075235 (0.500962195164151)	0.521529258533685 (0.460300409991217)	0.477696415482789 (0.403106765061646)
	SKernelG4	0.553736808591550 (0.500965049341630)	0.521531104299657 (0.460303578899614)	0.477698028674341 (0.403108998560352)
	SKernelG5	0.553736062860998 (0.500963622291300)	0.521530182146826 (0.460301994096480)	0.477697222420978 (0.403107881981425)
$n_2=100$	Epanechnikov	0.558734001910604 (0.511196289404968)	0.519533097425109 (0.460239148141440)	0.482611647483617 (0.414576659999132)
	Triweight	0.558732576064724 (0.511193670468877)	0.519531415100579 (0.460236645572016)	0.482609817002357 (0.414574258876556)
	SKernelG1	0.558732575491715 (0.511193669799727)	0.519531414913502 (0.460236645265480)	0.482609816662900 (0.414574258339206)
	SKernelG2	0.558732575091956 (0.511193669329033)	0.519531414779954 (0.460236645046808)	0.482609816422870 (0.414574257961754)
	SKernelG3	0.558732574092563 (0.511193668152310)	0.519531414446086 (0.460236644500134)	0.482609815822800 (0.414574257018135)
	SKernelG4	0.558733527677929 (0.511195417284018)	0.519532536462698 (0.460238314393160)	0.482611037675928 (0.414575860368294)
	SKernelG5	0.558733052099899 (0.511194543922038)	0.519531975842775 (0.460237480021191)	0.482610427328972 (0.414575059686449)
$n_3=150$	Epanechnikov	0.560891834405105 (0.514517535382731)	0.522094176646828 (0.464603986512316)	0.483372252803853 (0.416822034489933)
	Triweight	0.560890665759419 (0.514515795673643)	0.522094179521961 (0.464602234612618)	0.483370631850987 (0.416819878541148)
	SKernelG1	0.560890665229843 (0.514515794638693)	0.522094178965989 (0.464602233685958)	0.483370631445616 (0.416819878060080)
	SKernelG2	0.560890664862053 (0.514515793922176)	0.522094178579469 (0.464602233043773)	0.483370631161519 (0.416819877721513)

	SKernelG3	0.560890663942581 (0.514515792130900)	0.522094177613171 (0.464602231438321)	0.483370630451280 (0.416819876875104)
	SKernelG4	0.560891445009014 (0.514516956650491)	0.522095044028478 (0.464603403652450)	0.483371713291395 (0.416821316535714)
	SKernelG5	0.560891055413451 (0.514516376160805)	0.522094611888135 (0.464602819159472)	0.483371172723268 (0.416820597610723)

جدول (5) يبين المعدل لقيم المعيارين (MAE) و (MSE) ولجميع حجوم العينات ومستويات التلويث لمقارنة طريقة النواة الحصينة (RLPK) والشرائح الجزائية الحصينة (RPS) وأنحدار الشريحة الحصين (RSR) للأنموذج الأول

Sample Size حجوم العينات	طرائق التقدير الثلاث	مستويات التلويث الثلاث		
		10%	20%	30%
n_1=50	RLPK	0.552182677628186 (0.497987853181795)	0.514093857186375 (0.451740107344066)	0.472890417385481 (0.397028322465322)
	RPS	0.538134174160625 (0.473860757165212)	0.502248147272791 (0.429990828617491)	0.462963639876114 (0.379285140531454)
	RSR	0.494252757613744 (0.411193966551557)	0.464118029532197 (0.372720643229598)	0.428039635654853 (0.329029408229980)
n_2=100	RLPK	0.553926529654330 (0.502853521741159)	0.517165332831918 (0.457287545550743)	0.476139510917880 (0.405151278538561)
	RPS	0.542582229317600 (0.483540973480450)	0.507104079448146 (0.439355013866511)	0.467359006143270 (0.389545559131440)
	RSR	0.526922283383404 (0.460164554751303)	0.494020912219265 (0.418349850733305)	0.456073029838743 (0.370768708385820)
n_3=150	RLPK	0.556582116036679 (0.507763969466826)	0.517371520737847 (0.458132859697191)	0.476589173071062 (0.406528276210840)
	RPS	0.545865930085604 (0.489122329681300)	0.507888999885514 (0.440922411161987)	0.468081043418265 (0.391033534289272)
	RSR	0.539013253267200 (0.478740130687058)	0.501854557975388 (0.431547744606150)	0.463363727187112 (0.382856706987234)

جدول (6) يبين المعدل لقيم المعيارين (MAE) و (MSE) ولجميع حجوم العينات ومستويات التلويث لمقارنة طريقة النواة الحصينة (RLPK) والشرائح الجزائية الحصينة (RPS) وأنحدار الشريحة الحصين (RSR) للأنموذج الثاني

Sample Size حجوم العينات	طرائق التقدير الثلاث	مستويات التلويث الثلاث		
		10%	20%	30%
n_1=50	RLPK	0.551278289519607 (0.497138176837567)	0.512759499146262 (0.449501099784590)	0.473275548594472 (0.396740138543360)
	RPS	0.537273847232076 (0.473357986636949)	0.500722229993292 (0.428607503181588)	0.462346878171822 (0.377819715568827)
	RSR	0.492094370526183 (0.410377655049605)	0.462650151701324 (0.373141226129125)	0.427157754321176 (0.328872579499827)

n_2=100	RLPK	0.555970439283716 (0.506500464496937)	0.514237663610564 (0.452732138267907)	0.476144685738145 (0.405352753677434)
	RPS	0.544311202676510 (0.486442472819519)	0.503878494465002 (0.434485383043029)	0.467035306950199 (0.389095695229558)
	RSR	0.529084881054981 (0.463005868498495)	0.490637435775341 (0.412758412686543)	0.456238395051233 (0.370682901850268)
n_3=150	RLPK	0.555188685828958 (0.506116098820216)	0.517582525170232 (0.459985387872991)	0.480539530202167 (0.413526524213545)
	RPS	0.544194913670195 (0.486705699262586)	0.507561169453336 (0.442215726342324)	0.471968554417522 (0.397837746698252)
	RSS	0.537489413670595 (0.476421462121329)	0.502060196893957 (0.433004651706364)	0.467236645578959 (0.389575684712284)

جدول (7) يبين المعدل لقيم المعيارين (MAE) و (MSE) ولجميع حجوم العينات ومستويات التلويث لمقارنة طريقة النواة الحصينة (RLPK) والشرائح الجزائية الحصينة (RPS) وأنحدار الشريحة الحصين (RSR) للأنموذج الثالث

Sample Size حجوم العينات	طرائق التقدير الثلاث	مستويات التلويث الثلاث		
		10%	20%	30%
n_1=50	RLPK	0.554961775496029 (0.502079830183020)	0.517605312618680 (0.454610798202030)	0.482166962698017 (0.409763428634302)
	RPS	0.537297895250896 (0.472388134999466)	0.500975954817040 (0.426607949372100)	0.465928266293026 (0.384898608042349)
	RSR	0.492304399741873 (0.409535265249682)	0.462510002324010 (0.371856624591930)	0.432271952212960 (0.334928492767941)
n_2=100	RLPK	0.558572935588556 (0.511245330746551)	0.520790505107472 (0.462493330588669)	0.480948476688540 (0.411059874339644)
	RPS	0.542808978812518 (0.484232998268033)	0.506143576928414 (0.438463176791578)	0.467359006143265 (0.389545559131441)
	RSR	0.527101730277805 (0.460408240208546)	0.493077058997775 (0.417345149785998)	0.456073029850926 (0.370768708859802)
n_3=150	RLPK	0.561145247222113 (0.514863836665526)	0.521192691023334 (0.462845133648338)	0.482656757358574 (0.413719371224294)
	RPS	0.546012788494907 (0.489731613648223)	0.506638304575073 (0.439394071102712)	0.469390990013525 (0.392129377925265)
	RSS	0.539074433778907 (0.479411356008275)	0.500907088767684 (0.430235596183497)	0.464664280649894 (0.383684551653578)

الجانب التطبيقي

في هذا الجانب من البحث سيتم التطرق إلى تطبيق طريقة التقدير اللامعلمية الأفضل المستخرجة من الجانب التجريبي وذلك بأعتماد بيانات حقيقية خاصة بحركة التداول لشركة الخليج للتأمين المسجلة في سوق العراق للاوراق المالية. إذ أن البيانات التي تم اعتمادها في هذا الجانب تمثل أسعار الاسهم عند الاغلاق، والكمية المتداولة والذي تم الحصول عليها من الموقع الالكتروني للسوق اذ تم تسجيل قيم المتغيرات على مدى 100 يوماً.

نمذجة العلاقة بين أسعار الأسهم عند الاغلاق والقيمة المتداولة

تمت نمذجة العلاقة بين أسعار الأسهم عند الاغلاق Closing Price (CP) والقيمة المتداولة (حجم التداول) Trading Volume (TV) وذلك باعتبار أن أسعار الأسهم عند الاغلاق تمثل متغير الاستجابة والقيمة المتداولة تمثل المتغير التوضيحي. هنا نحتاج إلى إجراء تحويلاً رياضياً لكل من أسعار الأسهم وحجم التداول وذلك بحساب تقلبات الأسعار Price Volatility وكذلك تقلبات حجم التداول Volume Volatility وفق الصيغ الآتية:

$$PV = \frac{|R_x|}{\sigma_R} \qquad \dots (28)$$

$$VV = \frac{|\tilde{R}_x|}{\sigma_{\tilde{R}}}$$

إذ أن PV تمثل تقلبات الأسعار، و VV تمثل تقلبات حجم التداول، وأن كل من $R, \tilde{R}$ تمثل التغيّر في الأسعار وفي حجم التداول على التوالي، والذي يتم احتسابهما وفق الصيغ التالية:

$$R_x = \ln\left[\frac{CP_{(x)}}{CP_{(x-1)}}\right] \qquad \dots (29)$$

$$\tilde{R}_x = \ln\left[\frac{TV_{(x)}}{TV_{(x-1)}}\right]$$

وبذلك يكون إنموذج الانحدار اللامعلمي المراد تقديره بعد إعتماد التقلبات على أسعار الأسهم وحجم التداول بالصورة الآتية:

$$Y_{PV} = \alpha\,(x_{VV}) + \varepsilon \qquad \dots (30)$$

تقدير إنموذج الانحدار اللامعلمي الحصين

بأستخدام أسلوب تقدير إنموذج الانحدار اللامعلمي الأفضل والمتمثل بأسلوب انحدار الشريحة الحصين (RRS) والذي تم التوصل اليه في الجانب التجريبي من البحث، نقوم بتقدير إنموذج الانحدار اللامعلمي الذي يصف العلاقة بين أسعار الأسهم عند الاغلاق وحجم التداول والموضح بالصيغة (30) وذلك بعد إجراء التحويلات على كلا المتغيرين، علماً أن معامل المتغير التوضيحي x_{VV} والذي يمثل تقلبات حجم التداول يعتبر دالة بالنسبة

إلى سعر السهم عند الفتح. والجدول (8) يبين قيم معياري (MAE) و(MSE) لطرائق التقدير الحصينة الثلاث بالنسبة لحجم عينة(n= 100).

جدول (8) يبين المعدل لقيم المعيارين (MAE) و (MSE) لطرائق التقدير الحصينة الثلاث ولحجم عينة n=100

حجوم العينات	طرائق التقدير الثلاث	معايير المقارنة	
		MAE	MSE
n_1=50	RLPK	0.738027122720328	0.923802032722077
	RPS	0.325676336849908	0.350514886923690
	RSR	0.357817087765453	0.416601366256592

وفيما يخص القيم التقديرية لمتغير الاستجابة الذي يمثل تقلبات أسعار الأسهم عند الاغلاق وأيضاً القيم الحقيقية لهذا المتغير موضحة بالنتائج المبينة في الجدول (9) لتمثل قيم متغير الأستجابة الحقيقية والتقديرية، علماً أن تلك القيم التقديرية قد تم الحصول عليها وفق طرائق التقدير الحصينة الثلاث (RLPK)، (RPS) و(RSR).

جدول (9) يبين القيم الحقيقية والتقديرية لمتغير الاستجابة عند n= 100

Case No.	Y_{PV}	$\hat{Y}_{PV(RLPK)}$	$\hat{Y}_{PV(RPS)}$	$\hat{Y}_{PV(RSR)}$
1	1.532979617	1.098253964	0.876931979	0.552426587
2	0.369765988	1.098277858	0.786840067	0.56455118
3	0	1.098277858	0.363441088	0.395494168
4	0.766054879	1.098278061	0.684075797	0.574186307
5	0.387587721	1.098306271	0.423025088	0.636949096
6	0.78474647	1.098334468	0.873254764	0.669742228
7	0.825007097	1.098344547	0.713526674	0.523480285
8	1.322726791	1.098364662	0.778500348	0.658012112
9	1.441332006	1.098416636	0.602427956	0.438146423
10	0.494934808	1.098480588	0.338125781	0.530645881
11	0.480158406	1.098519893	0.26383198	0.268054745
12	0.480158406	1.098628749	0.44249064	0.592250319
13	0.494934808	1.098881879	0.382223218	0.633540049
14	0.494934808	1.098892229	0.527837625	0.670541971
15	0.946397199	1.098949105	0.662676739	0.682046343
16	0.453103543	1.099011093	0.550372648	0.821963207
17	0.869623248	1.09901453	0.811833113	0.869549752
18	0.417792635	1.09902217	0.452423751	0.413099844
19	1.287415883	1.099094904	1.263637941	1.514035993
20	1.694630345	1.099160792	1.577820644	0.739983844
21	0.78474647	1.09922309	0.783886588	0.784625621
22	0.387587721	1.099263296	0.156814339	0.195704621
23	0.397158749	1.099275989	0.38551262	-0.056845235
24	0.397158749	1.099568058	0.399820375	0.503789403
25	1.135820868	1.099700973	0.560316086	0.652728961
26	1.532979617	1.100006119	0.708215754	0.65399944
27	0	1.100118826	0.634573042	0.681806741
28	0.78474647	1.100175022	1.309302367	0.945578158
29	0.378467158	1.100214818	0.235525814	0.234662202
30	1.084733121	1.100413388	0.804493224	0.492946499
31	1.341119146	1.100500877	0.830055383	0.675403435

Case No.	Y_{PV}	$\hat{Y}_{PV(RLPK)}$	$\hat{Y}_{PV(RPS)}$	$\hat{Y}_{PV(RSR)}$
32	0.324942609	1.100509798	0.268196135	0.366251899
33	1.369687736	1.100590592	1.690447964	1.070199701
34	6.875057453	1.101009299	1.697241699	1.033399787
35	0.234808726	1.101072722	0.36963487	0.599464921
36	0.234808726	1.101419285	0.252344925	0.2566132
37	0.234808726	1.101832751	0.394795484	0.508030411
38	0	1.102011531	0.374028217	0.39787991
39	0	1.10203619	0.182085982	0.206003761
40	1.228391553	1.102410544	0.399864373	0.503736605
41	1.055074045	1.102759513	0.87600509	0.495780664
42	0.270327575	1.103176174	0.546073321	0.645905858
43	1.038044809	1.103207147	0.795535949	0.659134176
44	0	1.103350899	0.366686604	0.396236323
45	0.253298339	1.103961404	0.392659815	0.509380352
46	0	1.104574352	0.684407132	0.528764164
47	0	1.105036907	0.401599342	0.496059027
48	0	1.10570483	0.530342959	0.670050879
49	1.463200279	1.105767182	1.48050656	1.636386409
50	0	1.105801629	0.120225187	0.708066041
51	0	1.105981389	0.099824396	0.175924804
52	0.231430066	1.106237984	0.257843378	0.255668215
53	0.228147259	1.106357991	0.579909511	0.558179144
54	0.228147259	1.106380536	0.252260258	0.300438283
55	0	1.107645361	0.177384855	0.828451252
56	0	1.107750728	0.458533228	0.612907135
57	1.441332006	1.108058299	0.564037407	0.679785301
58	0	1.109449954	0.205127962	0.216763491
59	0.975093214	1.10961258	0.273824205	0.368693614
60	0.234808726	1.109819922	0.113347522	0.695399645
61	1.20990194	1.110510783	0.81949521	0.494770574
62	0.249371074	1.110677546	0.649010665	0.48711909
63	0.502669413	1.111332569	0.554249366	0.646383357
64	0.502669413	1.111996788	1.710294268	1.044130618
65	1.021556738	1.112139848	0.820565011	0.506438042
66	0	1.1128843	0.768417178	0.657383188
67	1.375548225	1.113286679	0.497743401	0.604556063
68	0.839361504	1.114718762	0.729483188	0.570266491
69	0.270327575	1.115230266	0.311098171	0.381797243
70	0	1.115233366	0.127753827	0.185148402
71	1.038044809	1.115469726	1.214557324	0.932031092
72	0.249371074	1.115540402	0.5710965	0.563321307
73	0	1.116113072	0.142143583	0.597613729
74	0	1.116476191	0.363557852	0.60017628
75	0.245563733	1.117437343	0.238422114	0.558613959
76	0.714967132	1.118539898	1.488492616	1.104412907
77	0.684533609	1.120488822	1.224685143	0.933493527
78	0.224956284	1.120749777	0.532525519	0.426246296
79	0.228147259	1.123059433	0.36494982	0.521662649
80	0.228147259	1.123835807	0.6439379	0.445565204
81	0.88154182	1.124800577	0.790241565	0.469976685
82	1.237850013	1.124822503	0.59581507	0.551536135
83	1.453747377	1.125002628	1.629634756	1.007830871
84	0.215897364	1.12595322	0.190232046	0.209625998
85	1.038044809	1.126405964	0.935373339	0.514714535
86	0.592120071	1.126612806	0.656473259	0.447865905
87	0.382973143	1.127171748	0.442804306	0.615695957
88	0.190346744	1.128105004	0.127530708	0.185071522
89	0.190346744	1.136284924	0.127530708	0.185071522
90	0	1.137291789	0.521445869	0.676340536
91	0	1.137758635	0.319538271	0.606421293

Case No.	Y_{PV}	$\hat{Y}_{PV(RLPK)}$	$\hat{Y}_{PV(RPS)}$	$\hat{Y}_{PV(RSR)}$
92	0.188120415	1.139841319	0.095526376	0.682597344
93	0	1.116704547	0.754213009	0.680407018
94	0.369765988	1.109586281	0.554120117	0.486383515
95	0	1.010674417	0.210037977	0.778225952
96	0.557886403	0.984958306	0.201136691	0.920535361
97	0.975093214	0.980729449	0.639615796	0.681874123
98	0	0.972790199	0.601903425	0.648884944
99	0.78474647	0.969043638	0.55682021	0.665970995

الاستنتاجات

على ضوء ما توصل إليه الباحث من نتائج المبحثين التجريبي والتطبيقي، يمكن طرح أهم الاستنتاجات التي تم استخلاصها وهي كالآتي:

1. في للجانب التجريبي ومن خلال مقارنة دوال النواة المعتمدة والمقترحة منها والتي تم استخدامها في طريقة النواة لمتعدد الحدود الموضعي الحصين (RLPK) تبيّن بأن دالة النواة المقترحة (SKernelG3) أنها أفضل دالة لمستويات التلويث المختلفة ولجميع نماذج المقارنة الثلاثة التي تم إعتمادها في تجارب المحاكاة، وذلك بالاعتماد على معياري متوسط الخطأ المطلق (MAE) ومتوسط مربعات الخطأ (MSE).
2. أيضاً وفي نفس المبحث ومن خلال مقارنة طرائق التقدير اللامعلمية الحصينة المعتمدة والمقترحة والمستخدمة في تقدير إنموذج الانحدار اللامعلمي، وجدنا بأن أفضل طريقة كانت طريقة انحدار الشريحة الحصين (RSR) المقترحة، إذ حققت أقل قيمة لمعياري متوسط الخطأ المطلق (MAE) ومتوسط مربعات الخطأ (MSE).
3. لوحظ من نتائج تجربة المحاكاة لتحديد أفضل دالة نواة أن قيم معياري متوسط الخطأ المطلق (MAE) ومتوسط مربعات الخطأ (MSE) لكافة تكرارات التجربة لكل حالة من مستويات التلويث الثلاث ونماذج المقارنة الثلاثة المعتمدة في التجربة، كانت صغيرة ومتقاربة مما يشير إلى تجانس قيم المعيارين لكل حالة من حالات تجربة المحاكاة.
4. بعد إيجاد أفضل دالة نواة حصينة، وهي الدالة المقترحة الثالثة (SKernelG3)، قمنا بمقارنة تلك الدالة مع طرائق التقدير الحصينة الأخرى لتحديد أي الطرائق كانت الأفضل، فكانت النتيجة بأن طريقة انحدار الشريحة الحصينة (RSR) هي الأفضل لتسجيلها أقل قيمة بالنسبة لمعياري متوسط الخطأ المطلق (MAE) ومتوسط مربعات الخطأ (MSE) بالنسبة لمستويات التلويث المختلفة لنماذج المقارنة الثلاثة.

5. وجد الباحث أنه كلما زادت نسبة تلويث البيانات أدى ذلك إلى تقليل قيم معدل متوسط الخطأ المطلق ومتوسط مربعات الخطأ مما يدل على تحسين في نتائج تقدير إنموذج الانحدار اللامعلمي.
6. أما فيما يخص الجانب التطبيقي من البحث، تم اعتماد حجم عينة n=100 ، وبالاعتماد على معيار العبور الشرعي المعمم (GCV) تبين ان أفضل عرض حزمة كان يساوي (0.036752) لحجم العينة n=100 والذي تم اعتماده في طريقة النواة لمتعدد الحدود الموضعي الحصين.
7. بالنسبة لأسلوب الشريحة الجزائية الحصينة (RPS) كانت قيمة معلمة التمهيد عند حجم عينة n=100 كانت تساوي (1.03E-06).
8. بمقارنة نتائج معياري متوسط الخطأ المطلق (MAE) ومتوسط مربعات الخطأ (MSE) لطرائق التمهيد الحصينة الثلاث نجد بأن أقل قيم سجلها هذين المعيارين كانت تخص أسلوب الشريحة الجزائية الحصين (RPS) وبذلك يعد هذا المعيار هو أفضل أسلوب تمهيدي يمكن أن يستخدم في تقدير إنموذج الانحدار اللامعلمي.

المصادر

1. Garcia, D., (2010), "***Robust Smoothing of Gridded Data in One and Higher Dimensions with Missing Values***". Computational Statistics & Data Analysis.
2. Härdle, Wolfgang, (1994), "***Applied Nonparametric Regression***". Cambridge: Cambridge University Press.
3. Kagerer, K., (2013), "***A Short Introduction to Splines in Least Squares Regression Analysis***". University of Regensburg Working Papers in Business, Economics and Management Information Systems, JEL Classification: C14, C51.
4. Lee, Jong S., & Cox, Dennis, D., (2010), "***Robust Smoothing: Smoothing Parameter Selection and Application to Fluorescence Spectroscopy***". Computational Statistics & Data Analysis, 54**.**
5. Schimek, Michael G., (2000), "***Smoothing and Regression: Approaches, Computation, and Application***". Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.
6. Wand, M. P., & Jones, M. C., (1995), "***Kernel Smoothing***". Chapman & Hall.
7. Wang, B, Shi, W., & Miao, Z., (2014), "***Comparative Analysis for Robust Penalized Spline Smoothing Methods***". Hindawi Publishing Corporation, Mathematical Problems in Engineering, Vol. 2014**.**
8. Wu, Hulin, & Zhang, Jin-Ting, (2006), "***Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling***

Approaches". Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.

9. Yao, F., & Lee, Thomas C.M., (2008), "***On Knot Placement for Penalized Spline Regression***". Journal of the Korean Statistical Society, 37.

Determination of the Best Smoothing Technique to Estimate a Nonparametric Regression Model with Practical Application

Abstract

Nonparametric regression technique represents a different way in regression analysis of parametric regression technique, but it does not mean that the use of one method prevent the use of the other. The methods of nonparametric regression can be used to assess the legality of supposed nonparametric model and vice versa, and matching the form of regression curve that we have it from nonparametric regression techniques may suggests the appropriate parametric regression model to be used in future studies.

The research contains the use of some Smoothing methods, Local Polynomial Kernel (LPK), Spline Regression (SR) and Penalized Spline (PS), to estimate a nonparametric regression model, in order to get that we depends on the results of a simulation study that described in experimental side. The comparison was made between the functions that were used in smoothing by using (MAE) and (MSE) criteria to reach the best estimator that represents the data that generated by simulation. In practical side we used real data from Iraqi Stock Market from Gulf Insurance Co.