مجلة جامعة تشرين للبحوث والدراسات العلمية – سلسلة العلوم الأساسية المجلد (40) العدد (4) 2018
Tishreen University Journal for Research and Scientific Studies - Basic Sciences Series Vol. (40) No. (4) 2018

تطبيق طرائق الانحدار اللاوسيطي كبدائل لطرائق الانحدار الخطي (دراسة مقارنة)

الدكتور محمد مزيد دريباتي*
الدكتور احمد يونسو**
ديمه احمد الشاخ***

(تاريخ الإيداع 13 / 11 / 2017. قُبِل للنشر في 18 / 7 /2018)

□ ملخّص □

تفرض طرائق الانحدار الخطية قيوداً شديدة على نماذج الانحدار وخاصة على حدود الخطأ حيث تفترض أنها مستقلة وتتبع التوزيع الطبيعي وهذا قد لا يتحقق في كثير من الدراسات مما يؤدي الى انحياز لا يمكن إهماله عن النموذج الفعلي مما يؤثر على مصداقية الدراسة.

يقدم هذا البحث مسألة تقدير دالة الانحدار باستخدام مقدّري النواة ناداريا واتسون و الجوارات الـ k الأكثر قرباً اللاوسيطيين كبدائل لمقدّرات الانحدار الخطية الوسيطية من خلال دراسة محاكاة على نموذج مفروض حيث قمنا بإجراء دراسة مقارنة بين هذه الطرائق باستخدام الحزمة الإحصائية **R** بغية معرفة أفضل هذه المقدّرات حيث تم استخدام معيارMSE متوسط مربعات الخطأ (Mean Squares Errors) لمعرفة المقدّر الأفضل. كما تشير نتائج دراسة المحاكاة إلى فعالية وكفاءة المقدّرات اللاوسيطية في تمثيل دالة الانحدار بالمقارنة مع مقدّرات الانحدار الخطية كما تشير إلى تقارب أداء هذين المقدّرين.

الكلمات المفتاحية: الانحدار الخطي،الانحدار اللاوسيطي، مقدّر النواة ناداريا واتسون، مقدّر الجوارات الـ k الأكثر قربًا، طريقة التحقق المتبادل.

* أستاذ مساعد في قسم الإحصاء الرياضي–كلية العلوم– جامعة تشرين–اللاذقية–سورية.
** مدرس في قسم الإحصاء الرياضي– كلية العلوم–جامعة دمشق–دمشق– سورية.
*** طالبة دراسات عليا(ماجستير) قسم الإحصاء الرياضي– كلية العلوم–جامعة تشرين–اللاذقية–سورية.

مجلة جامعة تشرين للبحوث والدراسات العلمية – سلسلة العلوم الأساسية المجلد (40) العدد (4) 2018
Tishreen University Journal for Research and Scientific Studies - Basic Sciences Series Vol. (40) No. (4) 2018

Applied Nonparametric Regression Methods as Alternatives to the Linear Regression Methods (Comparative Study)

Dr.Mohamed M. Deribati*
Dr. Ahmad Younso **
Dema A. Al-shakh ***

(Received 13 / 11 / 2017. Accepted 18 / 7 /2018)

□ ABSTRACT □

Linear regression methods impose strong constraints on regression models, especially on the error terms where it assumes that it is independent and follows normal distribution, and this may not be satisfied in many studies, leading to bias that cannot be ignored from the actual model, which affects the credibility of the study.

We present in this paper the problem of estimating the regression function using the Nadarya Watson kernel and k- nearest neighbor estimators as alternatives to the parametric linear regression estimators through a simulation study on an imposed model, where we conducted a comparative study between these methods using the statistical programming language R in order to know the best of these estimations. Where the mean squares errors (MSE) was used to determine the best estimate.

The results of the simulation study also indicate the effectiveness and efficiency of the nonparametric in the representation of the regression function as compared to linear regression estimators, and indicate the convergence of the performance of these two estimates.

Key words: Linear regression, Nonparametric regression, Nadaraya-Watson kernel estimator, k-nearest neighbor estimator, Cross validation method

* **Assistant Prof, Depart. Of Mathematical Statistics, Faculty of Science, Tishreen University, Lattakia, Syria.**
** **Professor, Depart. Of Mathematical Statistics, Faculty of Science, Damascus University, Damascus, Syria.**
*** **Postgraduate student, Depart. Of Mathematical Statistics, Tishreen University, Lattakia, Syria.**

مقدّمة:

يلعب تحليل الانحدار دوراً كبيراً في دراسة ظواهر مهمة في مختلف مجالات الحياة وذلك من خلال تحليل هذه الظواهر والتنبؤ بنتائجها المستقبلية. يعبّر نموذج الانحدار عن علاقة إحصائية بين متغير الظاهرة المدروسة والذي يسمى بالمتغير التابع (Dependent variable) مع متغيرات أخرى تسمى بالمتغيرات المستقلة (Independent variables) من جهة ثانية نرمز بـ Y لمتغير التابع وX لمتجه المتغيرات المستقلة. ننطلق في تحليل الانحدار من المتجه العشوائي (X,Y) الذي يأخذ قيمه في $\mathbb{R}^d \times \mathbb{R}$ ونفترض أن أي تغير في قيمX يؤدي إلى تغير في قيم Y وفق منهج معين، لذلك فإنه من المهم تقدير شكل و قوّة العلاقة المحتملة بين المتغير التابع وبقية المتغيرات المستقلة ، كما أن معرفة شكل هذه العلاقة يمّكن الباحث من التنبؤ بقيم جديدة للمتغير التابع من خلال معرفة قيم جديدة للمتغيرات المستقلة.

لتحليل الانحدار تطبيقات واسعة في شتى المجالات فعلى سبيل المثال قد يهتم الباحثون في الحقل **الطبي** بدراسة شكل وقوة العلاقة بين ارتفاع ضغط الدم وارتفاع نسبة الكولسترول في الدم أو شكل وقوة العلاقة المتوقعة بين أمراض القلب ونسبة الشحوم في الدم.

لنفرض أنه لدينا:

$$Y = m(X) + \varepsilon \qquad (1)$$

لتكن $\{(X_i, Y_i)\}_{i=1}^{n}$ عينة عشوائية للمتجه العشوائي (X, Y) ولنكتب نموذج الانحدار بالشكل:

$$Y_i = m(X_i) + \varepsilon_i \quad ; \qquad i = 1, .., n \qquad (2)$$

حيث تمثل $m(.): \mathbb{R}^d \longrightarrow \mathbb{R}$ دالة الانحدار حيث $m(X) = E(Y|X)$ و $\{\varepsilon_i\}_{i=1}^{n}$ متتالية حدود الخطأ العشوائي، تكون غير مرتبطة ومتماثلة التوزيع بمتوسط صفر وتباينσ^2. سنفترض في الدراسات التطبيقية أن X تخضع لسيطرة الباحث و أن التغير العشوائي في النموذج (1) ناتج فقط عن حدود الخطأ العشوائي. حيث نكتب النموذج بالشكل:

$$Y_i = m(x_i) + \varepsilon_i$$

أي أننا استبدلنا X_i بـ x_i للدلالة على خضوع X لسيطرة الباحث، حيث نكتب في هذه الحالة:

$$m(x) = E(Y|X = x) \qquad (3)$$

في أي نموذج انحدار نسعى إلى تقدير دالة الانحدار $m(x)$ وإجراء تنبؤات باستخدام النموذج المقدّر. يمكن تقدير دالة الانحدار $m(x)$ بعدّة طرائق منها وسيطية ومنها غير وسيطية[1].

إنَّ تقنيات تحليل الانحدار في النماذج الوسيطية تمثل طريقة للاستدلال عن $m(x)$، حيث يتم استخدامها عندما توجد معلومات قليلة عن شكل $m(x)$، والاستدلال عن $m(x)$ معادل للاستدلال عن عدد منته من الوسطاء نلخصها في المتجه $\boldsymbol{\beta}$، حيث يتم تقدير $\boldsymbol{\beta}$ باستخدام إحدى طرائق التقدير التقليدية كطريقة المربعات الصّغرى الاعتيادية (Ordinary Least Squares) أو طريقة الإمكانية العظمى (Maximum Likelihood) أو أسلوب بايز (Bayesian) في حالة توّفر معلومات أولية عن الوسطاء. إنَّ نتيجة التقدير هو منحني يُختار من أسرة منحنيات ليطابق البيانات. إنّ النموذج الوسيطي يقيّد بشروط عديدة حيث يفترض أن العينة تأتي من مجتمع معيّن له أسرة معروفة من التوزيعات مثل الأسرة الطبيعية (Gaussian). لكن هذه الافتراضات غالباً ما تكون قوية جداً لأن التوزيع الوسيطي المفترض لا يكون بالضرورة التوزيع الفعلي للمسألة المراد حلها [2]، إذ أنّ الافتراض الخاطئ للتوزيع يؤدي إلى استنتاجات غير صحيحة وتقديرات غير متسقة.

التّقنية الأخرى في مطابقة المنحنيات للبيانات هي تقنيات الانحدار اللاوسيطية، حيث أنّ هذه التّقنيات التي تقوم على تقدير دالة الانحدار ككل تعطي مرونة أكبر في التعامل مع البيانات لاسيما أنها لا ترتكز على قيود مسبقة شديدة كما هو الحال في الانحدار الوسيطي [3]. من أهم الطّرائق اللاوسيطية في تقدير دالة الانحدار نذكر طريقة النواة (Kernel Method) ، طريقة الجوّارات الـ k الأكثر قرباً (k-Nearest Neighbors) وطريقة الشرائح الممهدّة (Smoothing Splines) وغيرها.

أهمية البحث وأهدافه:

يهدف هذا البحث إلى تناول طريقتي النّواة و الجوّارات الـ k الأكثر قرباً اللاوسطيتين بغيّة معرفة مدى فعالية الطّرائق اللاوسيطية كبدائل للطرائق الوسيطية وسنرّكز تحديداً على نموذج الانحدار الخطي التقليدي عندما لا تتحقق الشروط الأساسية لهذا النموذج، ومن ثم إيجاد أفضل هذه الطرائق من خلال تطبيقها على نموذجي محاكاة باستخدام الحزمة الاحصائية **R**.

طرائق البحث ومواده:

سوف نطرح في هذه الورقة طريقة الانحدار الخطي التقليدية كإحدى الطرائق الوسيطية وكل من طريقتي النواة والجوارات الـ k الأكثر قرباً كطريقتين غير وسيطيتين. حيث سنقوم بصياغة نموذج محاكاة للتحقق من أداء هذه الطرائق باستخدام مقياس متوسط مربعات الخطأ **(MSE)** ، فضلاً عن مقارنة تقارب هذه الطرائق من المنحني الحقيقي لدالة الانحدار من خلال الرسوم **(Plots)** التي تعتمد على نتائج مستخلّصة من تجارب المحاكاة.

قبل البدء بدراسة المقارنة نقدّم فيما يلي طرائق الانحدار المستخدمة في هذا البحث:

الانحدار الخطي: Linear regression

في الانحدار الخطي نفترض أن:

$$m(x) = \beta_0 + x^T\boldsymbol{\beta} + \varepsilon \quad ; x \in \mathbb{R}^d \text{ و } \boldsymbol{\beta} \in \mathbb{R}^d$$

و ε تمثل حدود الخطأ العشوائي وتتبع التوزيع الطبيعي بحيث $E(\varepsilon) = 0$ و $var(\varepsilon) = \sigma^2$.

يتم استخدام طريقتي المربعات الصغرى أو الإمكانية العظمى لتقدير $\boldsymbol{\beta}$ وبالتالي تقدير $m(x)$، فمن أجل عينة للمتجه (X, Y) نحصل على تقدير لـ $\boldsymbol{\beta}$ و ليكن $\widehat{\boldsymbol{\beta}}$ ومنه يكون $\widehat{m}(x) = x^T\widehat{\boldsymbol{\beta}}$.

الانحدار اللاوسيطي: Nonparametric regression

ليكن $(\Omega, \mathcal{F}, P)$ فضاء احتمالي و (X, Y) متجهاً عشوائياً مستمراً معرفاً على هذا الفضاء و يأخذ قيمه من $\mathbb{R}^d \times \mathbb{R}$ و يملك كثافة احتمالية $f_{(X,Y)}(.)$ مجهولة. و لتكن الدالة $m(x)$ المذكورة في (3) هي دالة انحدار Y على X عندما $X = x$ أي أن $m(x) = E(Y|X = x)$. لنفرض أن $m(x)$ مجهولة ونسعى إلى تقديرها باستخدم عينات عشوائية للمتجه (X, Y).

مقدّر النواة ناداريا واتسون: Nadaraya-Watson kernel estimator

انطلاقاً من تعريف $m(x)$ يمكن كتابتها على الشكل التالي:

$$m(x) = E(Y|X = x)$$

$$= \int_{\mathbb{R}} yf(y|x)\,dy = \frac{\int_{\mathbb{R}} yf(y,x)\,dy}{f(x)} \qquad (4)$$

اقترح كل من ناداريا[4] ، واتسون[5] تقدير $m(x)$ باستبدال دالتي الكثافة $f(x,y)$ و $f(x)$ بمقدّراتها $\hat{f}(x,y)$ و $\hat{f}(x)$ بالترتيب باستخدام طريقة النواة لتقدير دالة الكثافة التي قدمت من قبل روزنبلات[6].

إن ناتج تقدير النواة لدالة الكثافة$f(x,y)$ هو

$$\hat{f}(x,y) = \frac{1}{nh^d h_y}\sum_{i=1}^{n} K_x\left(\frac{x-X_i}{h}\right)K_y\left(\frac{y-Y_i}{h_y}\right) \quad ; x \in \mathbb{R}^d$$

حيث$K: \mathbb{R}^d \longrightarrow \mathbb{R}$ دالة النواة وتعرف على انها دالة متناظرة موجبة وتكاملها موجبة تماماً بالنسبة لقياس لوبيغ. $h = h(n),\ h_y = h_y(n)$ تدعى وسطاء التمهيد أو النافذة (Bandwidth or window) وهي أعداد حقيقية موجبة تماماً تحقق أن $h \longrightarrow 0, h_y \longrightarrow 0$ عندما $n \longrightarrow \infty$ واختصاراً نقول h نافذة.

للحصول على مقدّر $m(x)$ لدينا:

$$\int_{\mathbb{R}} y\hat{f}(x,y)\,dy = \frac{1}{nh^d h_y}\sum_{i=1}^{n} K_x\left(\frac{x-X_i}{h}\right)\int_{\mathbb{R}} yK_y\left(\frac{y-Y_i}{h_y}\right)dy$$
$$= \frac{1}{nh^d}\sum_{i=1}^{n} K_x\left(\frac{x-X_i}{h}\right)\int_{\mathbb{R}}(Y_i + h_y s)K(s)\,ds \quad ; s \in \mathbb{R}$$
$$= \frac{1}{nh^d}\sum_{i=1}^{n} K_x\left(\frac{x-X_i}{h}\right)Y_i$$

حيث أن السطر الثاني من المساواة ينتج من تغيير المتحولات و السطر الثالث من حقيقة أن دالة النواة متناظرة K وتحقق $\int_{\mathbb{R}^d} uK(u)\,du = 0$ ، $\int_{\mathbb{R}^d} K(u)\,du = 1$.

كما يعطى مقدّر النواة لـ $f(x)$ كالتالي:

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K_x\left(\frac{x-X_i}{h}\right)$$

بتبديل توابع الكثافة المجهولة في العلاقة (4) بمقدّراتها ينتج لدينا مقدّر النواة لناداريا واتسون [13]:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \qquad (5)$$

كما يمكن كتابة العلاقة (5) بالشكل التالي[3]:

$$\widehat{m}(x) = \sum_{i=1}^{n} w_{ni}(x)\,Y_i \qquad (6)$$

حيث أن $w_{ni}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$ تحقق $\sum_{i=1}^{n} w_{ni}(x) = 1$. تسمى w_{ni} أوزاناً وبالتالي فإن $\widehat{m}(x)$ هو متوسط موزون لـ Y_i.

بملاحظة أن المقدّر خطي بالنسبة للمشاهدات $\{Y_i\}$ فإنه يدعى أيضا بالممهد الخطي، ولكن ليس بالضرورة أن يكون تابع انحدار خطي.

من الجدير بالذكر أن وسيط التمهيد h يلعب دورا مهماً في أداء مقدّر النواة. كما أن هناك العديد من النوى المستخدمة و أن اختيار إحدى هذه النّوى لا ينبغي أن يؤثر على تقديرات دالة الانحدار وعلى النتائج بشكل عام[3]. من أكثر النوى الشهيرة المستخدمة مع مقدّر النواة في تحليل الانحدار النواة الطبيعية و نواة إبانيشنكوف. سوف نستخدم في هذه الورقة النواة الطبيعية فقط و التي تعطى بالعلاقة التالية :

$$K(u) = (2\pi)^{-\frac{d}{2}}.exp\left(-\frac{\|u\|^2}{2}\right) \ ; \quad \forall\, u \in \mathbb{R}^d \qquad (7)$$

حيث أن الرمز $\|.\|$ هو النظيم الاقليدي والذي يعرف بالشكل :

$$\|u\| = \sqrt{u_1^2 + u_2^2 + \ldots + u_n^2} \quad (8)$$

طريقة الجوارات الـ k الأكثر قربا: k- Nearest Neighbor Estimation

لقد رأينا أن مقدّر النواة لدالة الانحدار هي عبارة عن المتوسط الموزون للمشاهدات Y_i في جوار مثبت لـ x. بدلا من هذا الجوار الثابت فإن مقدّر الجوارات الـ k الآكثر قرباً يستخدم جوارات متغيرة. أي أن قيم Y التي تستخدم في حساب المتوسط هي تلك القيم المقابلة لقيم X المشاهدة التي هي في حيز الجوار k الأكثر قرباً إلى النقطة x في المسافة الإقليدية التي نرغب بتقدير $m(.)$ عندها [10]. و بالتالي فإنه يمكن صياغة مقدّر الجوارات الـ k الآكثر قرب كالتالي [11−3]:

$$\widehat{m}_k(x) = \sum_{i=1}^{n} w_{ni}(x)\, Y_i$$

حيث أن الأوزان $\{w_{ni}(x)\}_{i=1}^{n}$ تعرف كالتالي:

$$w_{ni}(x) = \begin{cases} 1/k & ; i \in \boldsymbol{J_x} \\ 0 & ; \text{خلاف ذلك} \end{cases} \quad (9)$$

حيث $\boldsymbol{J_x}$ تمثل مجموعة الأدلة i التي تحقق أن X_i هي إحدى المشاهدات الـ k الأقرب الى x :

$$\boldsymbol{J_x} = \{i\,; X_i \; is\; one\; of\; the\; k\; nearest\; observation\; to\; x\;\}$$

حيث $k = k(n)$ متتالية من القيم الصحيحة الحقيقية. k تمثل وسيط التمهيد لمقدّر الجوارات وتحقق أن

$$k \to \infty\ ,\quad k/n \to 0\ \ as n \to \infty \quad (10)$$

حيث أن زيادة k نسبياً ضمن الشرط (10) تجعل المقدّر انعم.

وفقاً لروزنبلات[6] ، نادريا[4] و واتسون[5]هناك ارتباط بين مقدّر النواة و مقدّر الجوارات الـ k الآكثر قرباً . فإذا فرضنا أن $R = R_n(x)$ هو نصف قطر الجوار الذي مركزه x ويحوي الـ k مشاهدة لـ X الأكثر قرباً من x فإنه يتم الحصول على مقدر الجوارات من مقدّر النواة بتبديل النافذة h بـ R في مقدّر النواة مع استخدام النواة المنتظمة $K(u) = \frac{1}{2}\boldsymbol{I}(\|u\| \le 1)$، حيث $\boldsymbol{I}\{.\}$ هو تابع الإشارة في المجموعة و $\|.\|$ هو النظيم الأقليدي المعرف في (8). في هذه الحالة يصبح لدينا:

$$\widehat{m}_k(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{R}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{R}\right)}$$

$$= \frac{1}{k}\sum_{i=1}^{n} Y_i I\{\|x - X_i\| \le R\} \quad (11)$$

وبالتالي يمكن اعتبار $\widehat{m}_k(x)$ على انها مقدّر النواة المتضمن نافذة تعتمد على البيانات.

اختيار وسيط التمهيد: Smoothing parameter selection

في طريقة النواة يعتبر اختيار نصف قطر الجوار هو جزءاً اساسياً من تقريب دالة الانحدار اللاوسيطية إلى الدالة الأصلية. وللحصول على التقريب الملائم يجب إيجاد الطريقة المثلى لغرض الموازنة بين التحيز والتباين بحيث يكون الخطأ أقل ما يمكن.

إن اختيار وسيط التمهيد صغيراً سيؤدي إلى زيادة في تحيز المقدّر بينما اختيار كبيراً سيؤدي إلى زيادة التباين. لذلك من المهم إيجاد طريقة تضبط كلاً من التحيز والتباين معاً في آن واحد.

أما في طريقة الجوارات الـ k الأكثر قرباً فإن وسيط التمهيد هو عدد مجاورات x الأكثر قرباً والتي عددها k. هنا أيضاً اختيار k يؤثر في تغير كل من التحيز والتباين وينبغي ضبطهما معاً.

لاختيار كلاً من h و k سوف نستخدم **طريقة التحقق المتبادل (Cross Validation)** لتقييم نتائج الانحدار والنافذة المثلى. سوف يتم الحصول على القيمة الأفضل لـ k و h من خلال تقليل معيار التحقق المتبادل (CV Criterion) حيث تقوم الفكرة الرئيسية على حذف المشاهدة j ثم حساب مقدّر دالة الانحدار عند (X_j, Y_j) بالاعتماد على الـ$(n-1)$ مشاهدة الباقية من مشاهدات العينة [13-11-7].

فيكون من أجل طريقة النواة :

$$\hat{m}_j(X_j) = \frac{\sum_{i\neq j} Y_i K\left(\frac{X_j - X_i}{h}\right)}{\sum_{i\neq j} K\left(\frac{X_j - X_i}{h}\right)} \qquad (12)$$

ومن ثم يتم مقارنة (X_j) $\hat{m}$ مع Y_i من خلال مربع الفرق بينهما.

إن النافذة الأمثلية $\hat{h}_{op}$ هي هي القيمة التي تبلغ عندها دالة التحقق المتبادل التالية:

$$\mathrm{CV(h)} = \frac{1}{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\left(\mathrm{Y_j} - \hat{\mathrm{m}}_{\mathrm{j}}(\mathrm{X_j})\right)^2 \qquad (13)$$

قيمته الصغرى.

في طريقة الجوارات الـk الأكثر قرباً يكون:

$$\hat{\mathrm{m}}_{\mathrm{j}}(\mathrm{X_j}) = \sum_{\mathrm{i}\neq\mathrm{j}} \mathrm{w_{ni}}(\mathrm{X_j})\mathrm{Y_j} \qquad (14)$$

و

$$\mathrm{CV(h)} = \frac{1}{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\left(\mathrm{Y_j} - \hat{\mathrm{m}}_{\mathrm{j}}(\mathrm{X_j})\right)^2 \qquad (15)$$

النتائج والمناقشة:

سوف يتم استخدام أسلوب المحاكاة بطرائق مونتي كارلو بهدف مقارنة الطرائق المذكورة سابقاً من خلال نموذجي محاكاة ومن ثم استعراض أفضل هذه الطرائق باستخدام الحزمة الإحصائية **R** [10-9]. من أجل مقارنة الطرائق السابقة سيتم اعتماد معيار متوسط مربعات الخطأ (Mean Squares Error) حيث يعطى بالعلاقة التالية:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i} - \hat{\mathrm{m}}(\mathrm{X_i}))^2 \qquad (16)$$

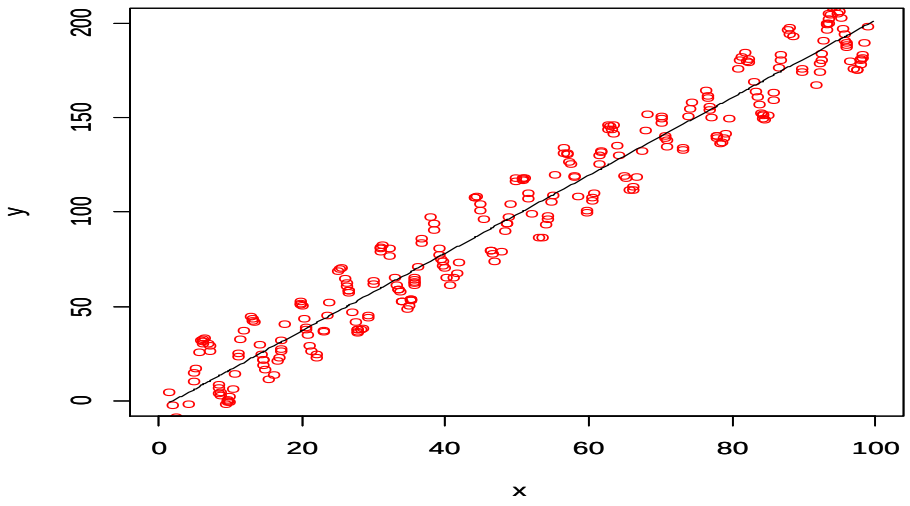
لقد أجريت دراسة المحاكاة من أجل مقارنة أداء المقدّرين المذكورين مع مقدّر الانحدار الخطي التقليدي.

أولا سيتم استخدام نموذج الانحدار المعطى بالصيغة التالية :

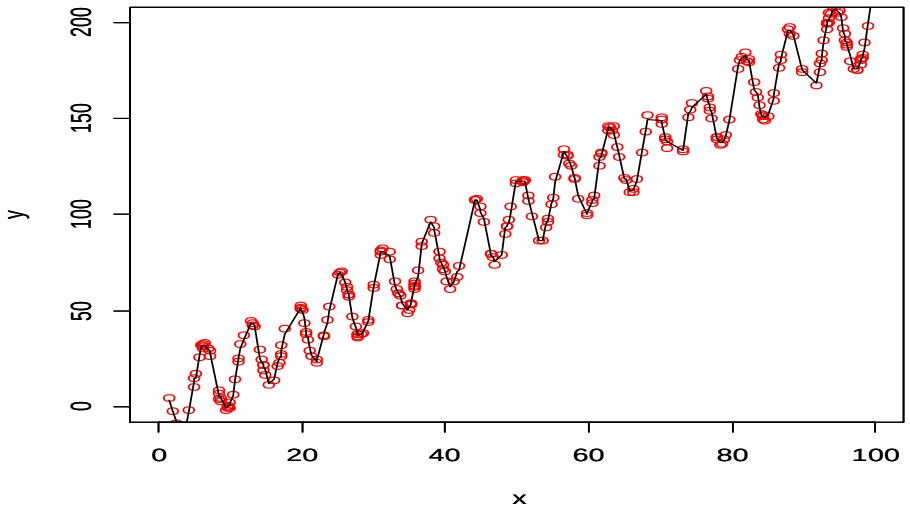
$$\mathrm{Y_i} = 2\mathrm{X_i} + 20\cos(\mathrm{X_i}) + \varepsilon_{\mathrm{i}} \qquad (17)$$

حيث تم سحب $\mathrm{X_i}$ من توزيع منتظم على المجال $[1,100]$ و ε_{i} لها توزيع طبيعي بمتوسط صفر وتباين 1. من أجل ذلك ولدنا عينات من القياسات (100,250,500) على الترتيب و من ثم قمنا بحساب قيمة النافذة h لمقدّر النواة نادريا واتسون، وقيمة الجوار k في مقدّر الجوارات الأكثر قرباً باستخدام طريقة التحقق المتبادل. حيث كان عدد مرات تكرار المحاكاة من أجل كل مقدّر 100.

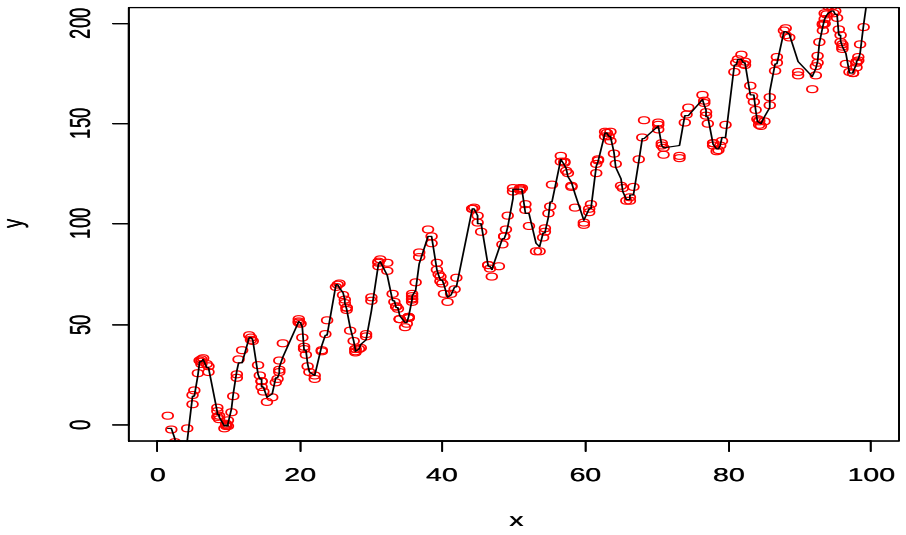
الرسم البياني لدالة الانحدار الحقيقية ومقدّرات دالة الانحدار على عينة مؤلفة من 300 مشاهدة موضح في الشكل (1) و (2) حيث يتضح من الشكل (1) أن لدالة الانحدار شكل خطي و أن التقدير الخطي لدالة الانحدار لا يلائم البيانات بشكل جيد بينما يتضح في الشكل (2) و (3) أن المقدّرات اللاوسيطية تبدو أكثر ملائمة للبيانات كما يبين تطابق بين مقدّري النواة والجوارات الـk الأكثر قرباً. حيث تم حساب **h** و**k** الأمثليتان باستخدام طريقة التحقق المتبادل.



شكل(1): الرسم البياني لدالة الانحدار الحقيقية مع المقدّر الخطي



شكل(2): مقدّر النواة لدالة الانحدار مع نافذ $h = 0.24$



شكل(3): مقدّر الجوارات الـ k الأكثر قرباً مع $k = 4$

باستخدام لغة البرمجة الإحصائية **R**، من أجل كل عينة سوف نقوم بتقدير دالة الانحدار باستخدام الأسلوب الخطي و طريقتي النواة و الجوارات الـ k الآكثر قرباً و من ثم احتساب متوسط مربعات الخطأ (MSE) المرتبط بكل مقدّر من أجل المقارنة بين الطرائق المذكورة. وتم توضيح النتائج في الجدول (1):

الجدول(1): قيم متوسط مربعات الخطأ المقابل لكل طريقة مع عينات من القياس (100,250,500)

حجم العينة N	طريقة الانحدار الخطية	طريقة النواة	طريقة الجوارات الـ k الآكثر قربا
100	194.4988	5.369254*	6.973538
250	198.246	2.376522*	2.861382
500	170.8895	1.338909*	2.166645

***قيمة (MSE) الأصغر في كل سطر**

ثانياً سنعيد المحاكاة السابقة على نموذج انحدار آخر يعطى بالشكل التالي:

$$Y_i = 1 - X_i + e^{-200\left(\frac{X_i - 1}{2}\right)^2} + \varepsilon_i \qquad (18)$$

في هذا النموذج تم سحب X_i من توزيع منتظم على المجال $[0,3]$. و ε_i كان لها توزيع طبيعي بمتوسط صفر وتباين (0.2). الجدول (2) يوضح النتائج التي توصلنا إليها من أجل هذا النموذج.

الجدول(2): قيم متوسط مربعات الخطأ المقابل لكل طريقة من اجل مع عينات من القياس (100,250,500)

حجم العينة N	طريقة الانحدار الخطية	طريقة النواة	طريقة الجوارات الـ k الآكثر قربا
100	1.747221	0.032228*	0.034771
250	1.765712	0.036259*	0.037146
500	1.781233	0.037294*	0.037277

***قيمة (MSE) الأصغر في كل سطر**

يبين الجدول (1) و (2) أنه من أجل جميع أحجام العينات، أن مقدّرات النواة تملك قيم لـ **MSE** اصغر من القيم المرتبطة بالمقدّرات الخطية ومقدّرات الجوارات الـ k الآكثر قرباً لكل عينة، كما يتضح تقارب قيم **MSE** في مقدّري النواة والجوارات الـ k الآكثر قرباً.

الاستنتاجات والتوصيات:

تم تقديم كلاً من مقدّري النواة لناداريا واتسون والجوارات الـ k الآكثر قرباً كبدائل لمقدّرات الانحدار الخطي الوسيطية. نتائج دراسة المحاكاة التي أجريت لتقييم أداء مقدّر الانحدار الخطي مع المقدّرين اللاوسيطيين المذكوران بينت أن مقدّر النواة ناداريا واتسون يعطي تقديرات أفضل من مقدّر الانحدار الخطي و مقدّر الجوارات الـ k الآكثر قرباً، كما بينت الدراسة أن كلاً من المقدّرين اللاوسيطيين لهما تقريبا نفس الأداء. وأخيراً يمكن القول أن المقدّرات اللاوسيطية هي الأفضل في تقدير دالة الانحدار.

المراجع

[1] WAND, M.P.; JONES, M.C. Kernel Smoothing, Chapman & Hall (1995) 224.
[2] ZHANG, J.S.; HUANG, X.F.; ZHOU, C.H. An Improved Kernel Regression Method Based on Taylor Expansion, Applied Mathematics and Computation, vol. 193 (2007) 419-429
[3] HÄRDLE, W. Applied Nonparametric Regression. Cambridge University Press, Cambridge (1990) 433.
[4] NADARAYA, E. A. On estimating regression. Theory Probab. Appl. 9 (1964) 141-142.
[5] WATSON, G. S.Smooth regression analysis. Sankhyà Ser. A 26 (1964) 359-372.
[6] ROSENBLATT, M. Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 21 (1956) 832-837.
[7] STONE, C. J. Cross-validatory choice and assessment of statistical predictions (with discussion).Journal of Royal Statistical Society 36 (1974)111–47.
[8] MACK, Y. P.; ROSENBLATT M. Multivariate k-nearest neighbor density estimates. Journal of Multivariate Analysis 9 (1979)1–19
[9] COHEN, Y.; COHEN, J.Y.Statistics and Data with R: An Applied Approach Through Examples. A John Wiley & Sons, Ltd. (2008)
[1o] CRAWLEY J. M. The R book. 2nd. ed., John Wiley & Sons, Ltd.,(2013) 1060.

[11] Györfi, L.; Kohler M.; Krzyzak, A.; Walk, H., A Distribution-Free Theory of Nonparametric Regression, Springer-Verlag, New York,2002,664.
[12] Hardle, W.; Muller, M.; Sperlich, S.;Werwatz,A., Nonparametricand Semiparametric Models, ser. Springer Series in Statistics. NewYork: Springer, 2004.
[13] Erdős, P.; Ormos, M.; Zibriczky, D. (2011) Non-parametric and semi-parametric asset pricing. Economic Modelling 28:(3) pp.1150-1162.