استخدام مقدر (Nadaraya-Watson) كيرنل في تقدير دالة الانحدار اللامعلمي

م.م محمد عبد الحسين محمد

هيئة التعليم التقني

المعهد التقني الديوانية

الملخص

تعد النماذج اللامعلمية جزءاً مهماً من الاحصاء اللامعلمي وهي تختلف عن النماذج المعلمية (*Parametric models*) في بناء هيكلية النموذج ، فهي لاتعتمد على محددات او فروض سابقة (*prior specified*) ولكنها تعتمد بشكل مباشر واساسي على البيانات (*Data*) ،كما ان مصطلح النماذج اللامعلمية لا يعني أنها لا تتضمن على معلمات (*parameters*) ولكن طبيعة هذه المعلمات وعددها يكون بشكل مرن (*flexible*) وغير ثابت (*not fixed*) ،لذلك يطلق على هذه النماذج بنماذج حرة التوزيع (*distribution free*) .ان طرق الاستدلال اللامعلمي هي عمليات رياضية لاختبار الفرضيات الاحصائية والتي لا تشترط وجود فرضيات حول التوزيعات التكرارية للمتغيرات ، لذلك فهي تكون اقل قوة من الاختبارات المعلمية ولكنها اكثر حصانة (*robust*) في حالة انتهاك الفروض الاساسية او عدم تحققها . في هذا البحث تم تقدير دالة الانحدار اللامعلمي بشكل مباشر دون الاعتماد على معلمات محددة باستخدام مجموعة من الطرائق اللامعلمية وهي طرائق ندارياـواتسن وقد تم استخدام اسلوب المحاكاة في تطبيق طرائق التقدير وفي اجراء المقارنات .

المقدمة [2],[1]

ان نماذج الانحدار المعلمي *(parametric regression)* تصف العلاقة بين متغير الاستجابة مع متغير واحد او مجموعة من المتغيرات التوضيحية وتستخدم هذه النماذج عندما توجد معلومات عن شكل هذه العلاقة وتحقق مجموعة من الافتراضات وتتم عملية تحليل هذه النماذج بتقدير معلمات النموذج باستخدام أي طريقة تقدير مناسبة مثل MLE اوOLS ومن ثم تقدير دالة الانحدار وتكون نتائج التقدير هي منحنى يختار من مجموعة من المنحنيات ليطابق البيانات وان هذا الاختيار مقيد بشروط عديدة لمطابقة الاشكال المتوقعة ،أما الأسلوب الآخر في مطابقة المنحنيات للبيانات هو طرائق الانحدار اللامعلمي (*nonparametric* regression) هذه الطرائق تسمح بمرونة عالية في الاشكال الممكنة لمنحنى الانحدار والافتراض على هذه الطرائق هو ان دالة العلاقة يجب ان تكون قابلة للاشتقاق ،وان هذه الطرائق تعتمد بشكل رئيس على البيانات حيث ان نوع البيانات يفسر الشكل الفعلي لمنحنى الانحدار .

هدف البحث

ان هذا البحث يهدف الى تطبيق مجموعة من طرائق نداريا-واتسن لتقدير دالة الانحدار اللامعلمي والتي لا تعتمد على معلمات محددة ومن ثم اجراء المقارنة بين الطرائق المستخدمة باستخدام اسلوب المحاكاة ($simulation$) .

الانحدار اللامعلمي ***Nonparametric Regression***: [1],[2],[3]

ان نماذج الانحدار المعلمي التقليدي تكتب بالصيغة التالية

$$y_i = f(\beta_j\, x_i') + \varepsilon_i$$

حيث ان :

y : المتغير التابع (متغير الاستجابة)

$\beta_j = (\beta_1, \beta_2, ..., \beta_k)$:هو متجه بمعلمات النموذج

$x_i' = (x_1, x_2, ..., x_k)$: هو متجه بمشاهدات المتغير التوضيحي ($Observation$)

ε_i : الاخطاء العشوائية والتي يفترض ان تتوزع طبيعياً بمتوسط صفر وتباين ثابت σ^2

اما نماذج الانحدار اللامعلمي فهي تكتب بشكل عام بالشكل الآتي :

$$\begin{aligned} y_i &= f(x_i') + \varepsilon_i \qquad \text{.................} \; 1 \\ &= f(x_{i1}, x_{i2}, ..., x_{ik}) + \varepsilon_i \end{aligned}$$

ان الانحدار اللامعلمي هو احد اشكال تحليل الانحدار، لكنه لايعتمد على نموذج ثابت ذي معلمات محددة وانما هو دالة تنشأ بموجب معلومات مستندة الى البيانات ($data$) وذلك لأنها الشيء الاساسي الذي يستند اليه في بناء النموذج ويستخدم الانحدار اللامعلمي لتقدير دالة الانحدار $f(.)$ بشكل مباشر وبدون وجود أي صيغة محددة لها وبعيداً عن تقدير معلمات النموذج كما هو الحال في الانحدار المعلمي ، وان معظم طرائق الانحدار اللامعلمي تفترض ان $\varepsilon \sim \mathrm{NIID}(0, \sigma_\varepsilon^2)$ وان الدالة $f(.)$ هي دالة مستمرة ($Continuous\ Function$) وممهدة ($smoothing$).

انحدار كيرنل (*Kernel Regression*) [1],[4],[6]

ان انحدار كيرنل (*Kernel*) هو طريقة احصائية لامعلمية لتقدير دالة التوقع الشرطي الهدف منه ايجاد علاقة لا خطية بين ازواج المتغيرات العشوائية كما انه طريقة مبسطة لايجاد هيكلية او نمط البيانات بدون الحاجة الى انموذج معلمي عن طريق سلسلة من الاوزان ، توصف دالة الوزن بواسطة دالة الكثافة مع معلمة قياس التي تعدل حجم وشكل الاوزان ، دالة الوزن هذه تسمى K (*krenel*) وهي دالة كثافة احتمالية حقيقية محددة مستمرة ومتماثلة حول الصفر تكاملها يساوي واحد فعلى فرض انه لدينا مجموعة من المشاهدات لمتغيرين عشوائيين بشكل ازواج مرتبة وبالشكل الاتي :

$$D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$$

فان دالة الانحدار كيرنل هي:

$$f(x)=E(y\,|\,x) \qquad \ldots\ldots\ldots\ldots 2$$

حيث ان

Y: هو المتغير المعتمد (*depended variable*)

X: هو المتغير التوضيحي (*independed variable*)

وان سلسلة الوزن (w_i) لتقديرات كيرنل (kernel) كالاتي

$$w_i(x)=k_h(x-X_i)/\hat{f}_h(x)$$

حيث ان

h هو عدد موجب يمثل عرض الحزمة (*bandwidth*) .

$\hat{f}_h(x)$ تقدير دالة الكثافة

K_h تمثل دالة كيرنل .

- **مقدر كيرنل نداريا-واتسن (*Nadaraya-Watson*) [5[,[6]**

ان مقدر (*Nadaraya-Watson*) كيرنل للدالة f يمكن تعريفه بالصيغة الاتية :

$$\hat{f}_h(x)=\sum_{i=1}^{n} w_i(x)y_i \qquad \ldots\ldots\ldots\ldots\ldots 3$$

حيث ان w_i تمثل اوزان تحسب من الصيغة الاتية :

$$w_i(x) = \frac{k(\frac{x-x_i}{h})}{\sum_{i=1}^{n} k(\frac{x-x_i}{h})} \qquad ....4$$

تقدير الدالة باستخدام $kernel$ [4],[5]

تعريف : مقدر الدالة نوع $kernel$ يعرف كما في الصيغة الاتية:

$$\hat{f}_h = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k(\frac{x-x_i}{h}) \qquad ....5$$

ان الحصول على مقدر $kernel$ يعتمد على دالة كيرنل k (حيث ان هناك عدة دوال كيرنل يمكن اعتمادها) وعلى قيمة $bandwidth$ (حيث ان لكل دالة كيرنل هناك قيمة مثلى لـ h يمكن استخدامها) ، كما ان

$$\int_{\forall t} k(t)dt = 1 \qquad ....6$$

وذلك لان مقدر $kernel$ يمثل مقدر اصلي ($Bona\ fide$) .

تعريف : ان مقدر الدالة الاحتمالية $\hat{f}$ والذي يكون غير سالب وتكامله واحد يدعى مقدر اصلي ($Bona\ fide$) وبعبارة اخرى :

$$\hat{f}(x) \ge 0 \ , \quad x \in \chi$$

$$\int_{\forall x} \hat{f}(x)dx = 1$$

ولصعوبة الحصول على مقدر اصلي ($Bona\ fide$) غير متحيز يتم الاعتماد على ايجاد الاتي:

- مقدر ($Bona\ fide$) يمتلك اقل mse .
- متتابعة $\{\hat{f}_n\}$ من المقدرات ($Bona\ fide$) والتي تكون غير متحيزة بشكل محاذي ($asymptotically\ unbiased$)

$$E\left[\hat{f}_n(x)\right] \to f(x)\ ,\ x \in \chi \quad as\ n \to \infty$$

ولتحديد المقدر $\hat{f}$ يمكن الاعتماد على احد المعايير الاتية :

1- متوسط مربعات الخطأ التجميعي (*MISE*) (*Mean Integrated Squared Error*)

2- متوسط مربعات الخطأ التجميعي المحاذي (*AMISE*) *(Asymptotic Mean Integrated Squared Error Integrated)*

ان عرض الحزمة h (*bandwidth*) هو بمثابة ثابت التمهيد في المقدر $\hat{f}$ فعندما تكون قيمته صغيرة فان منحني الدالة يكون خشن (غير ممهد) (*rough currey*) اما عندما تكون قيمته كبيرة نسبياً فسوف يكون المنحني اكثر تمهيداً (*more smoother*). ونظراً لاهمية ثابت التمهيد h في الحصول على مقدر *kernel* ($\hat{f}$) لذلك يمكن تعريف متوسط مربعات الخطأ التجميعي المحاذي كدالة في h $\{AMISE_{\hat{f}_n}(h)\}$ وحسب الصيغة التالية:

$$AMISE_{\hat{f}_n}(h) = \frac{1}{4}\sigma_k^2 h^4 R(f) + \frac{S(k)}{n_h} \qquad ....7$$

حيث ان

K : تمثل دالة كيرنل بمتوسط $M_k = 0$ وتباين محدد هو

$$\sigma_k^4 = \int x^2 k(x)dx \quad , 0 < \sigma_k^2 < \infty \qquad ....8$$

وان

$R(f)$: تمثل مقياس الخشونة للدالة f والذي يحسب من

$$R(f) = \int (f(x))^2 dx \qquad .... 9$$

S(k) : تمثل تباين k اذ ان:

$$S(k) = \int k^2(x)dx \qquad .... 10$$

ان القيمة المثلى لـ h (Optimal bandwidth) والتي تؤدي الى تصغير قيمة $\{AMISE_{\hat{f}_n}(h)\}$ يمكن الحصول عليها من الصيغة الاتية:

$$h^* = \left(\frac{S(k)}{n\sigma_k^4 R(f)}\right)^{1/5} \qquad .... 11$$

حيث ان

h^* هو (Optimal bandwidth) .

ووفق ذلك فان متوسط مربعات الخطأ التجميعي المحاذي والذي يمكن استخدامه في التطبيقات العملية يحسب بالشكل الاتي:

$$h^* = \left(\frac{3}{4n}\right)^{1/5} \sigma \qquad .... 12$$

حدود الثقة لمقدر (*Nadaraya Watson*) [3],[4],[5]

ان حدود الثقة لمقدر (نادرايا- واتسن) باحتمال $(1-\alpha)$ هي :

$$\left.\begin{aligned} \ell_n &= \hat{f}(x) - q\hat{s}e(x) \\ u_n(x) &= \hat{f}(x) + q\hat{s}e(x) \end{aligned}\right] \qquad .... 13$$

حيث ان

$$\hat{S}e(x) = \hat{\sigma}\sqrt{\sum_{i=1}^{n} w_i^2(x)}$$

$$w = 3h \quad , h \text{ is a bandwidth}$$

$$\hat{\sigma}^2 = \frac{1}{2(n-1)}\sum_{i=2}^{n-1}(y_{i+1} - y_i)^2$$

$$q = \phi^{-1}\left(\frac{1+(1-\alpha)^{1/n}}{2}\right)$$

بعض دوال نداريا- واتسن (*Nadaraya Watson*)الشائعة: [5],[6].

سوف نتناول بعض دوال نداريا- واتسن الشائعة حيث تم اخذ القيم المثلى h

($Optimal\ bandwidth$) المحسوب وفق المعادلة رقم (11) في حساب كل من هذه الدوال ،واهم هذه الدوال هي :

1- دالة $Uniform$ وصيغتها هي :

$$\hat{f}(x)=\frac{1}{2} \qquad |u|\leq 1 \qquad \ldots\ldots\ldots\ldots\ldots 14$$

2- دالة Triangle وصيغتها هي:

$$\hat{f}(x)=(1-|u|) \qquad |u|\leq 1 \qquad \ldots\ldots\ldots\ldots\ldots 15$$

3- دالة $Epanechnikov$ وصيغتها هي :

$$\hat{f}(x)=\frac{3}{4}(1-u^2) \qquad |u|\leq 1 \qquad \ldots\ldots\ldots\ldots\ldots 16$$

4- دالة $Quadratic$ وصيغتها هي :

$$\hat{f}(x)=\frac{15}{16}(1-u^2)^2 \qquad |u|\leq 1 \qquad \ldots\ldots\ldots\ldots\ldots 17$$

5- دالة $Triweight$ وصيغتها هي:

$$\hat{f}(x)=\frac{35}{32}(1-u^2)^3 \qquad |u|\leq 1 \qquad \ldots\ldots\ldots\ldots\ldots 18$$

6- دالة $Gaussian$ وصيغتها هي :

$$\hat{f}(x)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2) \qquad |u|\leq 1 \qquad \ldots\ldots\ldots\ldots\ldots 19$$

7- دالة $Cosinus$ وصيغتها هي :

$$\hat{f}(x)=\frac{\pi}{4}Cos(\frac{\pi}{2}u\ ) \qquad |u|\leq 1 \qquad \ldots\ldots\ldots\ldots\ldots 20$$

حيث ان : $u=(\frac{x-x_i}{h})$ في جميع الصيغ اعلاه .

الجانب التجريبي:

تم استخدام المحاكاة ($Simulation$) في توليد المشاهدات الخاصة بالدراسة وحسب احجام العينات (n=25,50,75,100,120) ، وذلك بفرض ان الخطأ يتوزع حسب التوزيع الطبيعي بمتوسط مساوٍ الى

الصفر وتباين σ_e^2 ،وقد تم تطبيق طرائق نداريا-واتسن المذكورة في الجانب النظري ولكافة احجام العينات واعتمد معيار *mse* في المفاضلة بين طرائق التقدير،وحسب الخطوات الاتية:

1- توليد مشاهدات الخطأ العشوائي حسب التوزيع الطبيعي وذلك على فرض ان:

$$\varepsilon \sim \mathrm{N}(0,\sigma_\varepsilon^2)$$

2- افتراض دالة معينة بالمتغير العشوائي X .

3- حساب قيم للمتغير المعتمد Y حسب المعادلة رقم (1)

4- تطبيق طرائق نداريا-واتسن المذكورة في الجانب النظري لتقدير الدالة $f(x)$ أي حساب الدالة $\hat{f}(x)$ ، فمثلاً لتطبيق طريقة *Epanechnikov* الموصوفة في المعادلة رقم (16) يكون بالشكل الاتي:

$$\hat{f}\,(x) = \sum w_i(x) y_i$$

$$w_i(x) = \frac{k_e(u)}{\sum k_e(u)}$$

$$k_e(u) = k_e\left(\frac{x-x_i}{h_e}\right) = \begin{cases} \frac{3}{4}\left(1-\left(\frac{x-x_i}{h_e}\right)^2\right) & if \ \left|\frac{x-x_i}{h_e}\right| \le 1 \\ 0 & otherwise \end{cases}$$

حيث ان :

$k_e(u)$: دالة (نداريا-واتسن كيرنل) *Epanechnikov*

h_e: ثابت التمهيد (*bandwidth*) للدالة *Epanechnikov*

5- حساب معيار *mse* حسب الصيغة الاتية :

$$mse = \frac{\sum\left[f(x) - \hat{f}(x)\right]^2}{n-2}$$

6- المقارنة بين طرائق التقدير نداريا-واتسن حسب معيار *mse* والنتائج موضحة في الجدول رقم (1)

7- تم اعادة تجربة المحاكاة لكل حالة 500 مرة لضمان العشوائية .

جدول رقم (1)

يمثل قيم *mse* لدوال نداريا- واتسن عند احجام عينات مختلفة وقيمة مثلى لـ *h*

n	*mse*						
	Uniform	*Triangle*	*Epanech.*	*Quartic*	*Triweight*	*Gaussian*	*Cosinus*
25	9.914	8.566	7.509	7.972	7.592	3.827	8.261
50	2.763	2.244	1.567	1.919	1.673	0.825	2.060
75	2.021	1.332	1.122	1.208	1.083	0.170	1.247
100	1.633	2.286	2.594	2.042	1.303	0.025	1.282
120	0.325	0.325	0.700	0.415	0.168	0.021	0.223

تحليل النتائج

من خلال الجدول رقم (1) نلاحظ الاتي:

1- عند الحجم n=25 نجد ان قيم *mse* مرتفعة نسبياً لكافة طرق نداريا-واتسن مقارنة مع باقي احجام العينات ، كما نجد ان هناك افضلية واضحة لطريقة *Gaussian* في التقدير على حساب بقية الطرائق المستخدمة حيث كانت قيمة *mse* لهذه الطريقة 3.827 بينما كانت اقل قيمة لبقية الطرق الاخرى هي 7.509 لطريقة *Epanechenikov* .

٢- عند الاحجام n=50,75,100 وبالرغم من وجود تحسن كبير في اداء كافة الطرائق مع زيادة احجام العينات، الا اننا في نفس الوقت نجد ان هناك تذبب في اداء هذه الطرائق وهذا واضح من خلال قيم mse ، فتارة نجدها تنخفض مع زيادة حجم العينة وتارة اخرى نجد ان قيمه ترتفع مع زيادة حجم العينة ، وعموماً تبقى طريقة Gaussian هي الافضل فعند الحجم n=100 كانت قيمة mse هي 0.025 بينما اقل قيمة لبقية الطرائق هي 1.303 لطريقة Triweight.

٣- عند الحجم n=120 نجد ان هناك تحسناً كبيراً في اداء كافة الطرائق وهذا واضح من انخفاض قيم mse وهذا يعني ان هذه الطرق كفوءة في العينات الكبيرة

الاستنتاجات

1- ان لزيادة حجم العينة دور كبير في تحسن اداء طرق نداريا-واتسن وهذا ناتج من اعتماد هذه الطرائق على معلومات العينة فكلما زاد حجم العينة زادت دقة التقدير.

٢- ان طرائق نداريا- واتسن هي طرق كفوءة في التقديرات اللامعلمية في العينات الكبيرة وهذا واضح من خلال انخفاض قيم mse لكافة الطرائق مع زيادة عدد البيانات .

٣- بشكل عام نجد ان طريقة Cosinus هي افضل طرق نداريا-واتسن (التي تم دراستها) في التقدير.

التوصيات

١- التعرف على كفاءة طرق نداريا- واتسن وذلك بمقارنتها مع طرق لامعلمية اخرى مثل طريقة C.V (Cross Validation) او طريقة الانحدار الخطي الموضعي (L.L.R)

المصادر *References*

١- يوسف ، خلود يوسف خمو يوسف ،(٢٠٠٤) "مقارنة اساليب بيز مع طرائق اخرى لتقدير منحنى الانحدار اللامعلمي" اطروحة دكتوراه في الاحصاء ، كلية الادارة والاقتصاد ، جامعة بغداد

2- John Fox.(2002),"Nonparamtric Regression ",Appendix to an R and s-plus Companion to Appiied Regression.

3- John Fox.(2004),"Nonparamtric Regression",McMaster ,Hamilton,Canada. E-mail jfox@mcmaster.ca

4- Li,Qi;racine,Jeffrey s.(2007),"Nonparametric Econometrics:Theory and practice ",Princeton University.

5- M. Amalia and Ricardo Cao(2005),"Comparison of Nadaraya-Watson and local linear methods ",Uiniversiry de Vigo (Spain). E-mail : amaliajp@uvigo.es

6- Nageswara S.V.Rao.(1996)"Nadaraya-Watson Estimator for Sensor Fusion problem ",center for Engineering system advanced Research ,Oak Ridge National Laboratory.