مقدر Nadaraya-Watson اسلوب تمهيدي لتقدير دالة الانحدار

أ. د. ظافر حسين رشيد أ. م. د.مناف يوسف حمود

جامعة بغداد- كلية الادارة والاقتصاد- قسم الاحصاء

م. م. سعد كاظم حمزة

المستخلص:

أن استعمال النماذج المعلمية وما يتبعها من أساليب تقدير يتطلب وجود العديد من الشروط الأولية الواجب توافرها كي تمثل تلك النماذج المجتمع تحت الدراسة تمثيلا ملائما الأمر الذي دفع الباحثون إلى البحث عن نماذج أكثر مرونة من النماذج المعلمية وتمثلت هذه النماذج بالنماذج اللامعلمية .

في هذا البحث تم مقارنة ما يسمى بمقدر Nadaraya-Watson في حالة استعمال معلمة عرض حزمة ثابتة ومتغيرة وذلك من خلال أسلوب المحاكاة مستخدمين بذلك نماذج وحجوم عينات متنوعة. ومن خلال تجارب المحاكاة وضحت النتائج وللانموذجين الاول والثاني افضلية مقدر NW ذو المعلمة التمهيدية الثابتة ولجميع الحالات، اما للانموذج الثالث فقد اوضحت النتائج افضلية مقدرNW ذو المعلمة التمهيدية المتغيرة.

Nadaraya-Watson Estimator a Smoothing Technique for Estimating Regression Function

Abstract:

The using of the parametric models and the subsequent estimation methods require the presence of many of the primary conditions to be met by those models to represent the population under study adequately, these prompting researchers to search for more flexible models of parametric models and these models were nonparametric models.

In this manuscript were compared to the so-called Nadaraya-Watson estimator in two cases (use of fixed bandwidth and variable) through simulation with different models and samples sizes. Through simulation experiments and the results showed that for the first and second models preferred NW with fixed bandwidth for all cases, whether for the third model the results showed a preference NW estimator with variable bandwidth.

صفحات البحث 283-291

1- المقدمة:

تكمن فلسفة الإحصاء من حيث آلية التطبيق إلى محاولة نمذجة الظواهر المختلفة بنماذج اقرب ما يمكن إلى الواقع الفعلي. إن هذه النماذج تقاس درجة قوتها بحسب درجة تقاربها مع الخواص الاستدلالية الإحصائية وان هذه النماذج هي على أشكال وأنواع مختلفة فمنها الاحتمالي والذي يعتمد في صياغتها على الاحتمالات الصرفة ومنها السببي والذي يقوم على السبب ونتيجة السبب وتأتي في مقدمة هذه النماذج ما يعرف بنماذج الانحدار ،إذ تقوم نماذج الانحدار باستكشاف العلاقة ما بين السبب والذي يعرف إحصائيا بالمتغير التوضيحي وبين نتيجة السبب أو ما يعرف بالمتغير المعتمد أو التابع.

وتعد نماذج الانحدار من أهم أصناف النظرية الإحصائية وذلك لما قدمته للباحثين في شتى المجالات العلمية والإنسانية من حلول عملية لمشاكلهم . وبسبب تنوع مجالات عملها كان لابد من تنوع أشكالها هي الأخرى ويعود أسباب هذا التنوع لاختلافات طبيعة الموارد المتاحة للباحثين من طبيعة البيانات وكذلك طبيعة المشكلة تحت الدراسة ،وعلى هذا الأساس قسمت نماذج الانحدار إلى صنفين أساسين بحسب طبيعة البيانات وهي :

1. نماذج انحدار معلميه وتقوم على أيجاد العلاقة بين السبب والنتيجة من خلال عدد من النقاط الأساسية التي تصف طبيعة تلك العلاقة مثل نقطة تقاطع خط الانحدار مع المحور الصادي والممثلة بالمعلمة β_0 وكذلك مقدار الميول الحدية لأنموذج الانحدار والممثلة بالمعلمات $\beta_1, \beta_2, \ldots, \beta_K$ ،وان استعمال هكذا نوع من نماذج الانحدار يتطلب العديد من الشروط الأولية التي يجب توافرها كي تكون قراءة هذه النماذج قراءة صحيحة وسليمة.
2. نماذج انحدار لا معلميه والتي تقوم على إيجاد العلاقة بين السبب ونتيجة السبب من خلال منحنى يصف تلك العلاقة، لذا فأن في الانحدار اللامعلمي يكون الباحث مهتما بإعطاء وصف عام للعلاقة وليس دراسة تفصيلية لتلك العلاقة .

وعلى الرغم من نماذج الانحدار اللامعلميه هي اضعف وصفا من نماذج الانحدار ألمعلميه ألا أنها في الوقت نفسه تحتاج إلى قيود أو شروط اقل من النماذج المعلمية وهذا الأمر تحديدا هو الذي جعل من نماذج الانحدار اللامعلمية أداة مرغوبة جدا لدى الباحثين لكون أن البيانات الفعلية ليست دوما لديها مواصفات مثالية وكان الباحثون على الدوام يطرحون مسألة مهمة وهي انه في حالة وجود بيانات غير مثالية هل يقفون عاجزين عن حل تلك المشاكل او القيام ببعض التنازلات للحصول على نماذج اقل فاعلية ولكنها في الوقت نفسه توفر حلولا منطقية.

يهدف هذا البحث إلى عرض احد المقدرات المستخدمة لتقدير دالة الانحدار اللامعلمية في حالة استعمال تلك المقدرات معلمات تمهيدية ثابتة ومتغيرة.

2- الجانب النظري:

يجمع الباحثون على إن ما يميز كل علم عن سائر العلوم الأخرى هو مجال عمل ذلك العلم .ويعد الإحصاء واحدا من أهم العلوم التطبيقية ،إذ لا يوجد هنالك نطاقا واضحا لهذا العلم اذ يعمل هذا العلم بشكل متوافق مع العلوم الطبية والهندسية وبحوث الفضاء وغيرها من المجالات التي لا مجال لسردها في هذا البحث.

ونتيجة لهذا برز تساؤل مهم تمثل بمعرفة الترادف ما بين الإحصاء وغالبية العلوم الأخرى سواء أكانت إنسانية أم علمية .والإجابة على التساؤل لم يكن منحسرا فقط لدى المفكرين في أساليب البحث العلمي وإنما كانت مطروحة لدى الإحصائيين أنفسهم وكانت دوما تنحسر بضرورة فهم ما الذي يقدمه الإحصاء للآخرين وهي ببساطة كآلاتي:

- الوصف.
- السيطرة.
- التنبوء أو الاستشراف.

ولتحقيق هذه الأهداف كان لابد للإحصاء من أن يصف الظواهر المختلفة من خلال أنموذج إحصائي مسيطر عليه ومن هنا اندفع الباحثون في المجالات المتنوعة في سبيل اغتنام اكبر قدر من الأساليب الإحصائية لتطوير علومهم ولكن في الوقت نفسه واجهت الإحصائيين مشكلة كبيرة تمثلت بأنه لا يمكن إيجاد أنموذجا أو أسلوبا إحصائيا يصف جميع الظواهر وذلك بسبب تنوع واختلاف شروط ومتطلبات هذه الظواهر.

وكان هذا الأمر لزاما على الإحصائيين توفير عددا كبيرا ومتنوعا من النماذج الإحصائية ليلائم كل منها عددا مختلفا من تلك الظواهر وقد ظهرت نتيجة هذه نماذج سميت بنماذج الانحدار ومن هذه النماذج ما يسمى بنماذج الانحدار اللامعلمي، والشكل العام لهذا الأنموذج هو:

$$Y_i = m(X_i) + \varepsilon_i \qquad \dots (1.1)$$

وتشترط هذه النماذج كون

$$E\varepsilon_i = 0 \ , \ E\varepsilon_i^2 = \sigma^2 \qquad \dots (1.2)$$

مع الإشارة إلى عدم وجود حاجة لتقدير معالم تلك الدالة كون تلك الدالة لا تحتوي على معالم ،أي أنها دالة مجهولة وكل ما نحتاج إليه هو تمهيد تلك النقاط لتقدير منحنى الانحدار [1][10][11].

2.1 – مقدر دالة الانحدار اللامعلمي أللبي:

لتكن $\{X_i\}_{i=1}^n$ تشير إلى عينة عشوائية من مجتمع معين له دالة كثافة f(x) ولتكن $\{Y_i\}_{i=1}^n$ تشير إلى قيم متغير الاستجابة وبافتراض f(x,y) تشير إلى دالة الكثافة الاحتمالية المشتركة للمتغيرين X, Y فأنه يشار لدالة الانحدار بـ $m(x) = E(Y/X=x)$ ولدالة التباين الشرطية بـ $\sigma^2(x) = VAR(Y/X=x)$ وهنالك العديد من الطرائق اللامعلمية المستخدمة لتقدير دالة الانحدار اللامعلمي ومنها الطرائق اللبية وقد تم التركيز هنا في هذا البحث على إحدى المقدرات وهو مقدر Nadaraya-Watson لكن مع استعمال معلمة تمهيدية ثابتة تارة ومعلمة تمهيدية متغيرة تارة أخرى.

2.2 – مقدر Nadaraya-Watson (NW) ذو المعلمة التمهيدية الثابتة :[6]

ويعد من مقدرات الانحدار اللامعلمي الاكثر شيوعا وسمي كذلك كون المعلمة التمهيدية المستخدمة كانت ثابتة، ومن صفات هذا المقدر انه يقدم تقديرا مستمرا للانحدار عندما تكون دالة اللب المستخدمة دالة مستمرة . [8]

ويعرف هذا المقدر رياضيا بالشكل الأتي:

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

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{K_h(X_i - x)}{n^{-1}\sum_{i=1}^{n} K_h(X_i - x)}\right] Y_i \qquad \dots (2.1)$$

$$= \frac{1}{n}\sum_{i=1}^{n} W_h(X_i) Y_i$$

ولبيان خصائص هذا المقدر لابد من توافر الشروط الاتية: [1]

1. دالة الانحدار قابلة للتفاضل مرتين.
2. $$\int u^2 K(u)\,du = \begin{cases} 1 & For\ \ q=0 \\ 0 & For\ \ q=1 \\ \neq 0 & For\ \ q=2 \end{cases}$$
3. $$K(u)=0 \quad For \quad (|u|>1)$$
4. $$|K(u)-K(v)| < |u-v|^{\gamma} \quad ,For\ some\ \ \gamma>0\ and\ some u,v,\gamma \in [-1,1]$$

علما ان دالة اللب تحقق شرط Holder او Lipschitz والذي يتحقق في حالة $\gamma = 1$.

5. $$|f(u)-f(v)| < |u-v|^{B} \quad ,For\ some\ \ B>0\ and\ \ for\ u,v \in [0,1]$$

وبتحقق الشروط المذكورة انفا فأن كلا من التحيز و التباين لمقدر NW يكونان:

$$Bias(\hat{m}^F(x)) = d_K\ h^2 \left(\frac{m''(x)}{2} + \frac{m'(x)\,f'(x)}{f(x)} \right) \qquad \text{... (2.2)}$$

$$Var(\hat{m}^F(x)) = \frac{\sigma^2(x)}{n\,h\,f(x)} C_K \qquad \text{... (2.3)}$$

اذ ان:

$$C_K = \int_{-\infty}^{\infty} K^2(u)\,du \qquad \text{... (2.4)}$$

$$d_K = \int_{-\infty}^{\infty} u^2 K(u)\,du \qquad \text{... (2.5)}$$

ومن المقادير المذكورة انفا نحصل على متوسط مربعات الخطأ المحاذي

$$AMSE(\hat{m}^F(x)) = \frac{\sigma^2(x)}{n h f(x)} C_K + d_K^2 h^4 \left(\frac{m''(x)}{2} + \frac{m'(x) f'(x)}{f(x)} \right)^2 \quad \text{... (2.6)}$$

وان المعلمة التمهيدية المثلى الناتجة من اشتقاق AMSE بالنسبة الى h تكون:

$$h_{opt} = \left(\frac{C_K \sigma^2(x) f^{-1}(x)}{d_K^2 \left(\frac{m''(x)}{2} + \frac{m'(x) f'(x)}{f(x)} \right)^2} \right)^{0.2} n^{-0.2} \quad \text{... (2.7)}$$

2.3 – مقدر Nadaraya-Watson (NW) ذو المعلمة التمهيدية المتغيرة:

ويستند هذا المقدر على استعمال معلمة تمهيدية متغيرة عوضا عن استعمال المعلمة التمهيدية الثابتة والتي تم توضيحها في المبحث السابق وسميت بالمتغيرة بسبب تغير المعلمة التمهيدية عند كل نقطة وذلك من خلال استعمال نوافذ كبيرة عند المناطق ذات الكثافة القليلة واستعمال نوافذ صغيرة عند المناطق ذات الكثافة الكثيفة ،وصيغة هذا المقدر تكون:

$$\hat{m}^V(x) = \frac{\sum_{i=1}^{n} K_{h(X_i)}(X_i - x) Y_i}{\sum_{i=1}^{n} K_{h(X_i)}(X_i - x)} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{K_{h(X_i)}(X_i - x)}{n^{-1} \sum_{i=1}^{n} K_{h(X_i)}(X_i - x)} \right] Y_i \quad \text{... (2.8)}$$

وقد اشار الباحث Kerm عام 2003 [13] الى ان قيمة $h(X_i)$ مساوية الى $h^f * \lambda_i$ ، اذ أن:

$$\lambda_i = \left(\frac{G}{f(X_i)} \right) \quad \text{... (2.9)}$$

كذلك اشار الباحث Takada عام 2001 الى نفس الطريقة مع كون

$$\lambda_i = \left(\frac{\hat{f}(X_i)}{G} \right)^{-\alpha} \quad \text{... (2.10)}$$

وان $\hat{f}(X_i)$ يشير الى مقدر اللب لدالة الكثافة الاحتمالية في حين يشير G الى الوسط الهندسي لدالة الكثافة وان

$$Log\, G = n^{-1} \sum_{i=1}^{n} Log\ \hat{f}(X_i) \quad \text{... (2.11)}$$

مع كون $\alpha = 0.5$. وخصائص مقدر NW المتغير تكون

$$Bias(\hat{m}^V(x)) = d_K\ h^2(X_i) \left(\frac{m''(x)}{2} + \frac{m'(x)\, f'(x)}{f(x)} \right) \quad \text{... (2.12)}$$

$$Var(\hat{m}^V(x)) = \frac{\sigma^2(x)}{n\, h(X_i)\, f(x)} C_K \quad \text{... (2.13)}$$

ومن جمع المقادير المذكورة انفا نحصل على متوسط مربعات الخطأ المحاذي:

$$AMSE(\hat{m}^V(x)) = \frac{\sigma^2(x)}{n h(X_i) f(x)} C_K + d_K^2\, h^4(X_i) \left(\frac{m''(x)}{2} + \frac{m'(x)\, f'(x)}{f(x)} \right)^2$$

... (2.14)

3 – الجانب التجريبي:

استعمل اسلوب المحاكاة للتجارب قيد الدراسة فقد كررت تجربة المحاكاة تكرارا مقداره 400 تجربة لكل انموذج من نماذج الانحدار المفترضة ،وان هدف هذه التجارب يتمثل بمقارنة مقدر دالة الانحدار وفق استعمال كلا من المعلمة التمهيدية الثابتة والمتغيرة مستخدمين معيار متوسط مربعات الخطأ المحاذي كمعيارا للخطأ، وان النماذج المستعملة تمثلت:

1. الانموذج الخطي من الدرجة الثالثة.

$$y_i = -0.03x_i^3 + 0.001x_i^2 + 1.25x_i + \varepsilon_i$$

2. انموذجا غير خطيا.

$$y_i = x_i + 2\exp(-16x_i^2) + \varepsilon_i$$

3. انموذجا غير خطيا يعتمد على دوال مثلثية.

$$y_i = \frac{Sin(2.5\pi x_i)}{1+3x_i^2} + \varepsilon_i$$

اذ يشير X الى المتغير التوضيحي والذي تم توليده كي يتوزع توزيعا طبيعيا بمتوسط صفر وتباين σ^2 وان قيم التباين هي 2,1.5,1 اما حجوم العينات المستعملة في تجارب المحاكاة فكانت 50,100,150 اما دالة اللب المستخدمة فكانت دالة Gaussian وقاعدة الابهام للباحث Silverman كقاعدة لاختيار المعلمة التمهيدية.

وقد تم وضع النتائج لتجارب المحاكاة في الجداول الاتية:

جدول (1)

يوضح قيم معدل متوسط مربعات الخطأ عند استعمال الانموذج الاول

n	σ^2 / Method	1	1.5	2
50	FNW	0.105	0.137	0.169
	VNW	0.198	0.222	0.246
100	FNW	0.0597	0.07528	0.059
	VNW	0.149	0.160	0.140
150	FNW	0.0445	0.0565	0.0684
	VNW	0.130	0.138	0.147

جدول (2)
يوضح قيم معدل متوسط مربعات الخطأ عند استعمال الانموذج الثاني

n	σ^2 / Method	1	1.5	2
50	FNW	0.300	0.383	0.367
	VNW	0.369	0.394	0.420
100	FNW	0.223	0.238	0.253
	VNW	0.282	0.293	0.304
150	FNW	0.200	0.213	0.225
	VNW	0.225	0.265	0.265

جدول (3)
يوضح قيم معدل متوسط مربعات الخطأ عند استعمال الانموذج الثالث

n	σ^2 / Method	1	1.5	2
50	FNW	0.212	0.245	0.278
	VNW	0.190	0.216	0.241
100	FNW	0.182	0.198	0.214
	VNW	0.164	0.176	0.188
150	FNW	0.176	0.188	0.200
	VNW	0.157	0.166	0.174

3.1– تفسير النتائج:

1. الانموذجين الاول والثاني:

1. قيم معدل متوسط مربعات الخطأ تتزايد مع تزايد تباين الخطأ ولجميع الطرائق المستعملة وفي كلا الحالتين (معلمة التمهيد الثابتة والمتغيرة).
2. قيم معدل متوسط مربعات الخطأ تتناقص مع تزايد حجوم العينات ولكلا الحالتين (معلمة التمهيد الثابتة والمتغيرة).
3. اظهرت النتائج افضلية استعمال المعلمة التمهيدية الثابتة ولجميع الحالات المستعملة من حجوم عينات ، تباين خطأ ،اي اظهرت النتائج تفوق مقدر NW الثابت.

2. الانموذج الثالث:

1. تزايد قيم معدل متوسط مربعات الخطأ ولجميع الطرائق ولكلا حالتي المعلمة التمهيدية الثابتة والمتغيرة وبتزايد قيمة تباين الخطأ ولجميع الطرائق المستعملة.
2. تناقص قيم معدل متوسط مربعات الخطأ مع تزايد حجوم العينات ولجميع الطرائق وقيم التباين المستعملة.
3. اظهرت النتائج افضلية مقدر NW ذو المعلمة التمهيدية المتغيرة ولجميع الحالات المستعملة.

4 – الاستنتاجات:

من خلال تنفيذ تجارب المحاكاة المصممة لمقارنة بعض طرائق تقدير دوال الانحدار اللامعلمي واستخدام معالم تمهيدية مختلفة تم التوصل الى الاستنتاجات الاتية:

1. قيم معدل متوسط مربعات الخطأ تتناسب عكسيا مع حجوم العينات ولجميع قيم التباين للخطأ العشوائي والنماذج المستعملة، اي تناقص تلك القيم مع تزايد حجوم العينات.
2. تتناسب قيم معدل متوسط مربعات الخطأ طرديا مع تزايد قيمة التباين للخطأ العشوائي σ_e^2.
3. توسيع حجم العينة يجعل من منحنيات المقدرات المستعملة اكثر قربا للمنحنى الحقيقي كون تلك المقدرات تتأثر بقلة البيانات وتشتتها.
4. من خلال ملاحظة الانموذج الثاني وضحت النتائج افضلية مقدر NW ذو المعلمة التمهيدية الثابتة ولجميع الحالات، اما للانموذج الثالث فقد اوضحت النتائج افضلية مقدرNW ذو المعلمة التمهيدية المتغيرة.

المصادر:

1. منـاف يوسـف حمـود (2000) " مقارنـة مقـدرات Kernel اللامعلميـة لتقـدير دوال الانحــدار" رسـالة ماجستير في الاحصاء / كلية الادارة والاقتصاد / جامعة بغداد .

2. **Cao R. (2002) "An Introduction to Nonparametric Curve Estimation" Universidad De coruna Spain .pp1-4.**
3. **Comniciu D. & Ramesh V. & Meer p. "The variable Bandwidth Mean Shift and Data Driven Scale Selection" pp.1-8.**
4. **Devroy L.P. (1997) "The Uniform Convergence of the Nadaraya–Watson Regression Function Estimation" Canadian J.Stat.–Vol.6-No.2 –PP.179-191.**
5. **Dinardo J. (2001) "Nonparametric Density and Regression Estimation" Journal of Economic Perspectives -Vol.15-No.4–PP.11-28.**
6. **Eubank R.L. & Shucany R.W. (1990) "Adaptive Bandwidth Choice for Kernel Regression" Department of Statistics - No. 5. PP.1-15.**
7. **Fan J. & Gijbels I. (1992) "Variable Bandwidth Linear Regression Smoothers" Ann. Stat.- No.4-PP.2008-2036.**
8. **Fan & Gijbels I. (1996) "A Study of Variable Bandwidth Selection for Local Polynomial Regression" Statistica Sinica-PP.113-127.**
9. **Fox J. (2004) "Nonparametric Regression" Dept. of Sociology – Canada – PP.1- 10.**
10. **Hens .N. (2005) "Non and Semi-Parametric Techniques for Handling Missing data".**
11. **Hua .Q.W. "Probability Density Estimation with Missing Data at Random When Covariables Are Present" Chinese Academy of Science.**
12. **John & Ronen.K&David.A. (2005) "Modeling the United States Stock Market with Kernel Regression "IASTED.**

www.actapress.com/PDFViewer.aspx?paperId=20848

13. **Kerm V.P. (2003) "Adaptive Kernel Density Estimation" J.Royls S.S- pp1-3.**

cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonparametric