تقدير دالة الأنحدار اللامعلمي باستخدام بعض الطرائق اللامعلمية الرتيبة

م. م. ياسمين عبد الرحمن محمد   أ. م. د. دجلة ابراهيم

المستخلص

تم في هذا البحث دراسة الطرائق اللامعلمية الرتيبة لتقدير دالة الأنحدار اللامعلمي، ومعالجة القيم الشاذة الموجودة في دالة الأنحدار اللامعلمي لجعل الدالة رتيبة (متزايدة أو متناقصة).

لذا سنقوم أولاً بتقدير دالة الأنحدار اللامعلمي بإستخدام ممهد Kernel ومن ثم تطبيق الطرائق الرتيبة لجعل الدالة متزايدة إذ سنتناول ثلاث طرائق للتقدير:-

1- طريقة (Mukerjee-stern) إذ سيتم الأستفادة من الحدود الدنيا والحدود العليا للمجاميع الجزئية للبيانات لتعديل مقدر Kernel بإستخدام دالة تقلص (Shrunken).

2- إعتماداً على الطريقة الأولى سيتم أستخدام الحالة الخاصة لدالة التقلص (Shrunken) عندما $\alpha = 0.5$ بوصفها طريقة أخرى مستقلة عن الطريقة الأولى.

3- خوارزمية الأنحدار الرتيب ذو المربعات الصغرى (LSIR) لمعالجة القيم الشاذة.

وسيتم في هذا البحث مقارنة بين هذه الطرائق من خلال إيجاد متوسط مربعات الخطأ والكفاءة النسبية لكل مقدر ولكل أنموذج في الجانب التجريبي من خلال أسلوب محاكاة مونتي كارلو (Monte Carlo)، وتم ايضا مقارنة الطرائق من خلال التطبيق على بيانات لخمسة وعشرين مريضاً مصابين بضغط الدم (العالي والواطئ) وتم التوصل الى أن طريقة (Mukerjee-stern) هي الأفضل من بين الطرائق الأخرى.

ABSTRACT:-

This research was concerning to study monotone nonparametric methods for estimating the nonparametric regression function (i.e treatment outlier) to achieve a monotone function (increasing or decreasing).

So we will use the monotone methods to treatment outlier but after estimate the regression function with use kernel estimator (Nadarya - Watson) these methods are:-

1- Mukerjee method takes averages of maximums and minimum of subsets of the data was used to adjust the initial kernel regression estimates and use the researcher special case when $\alpha = 0.5$.

2- Algorithm least square isotonic regression.

In the experimental aspect comparison was done of which is the best methods through the simulation procedure using Mote Carlo method using five models.

While in the application aspect practical application was done on data represent the measurements for blood pressure patients.

In both aspects we use two of the important statistical measures which are Mean square error (MSE) and efficiency. We find through the application that the best method is Mukerjee method for general case as it has minimum Mean square error and maximum efficiency.

1-1 المقدمة:-

يعد الانحدار من اكثر الاساليب الاحصائية استعمالا في مختلف مجالات العلوم إذ يحدد بوضع العلاقة بين المتغيرات على شكل معادلة ويستدل على اهمية وقوة واتجاه العلاقة تلك من خلال تقدير معاملات معادلة الانحدار.

فعند تطبيق اسلوب الانحدار البسيط على بيانات يتوقع في حالات معينة ان تكون دالة الاستجابة رتيبة متزايدة او رتيبة متناقصة إذ ان الانحدار اللامعلمي يستعمل للبيانات التي تتضمن اتجاهات غير واضحة (متزايدة او متناقصة) وقيم شاذة لذا سنقوم بدراسة العلاقة بين المتغيرات على وفق أنموذج رياضي معين.

فإذا جمعنا العينة العشوائية المكونة من n من المشاهدات $\{(X_i, Y_i)\}_{i=1}^{n}$ فأن أنموذج الانحدار:-

$$Y_i = \mu(X_i) + \varepsilon_i \qquad i = 1,....., n$$

إذ أن $\mu(X)$ تمثل دالة الأنحدار المراد تقديرها إذ يتم أفتراضها على أنها دالة متزايدة في X، أما ε فأنه يمثل الخطأ العشوائي بوسط حسابي صفر وتباين واحد.

تم تقدير دالة الأنحدار بأستعمال ممهد Kernel للحالة الأولية {Nadaraya-watson} لتقدير دالة الأنحدار ومن ثم نلجأ إلى الطرائق المعدلة لتعديل ممهد Kernel للحالة الأولية وجعله رتيباً.

التقنية الأولى هي مقترحة من قبل الباحثين Mukerjee & stern للأنحدار اللامعلمي الرتيب. إذ سيتم الأستفادة من التقنية الأعتيادية البسيطة للحدود الدنيا والحدود العليا للمجاميع الجزئية للبيانات لتعديل ممهد Kernel وجعله رتيباً.

أما التقنية الرتيبة الثانية فهي تتمثل بخوارزمية -pool-adjacent} violators} إذ أن هذه الخوارزمية تعمل بكفاءة في حالة وجود متغير مستقل واحد أما في حالة وجود أكثر من متغير مستقل فأن تطبيق الخوارزمية يؤدي الى اعطاء نتائج غير دقيقة.

الجانب النظري

1-2 الطرائق اللامعلمية الرتيبة لتقدير دالة الانحدار

سيتم استعمال ممهد (Nadaraya-watson) [4] وهو احد ممهدات kernel لتقدير دالة الانحدار اللامعلمي:-

$$T(x)=\frac{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)y_i}{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)}$$

إذ أن k(.) هنا دالة kernel إذ استخدمنا دالة (Gaussian Kernel) لتقدير $T(x)$ ذو عرض حزمة (h) ومساوية لـ (0.1).

وبعد تقدير دالة الانحدار بممهد kernel سيتم معالجة القيم الشاذة لجعل دالة الانحدار رتيبة (متزايدة) باستخدام الطرائق اللامعلمية الرتيبة الاتية:-

1-2-1 طريقة (Mukerjee-stern):- [4] [5][7]

وجد الباحثين Mukerjee & stern بانه يمكن تطبيق الأنحدار اللامعلمي أولاً على البيانات متبعاً بالتعديلات التي تجعل الدالة المقدرة رتيبة، فإذا بدأنا بدراسة رُتبية البيانات الأولية على (X_i) ومن ثم مقدر دالة الأنحدار الرتيبة الناتجة بأستخدام ممهد Kernel الملائم ينتج عنه مقدرات متزايدة {increasing} مع خصائص مقاربة ومماثلة للمقدرات الموجودة في الأنحدار الرتيب {IR}.

وبعد تقدير دالة الأنحدار بأستعمال ممهد (N.W) Kernel سيجري تعديل هذا الممهد مع الأخذ بالاعتبار الحالة الاتية:-

$$Y_i/X_i=x_i\sim U(G_1(x_i),G_2(x_i))\qquad i=1,2,......,n$$

إذ ان:-

$$G_1(x_i) = \min \{T(x') : x' \geq x\}$$

$$G_2(x_i) = \max \{T(x') : x' \leq x\}$$

وعليه فان الشكل العام لطريقة (Mukerjee-stern) الرتيبة (المتزايدة) هي:-

$$G^{\alpha}(x_i) = (1-\alpha)G_1(x_i) + \alpha G_2(x_i)$$

وان معلمة التقلص هي:-

$$\alpha = \frac{\sum_{i=1}^{n}[(T(x_i) - G_1(x_i))(G_2(x_i) - G_1(x_i))]}{\sum_{i=1}^{n}[G_2(x_i) - G_1(x_i)]^2}$$

إن حساب $G^{\alpha}(x_i)$ بسيط جداً حتى للقيم الكبيرة وعليه فأن الخاصية المتزايدة {increasing} للدالة $G^{\alpha}(x_i)$ يتم أثباتها بمشاهدة المعلمتين $G_1(x_i) \& G_2(x_i)$ كلتاهما متزايدتان.

كذلك تم استعمال حالة خاصة لدالة التقلص وهي عندما $\alpha = 0.5$:-

$$G^{0.5}(x_i) = \frac{G_1(x_i) + G_2(x_i)}{2}$$

اذن نجد ان $G^{0.5}(x)$ $(\alpha = 0.5)$ تكون أيضاً متزايدة {increasing} لكون المعلمتين متزايدتين.

2-2-1 الأنحدار الرتيب ذو المربعات الصغرى (LSIR):- [6][3] [2] [1]

إن المقياس الحقيقي للأنحدار الرتيب {IR} هو الأنحدار الرتيب ذو المربعات الصغرى {(Least Square Isotonic Regression),(LSIR)} الذي يستعمل بسهوله ليكون دالة الأنحدار الرتيبة في حالة وجود متغير مستقل واحد {one-dimension}.

لقد حصلت خوارزميات حساب دالة الأنحدار الرتيب ذات المربعات الصغرى {LSIR} على أهتمام كبير في الأدبيات . وكل الخوارزميات تعمل على نحو فاعل جداً في الحالات التي يوجد فيها متغير مستقل واحد ، إذ ان الخوارزمية الواسعة الاستعمال هي خوارزمية {PAV} القابلة للتطبيق في حالة الترتيب الخطي البسيط (متغير مستقل واحد)، أما في حالة وجود أكثر من متغير مستقل فأن هذه الخوارزمية تكون غير كفوءة إذ تنفيذها يكون صعباً بسبب العدد الكبير من المجاميع الدنيا الموجودة أو لانها تتضمن تقنيات البحث التي تتطلب كمية معنوية من الفحص واعادة التعديل.

ولوصف الخوارزمية (PAV) سيتم الاشارة الى مجموعة العناصر المتتالية لـ X_i بأسـم المجاميع والخوارزمية تبدأ بأدق تقسيم ممكن إلى مجاميع، وربط المجاميع سوية خطوة خطوة إلى أن يتم التوصل إلى التقسيم النهائي.

وبالبدء بـ $T(x)$ نتحرك إلى اليمين فإذا كان:-

$$T(x_1) \le T(x_2) \le \dots\dots\dots\dots \le T(x_n)$$

فإن هذا التقسيم الأولي يكون أيضاً التقسيم النهائي وأن:-

$$T^*(x_i) = T(x_i) \qquad i = 1,2,\dots, n$$

وإذا لا نتوقف عند المكان الأول الذي تكون فيه قيمـة $T(x_i) > T(x_{i+1})$ وطالما أن $T(x_i)$ تخالف الأفتراض الرتيب فأننا نجمع $T(x_i)$ مع $T(x_{i+1})$ أي نربط النقطتين x_i , x_{i+1} في المجموعة $\{x_i , x_{i+1}\}$ ونستبدلهم بمعدلهم الأعتيادي:-

$$Av\{i,(i+1)\} = \frac{[w(x_i)T(x_i) + w(x_{i+1})T(x_{i+1})]}{[w(x_i) + w(x_{i+1})]}$$

وبعدها نتحرك إلى اليسار للتأكد من أن $T(x_{i-1}) \le Av(i,i+1)$ وفي حالـة عـدم تحقق ذلك فأننا نجمع $T(x_{i-1})$ مع $Av(i,i+1)$ مستبدليهم بمعدلهم الأعتيادي. ونستمر هكذا حتى نصل إلى الرتابة المطلوبة. وبعدها ننطلق إلى اليمين ونستمر بنفس العمليـة إلـى أن نصل إلى النهاية اليمنى.

وبعد الوصول إلى الترتيب المتزايد المطلوب فأن الأنحدار الرتيب {IR} يكون:-

$$T^*(x_i) = Av(s,p) = \sum_{r=s}^{p} T(x_r)w(x_r) / \sum_{r=s}^{p} w(x_r)$$

وبما أن $T(x)$ هي دالة متزايدة في X_i لدالـة الأنحـدار الرتيبـة $\mu(x_i)$، فـأن الأنحدار الرتيب $T^*(x)$ هو مقدرات المربعات الصغرى لدالة الأنحدار الرتيبة {LSIR}.

الجانب التجريبي

1-3 وصف تجربة المحاكاة:-

تم اعداد برنامج بلغة (Visual basic)، لغرض القيام بمحاكاة التجارب المطلوب دراستها، بعدما تتم عملية توليد الأرقام العشوائية التي تتبع التوزيع المنتظم القياسي $U(0,1)$ وتم ترتيب الارقام العشوائية بشكل تصاعدي، أما المرحلة الثانية من العمل فسيتم توليد قيم المتغير العشوائي بالاعتماد على التوزيعات التالية:-

(التوزيع المنتظم القياسي، التوزيع الاسي، التوزيع الطبيعي القياسي، توزيع t بدرجة حرية 3)

وباحجام عينات (10,50,100,150,200,250)، أما النماذج اللامعلمية التي يتم الأعتماد عليها في تجربة المحاكاة هي كالاتي:-

$$Y_i = e^{X_i} + \varepsilon_i \qquad ......(1)$$

$$Y_i = Sin\left(\frac{\pi}{2} X_i\right) + \varepsilon_i \qquad ......(2)$$

$$Y_i = e^{3(X_i - 1)^2} + \varepsilon_i \qquad ...(3)$$

$$Y_i = \frac{16}{9}(X_i - 1/4)^2 + \varepsilon_i \qquad ...(4)$$

$$Y_i = f(X_i) + \varepsilon_i \qquad .....(5)$$

$Where\ :-$

$$f(X_i) = \begin{cases} X_i & if \quad X_i \in [0,1/3] \\ 7X_i - 2 & if \quad X_i \in [1/3,2/3] \\ X_i + 2 & if \quad X_i \in [2/3,1] \end{cases}$$

تم تكرار تجربة المحاكاة بمقدار (600) تجربة لكل أنموذج من نماذج الأنحدار الأفتراضية، وتم عرض النتائج في الجداول (5),(4),(3),(2),(1) ومن خلال تلك الجداول نجد ان قيم متوسط مربعات الخطأ تتناسب عكسياً مع احجام العينات ولكل النماذج والتوزيعات إلا في النموذج الخامس التوزيع المنتظم نجد ان قيم متوسط مربعات الخطأ تتناسب طردياً مع احجام العينات، وايضاً نلاحظ ان قيم الكفاءة النسبية تتناسب طردياً مع احجام العينات في كل النماذج والتوزيعات إلا في النموذج الخامس التوزيع المنتظم نجد ان قيم الكفاءة تتناسب عكسياً مع احجام العينات، وبملاحظة قيم متوسط مربعات الخطأ للمقدرات نجد ان مقدر $G^{\alpha}(x)$ يمتلك متوسط مربعات خطأ اقل من المقدرين $(T^{*}(x), G^{0.5}(x))$ على التوالي.

جدول (1) يبين قيم متوسط مربعات الخطأ والكفاءة النسبية للأنموذج الأول

distribution	Sample size	$G^{\alpha}(x_i)$		$G^{0.5}(x_i)$		$T^*(x_i)$	
		MSE	eff	MSE	eff	MSE	eff
Uniform	10	0.0624392		0.0627613	0.9948678	0.0624723	0.9994701
	100	0.0245263		0.0245352	0.9996372	0.0245268	0.9999796
	200	0.0220613		0.0220622	0.9999592	0.0220612	1.0000045
Exponential	10	0.6193515		0.6908510	0.8965051	0.6333905	0.9778351
	100	0.4353343		0.4405033	0.9882656	0.4349712	1.0008347
	200	0.4283775		0.4313555	0.9930961	0.4288881	0.9988094
Normal	10	0.3246455		0.3430044	0.9464773	0.3283077	0.9888452
	100	0.2052183		0.2057437	0.9974463	0.2052272	0.9999566
	200	0.1996819		0.1998793	0.9990124	0.1996841	0.9999889
t_3	10	0.9092338		1.0731951	0.8472213	0.9381010	0.9692280
	100	0.5930866		0.6109807	0.9707124	0.5953090	0.9962668
	200	0.5663625		0.5743357	0.9861175	0.5674718	0.9980451

جدول (2) يبين قيم متوسط مربعات الخطأ والكفاءة النسبية للأنموذج الثاني

distribution	Sample size	$G^{\alpha}(x_i)$		$G^{0.5}(x_i)$		$T^{*}(x_i)$	
		MSE	eff	MSE	eff	MSE	eff
Uniform	10	0.1129555		0.1133129	0.9968459	0.1130874	0.9988336
	100	0.0740838		0.0740870	0.9999568	0.0740839	0.9999986
	200	0.0710138		0.0710145	0.9999901	0.0710138	1.0000000
Exponential	10	0.7705983		0.8563236	0.8998914	0.7943359	0.9701164
	100	0.6028023		0.6081931	0.9911363	0.6025554	1.0004097
	200	0.5975910		0.6002199	0.9956201	0.5980811	0.9991805
Normal	10	0.4361689		0.4607192	0.9467130	0.4430349	0.9845023
	100	0.3255221		0.3258710	0.9989293	0.3255260	0.9999988
	200	0.3199150		0.3200310	0.9996375	0.3199160	0.9999968
t_3	10	1.0762772		1.2524097	0.8593651	1.1109049	0.9688292
	100	0.7851461		0.8045948	0.9758279	0.7882343	0.9960821
	200	0.7588128		0.7671749	0.9891001	0.7600742	0.9983404

جدول (3) يبين قيم متوسط مربعات الخطأ والكفاءة النسبية للأنموذج الثالث

distribution	Sample size	$G^{\alpha}(x_i)$		$G^{0.5}(x_i)$		$T^*(x_i)$	
		MSE	eff	MSE	eff	MSE	eff
Uniform	10	0.1478229		0.1525361	0.9691010	0.1482178	0.9973356
	100	0.1152960		0.1160814	0.9932340	0.1152759	1.0001743
	200	0.1126029		0.1128652	0.9973759	0.1125962	1.0000595
Exponential	10	0.8568340		0.9459886	0.9057551	0.8744069	0.9799030
	100	0.7104657		0.7340047	0.9679307	0.7111875	0.9989850
	200	0.7084897		0.7256410	0.9763639	0.7099855	0.9978931
Normal	10	0.4939838		0.5411760	0.9127969	0.5055855	0.9770529
	100	0.4066888		0.4136785	0.9831035	0.4070402	0.9991366
	200	0.4034065		0.4067649	0.9917436	0.4035709	0.9995926
t_3	10	1.1642533		1.3768343	0.8456023	1.2035469	0.9673526
	100	0.9046217		0.9505175	0.9517149	0.9107757	0.9932431
	200	0.8825455		0.9116447	0.9680805	0.8856717	0.9964702

جدول (4) يبين قيم متوسط مربعات الخطأ والكفاءة النسبية للأنموذج الرابع

distribution	Sample size	$G^{\alpha}(x_i)$		$G^{0.5}(x_i)$		$T^{*}(x_i)$	
		MSE	eff	MSE	eff	MSE	eff
Uniform	10	0.1427549		0.1504237	0.9490186	0.1434626	0.9950670
	100	0.1094511		0.1127802	0.9704815	0.1095181	0.9993882
	200	0.1087398		0.1113209	0.9768138	0.1086550	1.0007804
Exponential	10	0.8366866		0.9622946	0.8694703	0.8665247	0.9655657
	100	0.7005003		0.7341889	0.9541145	0.7011181	0.9991188
	200	0.6993363		0.7273569	0.9614761	0.7001078	0.9988980
Normal	10	0.4862740		0.5392817	0.9017068	0.4974910	0.9774528
	100	0.3992496		0.4131378	0.9663836	0.3993061	0.9998585
	200	0.3963677		0.4056122	0.9772085	0.3962346	1.0003359
t_3	10	1.1511239		1.3720413	0.8389863	1.1892906	0.9679080
	100	0.8921582		0.9511537	0.9379747	0.8991733	0.9921982
	200	0.8717041		0.9128813	0.9548931	0.8749047	0.9963417

جدول (5) يبين قيم متوسط مربعات الخطأ والكفاءة النسبية للأنموذج الخامس

distribution	Sample size	$G^{\alpha}(x_i)$		$G^{0.5}(x_i)$		$T^{*}(x_i)$	
		MSE	eff	MSE	eff	MSE	eff
Uniform	10	0.0667667		0.0663085	1.0020383	0.0667736	0.9998956
	100	0.1127050		0.1126840	1.0001866	0.1127053	0.9999973
	200	0.1198571		0.1198542	1.0000239	0.1198573	0.9999985
Exponential	10	0.2551770		0.2849084	0.8956457	0.2609269	0.9779636
	100	0.0443288		0.0452793	0.9790081	0.0442394	1.0020196
	200	0.0349333		0.0354221	0.9862009	0.0349577	0.9993042
Normal	10	0.0927561		0.0943778	0.9828170	0.0932408	0.9948015
	100	0.0802410		0.0800767	1.0020519	0.0802869	0.9994278
	200	0.0059689		0.0059639	1.0008353	0.0059702	0.9997778
t_3	10	0.4482870		0.5496765	0.8155468	0.4568863	0.9811785
	100	0.1036572		0.1099340	0.9429043	0.1043804	0.9930722
	200	0.0792481		0.0810953	0.9772208	0.0795130	0.9966685

وتم تطبيق الطرائق الرتيبة على تجربة طبية تضمنت بيانات تخص ضغط الدم اذ سيتم دراسة تأثير العمر على الاصابة بضغط الدم، البيانات التي تم استخدامها تمثل بيانات لخمسة وعشرين مريضاً ولكلا الجنسين مصابين بضغط الدم العالي والواطئ.

اذ ان عمر المريض يمثل المتغير التوضيحي، وان ضغط الدم العالي والواطئ يمثلان متغيرا الاستجابة، لذا سيتم استخدام انموذجين للانحدار البسيط، الانموذج الاول يمثل تأثير العمر على الاصابة بضغط الدم العالي والانموذج الثاني يمثل تأثير العمر على الاصابة بضغط الدم الواطئ.

جدول (6) يبين متوسط مربعات الخطأ والكفاءة النسبية
لمرضى مصابين بضغط الدم العالي والواطئ

Model	$G^{\alpha}(x)$		$G^{0.5}(x_i)$		$T^*(x_i)$	
	MSE	eff	MSE	eff	MSE	eff
ضغط العالي	1.5903440		1.6457970	0.9663062	1.5949327	0.9971227
ضغط الواطئ	9.9345440		1.0154246	0.9783637	1.0002112	0.9932444

الاستنتاجات

1- افضل مقدر كان $G^{\alpha}(x)$ تبعه مقدر $T^*(x)$ ثم مقدر $G^{0.5}(x)$.

2- كحالة عامة نجد ان افضل توزيع كان التوزيع المنتظم تبعه التوزيع الطبيعي ثم التوزيع الأسي ومن ثم توزيع t_3.

التوصيات

1- تطبيق الطرائق الرتيبة على الأنحدار الخطي المتعدد.

2- تطبيق تقنيات الأنحدار اللامعلمي مثل (B-Splin، K-NN،.....الخ) على البيانات اولاً بدلاً من ممهد Kernel ومن ثم تطبيق الطرائق الرتيبة.

3- مقارنة الطرائق الرتيبة لدالة الأنحدار اللامعلمي في حالة التصميمين الثابت والعشوائي

المصادر الأجنبية:-

1- Barlow, R.E.; and Brunk, H.D. (1972) "The Isotonic Regression Problem and its Dual". Journal of the American Statistical Association, 67, 337, 140-147.
2- Barlow, R.; Batholomew, D.; Bremmer, J.; and Brunk, H. (1972) "Statistical Inference under order Restrictions", John Wiley and Sons, New York.
3- Dykstra, R.L and Robertson, T. (1982) "An algorithm for isotonic regression for two or more independent variables[1]". Annals of Statistics, 10, 3, 708-716.
4- Mukarjee, H. and Stern, S. (1994) "Feasible nonparametric estimation of multiargument monotone functions". Journal of American Statistical Association, 89, 425, 77-80.
5- Mukerjee,H. (1988) "Monotone nonparametric regression". The Annals of Statistics,16,741-750.
6- Robertson, T.; Wright, F.; Dykstra, R. (1988) "Order-Restricted Statistical Inference" John Wiley and Sons; New York.
7- Strand, M. (2003) "Comparison of methods for monotone nonparametric multiple regression".Biometrics (2003),32,1, 165-178.

المصادر العربية:-

8- الجواد: ياسمين عبد الرحمن محمد، (2007) "تقدير دالة الانحدار اللامعلمي باستخدام بعض الطرائق اللامعلمية الرتيبة مع تطبيق عملي للمقارنة بينها" رسالة ماجستير في علوم الأحصاء كلية الأدارة والأقتصاد-جامعة بغداد.