التعدد الخطي في الأنحدار المتعدد اللامعلمي ، الكشف و المعالجة بأستعمال المحاكاة

أ.م.د. لقاء علي محمد / كلية الأدارة و الأقتصاد / جامعة بغداد
الباحث/ صابرين حسين كاظم

تاريخ التقديم:2017/3/6
تاريخ القبول:2017/5/9

المستخلص

يعتبر تحليل الأنحدار هو الحجر الأساس لعلم الأحصاء ، و الذي يعتمد في الغالب على طريقة المربعات الصغرى الأعتيادية **Ordinary Least Square Method** ، لكن كما هو معروف ان الطريقة المذكورة انفآ لها عدة شروط كي تعمل بدقة و بنتائج يمكن الأعتماد عليها ، اضافة الى إن عدم توفر بعض من شروطها يجعل من المستحيل اتمام العمل و تحليل النماذج و من ضمن تلك الشروط هي عدم وجود مشكلة التعدد الخطي (**Multi-CoLinearity**) و نحن في صدد الكشف عن وجود تلك المشكلة بين المتغيرات التوضيحية بأستعمال اختبار فيرار كلوبر، بالأضافة الى شرط خطية البيانات و لعدم توفر الشرط الأخير تم اللجوء الى الأنحدار اللامعلمي (**Nonparametric Regression**) و معالجة المشكلة بإستعمال دالة انحدار الحرف اللبي **Kernel Ridge Regression** و التي تعتمد على تقدير عرض الحزمة (معلمة التمهيد) و لذلك تم اللجوء الى طريقتين مختلفتين لتقدير المعلمة التمهيدية و هما طريقة قاعدة الأبهام **Rule of thumb (RULE)** و الطريقة التمهيدية **Bootstrap (BOOT)** و المقارنة بين تلك الطرق بأستعمال اسلوب المحاكاة .

المصطلحات الرئيسية للبحث/ اختبار **Fararr – Gluaber** ، احصاءة مربع كاي ، تقدير عرض الحزمة **h** ، **RULE** ، **BOOT** ، انحدار الحرف اللبي **KRR** .

*بحث مستل من رسالة ماجستير

الجانب النظري :-

1- **المقدمة** : 1- Introduction :

ان تحليل الأنحدار هو من اساسيات علم الأحصاء و هو من الأساليب المهمة لتحليل العلاقة بين متغيرين او اكثر [7] . و حتى يتم اجراء هذا التحليل و تفسير العلاقة بين المتغيرات يجب ان نختبر فيما اذا كانت هنالك مشكلة عدم تجانس التباين او مشكلة الأرتباط الذاتي أو مشكلة التعدد الخطي ، كما وقد اعتبر الباحث **Kmenta** المشكلة الأخيرة مشكلة درجة وليست مشكلة نوع حيث اشار بأنها مشكلة وجود علاقة بين المتغيرات التوضيحية و بأختفاء تلك العلاقة يمكن استعمال **p** من النماذج الخطية البسيطة و بالتالي لا داعي لأستعمال نموذج خطي متعدد [2] ، ولذك نحن في صدد اختبار مشكلة التعدد الخطي **Multi-colinearity** و التي يتم معالجتها عادة بأستعمال انحدار الحرف الأعتيادي **Ridge Regression** لكن عند وجود بيانات لاخطية تضطرنا الى اللجوء الى أسلوب النمذجة اللامعلمية (**Nonparametric Modeling**) ، حيث إن الحاجة الى المرونة في تحليل البيانات و أدراك الأحصائيين بعدم توافق التقدير المعلمي في تقدير منحنى الأنحدار في حالة لاخطية البيانات و كذلك التطور التكنلوجي ماديا و برمجيا، ادى كل ذلك الى تطور طرق التمهيد اللامعلمية خلال العقدين الماضيين [4] .

و من اهم الطرق اللامعلمية لمعالجة تلك المشكلة هي انحدار الحرف اللبي **Kernel Ridge Regression** **(KRR)** .

2- **هدف البحث** Goal of the Search

أن الهدف من البحث هو الكشف عن مشكلة التعدد الخطي **Multi-Colinearity Problem** بأستعمال اختبار فيرار كلوبر **Ferarr – Gluaber Test** و معالجتها عندما تكون البيانات لاخطية **Nonlinear** و ذلك بإستعمال دالة أنحدار الحرف اللبي **Kernel Ridge Regression** وذلك بأستعمال الطرائق اللبية **Kernel** و تطبيق تلك الطرق بإستعمال اسلوب المحاكاة .

3- **مشكلة التعدد الخطي** Multi – Co Linearty

يقصد بمشكلة التعدد الخطي هو وجود علاقة خطية بين المتغيرات التوضيحية اي حدوث ازدواجية بين تلك المتغيرات او البعض منها. و يظهر الأزدواج الخطي بين المتغيرات التوضيحية عند تساوي قيمة احد المتغيرات التوضيحية (المستقلة) لكافة المشاهدات او عندما يرتبط متغيرين او اكثر بعلاقة خطية و يصعب فصلهما .

إن مشكلة التعدد الخطي نالت الأهتمام من قبل عدد كبير من الأحصائيين و التي ظهرت منذ عام **1934**م من خلال البحث المقدم من قبل العالم (**Frisch**) ، و الذي أشار الى وجود علاقة خطية بين متغيرين أو أكثر من المتغيرات التوضيحية [5] .

4- **اختبار وجود مشكلة التعدد الخطي** Test The Problem Of Multi – Linear [1]

قبل البدأ بتقدير دالة الأنحدار المجهولة في تحليل الأنحدار اللامعلمي يجب اختبار فيما اذا كانت لدينا مشكلة الأزدواجية (اي مشكلة التعدد الخطي) بين المتغيرات التوضيحية أم لا . و هنالك عدة اختبارات للقيام بذلك ، و منها اختبار **Farrar – Glauber** .

1-4 **اختبار فيرار كلوبر** [5] [8] Farrar – Glauer Test

هو اختبار احصائي لمشكلة التعدد الخطي تطور من قبل الباحثين **Farrar** و **Glauber** في عام **1967**و هو عبارة عن مجموعة من ثلاث اختبارات الأختبار الأول يستند الى أحصاءة مربع كاي χ^2 و ذلك لأختبار الفرضية التالية :

H0 : X_j is *orthogonal*

H1 : X_j is *not orthogonal*

و تكتب الصيغة العامة له بالشكل التالي [8] : -

$$\chi_0^2 = -\left[n - 1 - \frac{1}{6}(2 * p + 5)\right] \ln|R|$$

عندما **n** يمثل حجم العينة **Sample Size** .
p تمثل عدد المتغيرات التوضيحية **Independent Variable** .
بينما يشير الرمز **R** الى مصفوفة معاملات الأرتباط التالية : -

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ \vdots & & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

كما يوصف معامل الأرتباط بالصيغة الاتية : -

$$r_{12} = \frac{\sum_{i=1}^{n} x_{i1} x_{i2}}{\sqrt{\sum_{i=1}^{n} x_{i1}^2 \ \ x_{i2}^2}}$$

و تقارن قيمة χ^2 المحسوبة مع χ^2 الجدولية بمستوى دلالة معين و درجة حرية ***(p(p-1)/2)*** [5] .
اما ثاني اختبار فهو اختبار **F** والذي يحدد موقع المتغيرات التي تسبب مشكلة التعدد الخطي ، بينما الأختبار الثالث فهو اختبار **t** و الذي يوضح نمط الأزدواجية الخطية [8] .
وبعد التأكد من ظهور مشكلة التعدد الخطي على الباحث معالجة تلك المشكلة بأحدى الطرق المناسبة أحصائيا مثل انحدار الحرف اللبي ***Kernel Ridge Regression*** .

5- **انحدار الحرف اللبي** Kernel Ridge Regression

هو الشكل اللاخطي لأنحدار الحرف و يرمز له اختصارا بالرمز **KRR** و يمكن كتابة الصيغة العامة بالشكل التالي [13] :

$$\hat{f} := \operatorname{argmin}_{f \in H} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \|f\|_H^2\right)$$

حيث ان λ هي معلمة التنظيم **Regularization Parameter** و التي سيتم التحدث عنها لاحقاً .
و على فرض بأن $x_1, x_2, \ldots\ldots x_n$ هي متجهات عشوائية مستقلة تمتلك نفس التوزيع (**iid**) حيث ان x_iهو متجه من المتغيرات التوضيحية X_j [6] .
و يشير الرمز $\|f\|$ الى طول المتجه f ، حيث ان الدالة **f** تستخرج بالشكل التالي : -

$$f^*(.) = \sum \alpha_i \ K(., X_i)$$

$$\alpha = (\alpha_1, \alpha_2, \ldots\ldots, \alpha_n)$$

حيث إن α تمثل متجهات معاملات انحدار الحرف **Kernel Ridge Regression Vector** و يمكن كتابة الصيغة العامة لها بالشكل الآتي :

$$= (K + \lambda I_n)^{-1} \, Y\alpha^*$$

و يشير الرمز **K** الى مصفوفة **Kernel** و التي يمكن ان توصف بالشكل التالي [6] :-

$$K_{ij} = k(x_i, x_j)$$

و يتم تكوينها بأستعمال دوال كيرنال و قد تم استعمال دالة **Biweight** كما في الجدول التالي [12] :

Kernel	K(u)	Range
Biweight	(3/4) $(1 - u^2)^2$	I (l u l ≤1)

حيث ان :

$$u = \left(\frac{x - X_i}{h}\right)$$

عندما h : تمثل عرض الحزمة .

6- **عرض الحزمة** [1] [7] *BandWidth (Smoothing Parameter)*

تعرف بمعلمة التمهيد او بعرض الحزمة و قد تسمى بمعلمة الأنتشار و هي قد تقترب من الصفر لكن كلما اقتربت قيمتها من الصفر ازداد التباين و قل التحيز بينما بتزايدها يقل التباين و يزداد التحيز و لذلك يجب على الباحث تقديرها بطريقة توازن بين التباين و التحيز و قد تم اللجوء الى طريقتين لتقدير المعلمة المذكورة و هي طريقة قاعدة الأبهام ***Rule Of Thumb (RULE)*** و الطريقة التمهيدية ***Bootstrap (BOOT)*** و كما يلي:-

1-6 **طريقة قاعدة الأبهام** [9] Rule – Of –Thumb Band width

و تدعى بقاعدة التوزيع الطبيعي (**Normal Distribution Rule**) و المقترحة من قبل الباحث **Silverman** ، و تكتب الصيغة العامة لها بالشكل التالي : -

$$h = \hat{\sigma}\, CV(k)\; n^{-\frac{1}{2v+1}}$$

حيث تدل **v** على درجة **kernel** و التي تمثل اول عزم غير صفري و ذلك كون العزوم من الدرجة الفردية تساوي صفر لذلك تكن **v** دائما عدداً زوجيا ، بينما يشير المقدار $\hat{\sigma}$ الى قيمة الأنحراف المعياري للعينة .

وتعتبر **Cv(k)** ثابتة و كما في الجدول التالي و المعتمدة على **v (درجة Kernel)** :-

درجة Kernel	V=2	V=4	V=6
Biweight	2.78	3.39	3.84

2-6 **الطريقة التمهيدية** Bootstrab Smoothing

تعتمد هذه الطريقة على تقليل معيار **MISE(h)** و الذي ينطوي على الدالة المجهولة f_x و بالتالي يصعب ايجاده و كما في الصيغة التالية [10] :

$$\text{MISE (h)} = E \int [\, \hat{f}_h(x; h) - f(x; h)\,]^2\, dx$$

و لذلك نلجأ الى ما افترضه **Peter Hall** في عام **1990** بأن هنالك عينات جديدة و لتكن x_i^* بحجم $n_i \leq n$ مسحوبة من العينة الأصلية x_i وبأيجاد الدالة اللبية و لتكن l بحيث ان $k \neq l$ لتلك العينات الجديدة و بقيمة افتراضية لعرض الحزمة و لتكن **g** بحيث ان $g \neq h$ نستطيع أيجاد *(h)*MISE* بدلا من **MISE (h)** وحسب الصيغة التالية:-

$$\text{MISE}^*(h) = E^* \int [\, \hat{f}_h^*(x) - \hat{f}_g(x)\,]^2\, dx$$

حيث ان :

$$\hat{f}_h^*(x) = n^{-1} \sum_{i=1}^{n} k_h \left(\frac{x - X_i}{h}\right)$$

و أن [14] :

$$\hat{f}_g(x) = n^{-1} \sum_{i=1}^{n} l_g \left(\frac{x - X_i}{g}\right)$$

و تحسنب قيمة معلمة التمهيد النهائية بتقليل $MISE^*(h)$ اي ان [10] :-

$$= argmin_{h>0}\ MISE^*(h)\hat{h}^*$$

لكن عندما تستخدم تقديرات **bootstrap** فإن **MSE** و **MISE** يأخذ عرض الحزمة عادة بحجم $n^{-1/(2r+1)}$
و بالتالي فإن

$$= n^{-1/(2r+1)}\ \hat{h}^*\hat{h}_{boot}$$

عندما $r = 2$ [11] .

7- **معلمة التنظيم** *Regularization parameter*

و تدعى بمعلمة الضبط **Tuning Parameter** او بمعلمة التنظيم **Regularization Parameter** و هي التي تتحكم بكمية التنظيم و يرمز لها بالرمز λ و الصيغة العامة لها بالشكل الآتي :-

$$\lambda_n = 4\sigma R\sqrt{\frac{\log p}{n}}$$

عندما

$$R = max_j \frac{\|x_j\|}{\sqrt{n}}$$

و تعتمد بذلك معلمة التنظيم على قيمة الأنحراف المعياري و عدد المتغيرات التوضيحية [15] .

8- **الجانب التجريبي**:

لعرض الجانب النظري و تطبيق الطرق التي تطرقنا اليها انفآ و المقارنة بينها تم استعمال تجارب المحاكاة وتكرار التجربة و لتوضيح الجانب التجريبي قسم الى عدة مراحل و كما يلي :-

المرحلة الأولى:-

تحديد قيم أفتراضية مختلفة والتي تعتمد عليها المراحل التالية و هي حجوم العينات (n=70,150) و قيم مختلفة للأنحراف المعياري (sd=2.5,5) بالأضافة الى ابعاد مختلفة للمتغيرات التوضيحية (p=9,12) و عرض حزمة افتراضي كقيمة اولية و هو (h=0.1) كما و تم احتساب المتغير المعتمد النموذج التالي :-

[3] **a)** $Y_i = x_i + 2\exp(-16\, x_i^2) + e_i$

المرحلة الثانية:-

تقدير المعالم اللازمة و كما يلي :-

1- تقدير معلمة التمهيد (عرض الحزمة *Bandwidth parameter*) و قد استعملت طريقتين مختلفتين لتقدير المعلمة وهي (طريقة قاعدة الأبهام *RULE* ، طريقة التمهيدية *BOOT*) .

2- تقدير معلمة التنظيم (معلمة الضبط) *Regularzion parameter (Tunning Parameter)* .

كما و تم رسم المتغير المعتمد Y_i مع دالة انحدار الحرف اللبي التقديرية بأستعمال برنامج *Excel* لبعض الحالات المفترضة حيث لا مجال لعرض كل الحالات .

الجدول رقم (1) يوضح قيمة متوسط مربعات الخطأ في حالة كون عدد المتغيرات التوضيحية (p=9)

sample	standard devation	Sd=2.5	Sd=5
	function method	Biw	Biw
n=70	RULE	0.013	0.4178
	BOOT	0.0128	0.4181
n=150	RULE	0.0463	0.3451
	BOOT	0.0463	0.3451

يتبين من الجدول رقم (1) ما يلي:-

- يتبين بأن طريقة Bootstrap بالنسبة لدالة Biweight تعطي اقل قيمة لمعيار الأختبار Mean Square Error (MSE) عندما تكن قيمة الأنحراف المعياري sd=2.5 و حجم عينة n=70 و بالتالي فهي الطريقة الأفضل .
- بينما تكن طريقة Rule Of Thumb الطريقة الأفضل عندما تكن قيمة الأنحراف المعياري sd=5 عند حجم عينة n=70 .
- يلاحظ بأن قيم معيار الأختبار MSE تتعادل لطريقتي Bootstrap و Rule Of Thumb عند حجم عينة n=150 مما يعني الحصول على الأستقرارية لكلتا الطريقتين من حيث الأفضلية .

الجدول رقم (2) يوضح قيمة متوسط مربعات الخطأ في حالة كون عدد المتغيرات التوضيحية (p=12)

sample	standard devation	Sd=2.5	Sd=5
	Function Method	Biw	Biw
n=70	RULE	0.1946	0.5498
	BOOT	0.1947	0.5497
n=150	RULE	0.0586	0.1883
	BOOT	0.0586	0.1883

- يتضح من خلال نتائج المحاكاة و كما في الجدول رقم (2) بأن طريقة Bootstrap تعطي قيمة اقل لمعيار الأختبار MSE من طريقة Rule of Thumb وبالتالي فهي الطريقة الأفضل بالنسبة لدالة Gaussian عندما تكن قيمة الأنحراف المعياري sd=2.5 و حجم عينة n=70 و بالتالي فهي الطريقة الأفضل .
- بينما تكن طريقة Rule Of Thumb الطريقة الأفضل عندما تكن قيمة الأنحراف المعياري sd=5 عند حجم عينة n=70 .
- يلاحظ بأن قيم معيار الأختبار MSE تتعادل لطريقتي Bootstrap و Rule Of Thumb عند حجم عينة n=150 مما يعني الحصول على الأستقرارية لكلتا الطريقتين من حيث الأفضلية .

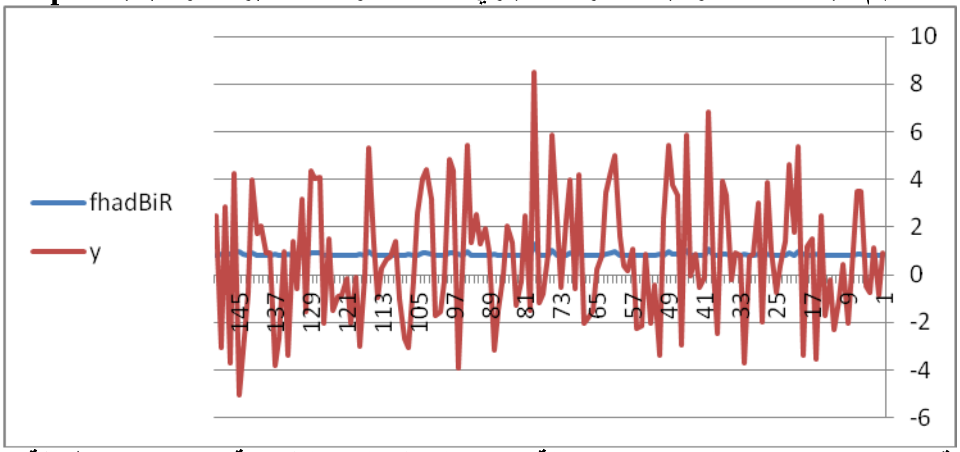
9- **الأستنتاجات** :-

- نستنتج بأن الأفضلية لطريقة Bootstrap و بذلك فقد اثبتت اهميتها عند استعمال قيمة الأنحراف المعياري sd=2.5.
- نستنتج بأن طريقة Rule Of Thumb هي الأفضل عندما تكن قيمة الأنحراف المعياري sd=5 .
- نستنتج بأن قيمة الأنحراف المعياري sd=2.5 هو من يعطي اقل قيمة لمعيار الأختبار MSE في جميع الحالات المفترضة و بالتالي فهو الأنسب اختياراً .
- نستنتج الحصول على الأستقرارية عند حجم العينة n=150 وفي جميع الحالات المفترضة بينما.

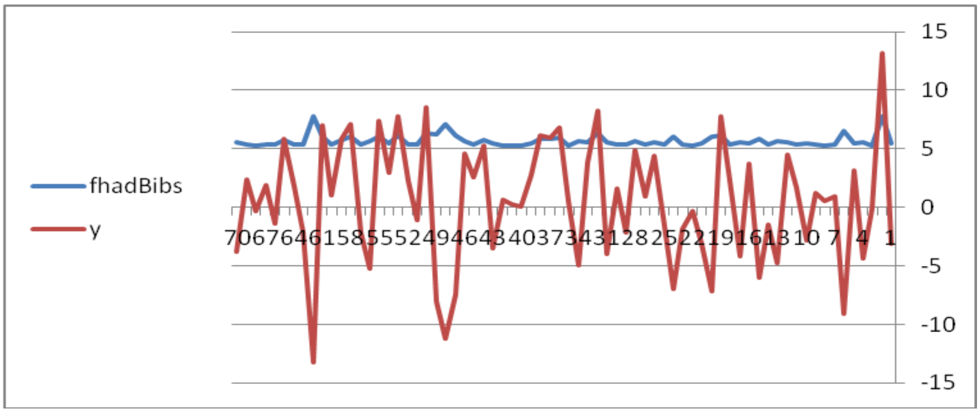
10- **التوصيات** :-

- نوصي بأستعمال طريقة Bootstrap لتقدير عرض الحزمة h عند استعمال دالة Biweight وذلك لأنها أثبتت افضليتها من حيث استقرارها و من حيث اعطائها اقل قيمة لمعيار الأختبار Mean Square Error (MSE) .
- نوصي بأختيار قيمة انحراف معياري اقل و ليكن sd=0.5,1 .
- نوصي بأختيار حجم عينة كبير للحصول على الأتزان من حيث الأفضلية .

الشكل رقم (1) يوضح المتغير المعتمد Y مع دالة انحدار الحرف اللبي التقديرية عند استعمال طريقة RULE لحجم عينة n=150 و قيمة انحراف معياري sd=2.5 و عدد متغيرات توضيحية p=9.



الشكل رقم (2) يوضح المتغير المعتمد Y مع دالة انحدار الحرف اللبي التقديرية عند استعمال طريقة BOOT لحجم عينة n=70 و قيمة انحراف معياري sd=5 و عدد متغيرات توضيحية p=12



المصادر

المصادر العربية:-

1- البياتي ، صابرين حسين كاظم " استعمال دالة الأنحدار الحرف اللبي في معالجة مشكلة التعدد الخطي مع تطبيق عملي " رسالة ماجستير قيد المناقشة .
2- السعدون ، فوزية غالب عمر و الثعلبي ، ساهرة حسين زين " تحليل الأنحدار " لعام 2014 م الطبعة 932 ، طبع في مديرية دار الكتب للطباعة و النشر دامعة البصرة .
3- حمود ، مناف يوسف " مقدر *Nadarea – Watson* اسلوب تمهيدي لتقدير دالة الأنحدار " مجلة العلوم الأقتصادية العدد (65) 2001المجلد (18)ص ص *[283-291]*.
4- الشاروط ، محمد حبيب " مقارنة بعض طرائق تمهيد الأنحدار اللامعلمي بأستخدام المحاكاة " بحث مقدم الى كلية الأدارة و الأقتصاد / جامعة القادسية .
5- كاظم ، اموري هادي و الدليمي ، محمد مناجد "مقدمة في تحليل الأنحدار الخطي"1988 الطبعة 1001 ، طبع في مديرية دار الكتب للطباعة و النشر /جامعة الموصل .

المصادر الأجنبية:-

6- Rudelson , Mark (2015) "Spectral Norm of Random Kernal Matrices WithApplications To Privacy"
7- Rs – Ec2 – Lecture12 .
8- Olawuwo , Simeon ;Ogunleye , Timothy A. ; Ojo, Thompson O & Adejumo, Adebowale O.(2014)" Comparison of Classical Least Squares (CLS), Ridge and Principal Component Methods of Regression Analyses using Gynecological Data".
9- Hansen , Bruce .E.(2009) "Lecture Noteson Nonparametrics" University Of Wisconsin.
10- A. Delaigle And I. Gijbels (2004) " Bootstrap Bandwidth Selection In Kernel Density Estimation From A Contaminated Sample" .
11- Hall , Peter (1990) "Using The Bootstrap To Estimate Mean Square Erorr and Select Smoothing Parameter in Nonparametric Problems".
12- Berwin A.Turlach " Bandwidth Selection In Kernel Density Estimation : A Review" .
13- Yuchen Zhang , John Duchi , Martin Wainwright " Divide and Conquer Kernel Ridge Regression " *University of California* ,Year (2013) .
14- J. S. Marron (1990) " Bootstrap Bandwidth Selection " University of North Carolina
15- Ryota Tomioka "Introduction To The analysis of Learning algorithms : ridge regression and lasso " University of Tokyo .

Multi – Linear in Multiple Nonparametric Regression , Detection and Treatment Using Simulation

ABSTRACT:

It is the regression analysis is the foundation stone of knowledge of statistics , which mostly depends on the ordinary least square method , but as is well known that the way the above mentioned her several conditions to operate accurately and the results can be unreliable , add to that the lack of certain conditions make it impossible to complete the work and analysis method and among those conditions are the multi-co linearity problem , and we are in the process of detected that problem between the independent variables using farrar –glauber test , in addition to the requirement linearity data and the lack of the condition last has been resorting to the nonparametric regression and processor the problem using kernel ridge regression function and that depend on estimate band width (smoothing parameter) therefore has been resorting to two different ways to estimate the parameter and are Rule of thumb (RULE) and Bootstrap (BOOT) and comparison between those ways using the style of simulation .

Keyword: Ferarr – Glauber Test , Chi-square Statistic , bandwidth estimating h , RULE ,BOOT ,Kernel ridge regression KRR .