صفحة ٢٢-٢٦

تقنية ROBUST – LOESS في تحليل الانحدار

زينب حسن راضي

جامعة القادسية / علوم الحاسوب وتكنولوجيا المعلومات / قسم الرياضيات

استلام البحث: ٢٠١٦/٢/٩ إرسال التعديلات: ٢٠١٦/٤/١٧ قبول النشر ٢٠١٦/٥/٢٦

المستخلص :

تناولنا في هذا البحث دراسة أستخدام تقنية **robust-loess** في تقدير دالة الانحدار اللامعلمي ، حيث أن هذه الطريقة تضيف الحصانة الى طريقة **loess** كونها طريقة غير حصينة بسبب اعتمادها استخدام المربعات الصغرى في التمهيد والتي تتأثر بوجود الشواذ ويتم تنفيذ **robust-loess** باستخدام تقنية **Last Absolute Residuals** وتقنية **bisquare** التي تضيف المتانة الى المربعات الصغرى الموزونة في **loess** .

١) المقدمة :

تحليل الانحدار هو اداة احصائية تقوم ببناء نموذج احصائي لتخمين العلاقة بين المتغيرات وقد استخدمنا في هذا الاسلوب تمهيد مخطط التشتت لانه يساعدنا في رؤية تلك العلاقة فهو يلخص متوسط البيانات باستخدام دالة تمهيد للنقاط اضافة الى انه يوفر لنا تقدير او تنبؤ لكل قيمة معطاة من x يطلق على هذه الطريقة أسم **Loess** وهي احدى ادوات تحليل الانحدار اللامعلمي وقد اضفنا لها تقنية **robust** التي أستخدمت من قبل باحثين منهم [1] خلود يوسف خمو و **(2004)** [2] زينب حسن راضي **(2015)** و **William(1979)**[3] و **William** [4] **and Devlin(1988)** ، صيغة نموذج الانحدار اللامعلمي[1][2] :

$$y_i = g(x_i) + \varepsilon_i \qquad i = 1, 2, \dots n$$

حيث ε_i يمثل الخطأ العشوائي و $g(x_i)$ هي دالة الانحدار المجهولة والتي نريد تقديرها أو تمهيدها و y_i المتغير المعتمد .

٢) هدف البحث :

يهدف البحث الى أضافة أستخدام تقنية **Robust** الى طريقة **Loess** لتمهيد الدالة اللامعلمية لاضافة المتانة والدقة في التقدير للوصول الى التقارب من المنحنى الحقيقي.

٣) البواقي Residuals :

من طبيعة تحليل الانحدار (سواء كان معلمي او لامعلمي) هو وجود البواقي التي تزيد من معرفتنا بمدى تشتت الخطأ العشوائي حول خط الانحدار(والتي هي الفروق بين القيمة التي نحسبها من نموذج الانحدار والقيمة الحقيقية) وتعرف بالصيغة التالية :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \qquad i = 1, 2, \dots n$$

حيث ان y_i قيم المتغير المعتمد نسبة الى المتغير المستقل x_i و $\hat{y}_i$ قيمة متغير التنبؤ. ويتم العمل على عرض هذه الفروق او الاختلافات باستخدام احدى الطرق مثل المدرج التكراري او **box plot** او استخدام مخطط التشتت البواقي .

٤) تمهيد Loess

هي طريقة لامعلمية (بمعنى ليس لها تحديدات اولية لوصف شكل العلاقة بين المتغيرات) وهي اختصار لمصطلح الانحدار الموضعي الموزون **Cleveland(1979)**[3] والمصطلحان **Loess** و **Lowess** مشتقان من مصطلح **locally weight scatter plot smoothing** وهما الطريقتان مختلفتان بالموديل المستخدم للانحدار، حيث تستخدم **Lowess** دالة أنحدار متعددة حدود خطية اما **Losee** فتستخدم دالة متعددة حدود تربيعية. ان عملية التمهيد هنا تعتبر موضعية لان كل قيمة ممهدة تحدد بواسطة جوار الاقرب للبيانات الواقعة ضمن الفترة و موزونة لان دالة انحدار الوزن تعرف لكل نقاط البيانات الواقعة ضمن الفترة. وبأستطاعتنا ان نستخدم دالة الوزن الحصينة لجعلها مقاومة للقيم الشاذة او المتطرفة كما في طريقة **Robust-Loess** .

الخطوات المبينة ادناه توضح خطوات وصف Loess في عملية التمهيد:

a) نحسب الانحدار الموزون لكل نقطة من نقاط البيانات ضمن المجال ،حيث الوزن يعطى بأستخدام دالة الوزن الثلاثية :

$$w_i = (1 - \left| \frac{x - x_i}{d(x)} \right|^3)$$

حيث X تمثل قيمة التنبؤ المرتبطة بالتمهيد و X_i قيم نقاط الجوار الى X المعرفة في الفترة و $d(X)$ طول المسافة على المحور السيني لمعظم قيم المؤشر في الفترة .

ويمتاز الوزن بالخصائص التالية :

- نقاط البيانات الممهدة تمتلك وزن اكبر ولها التاثير الاكبر على التمهيد
- نقاط البيانات الشاذة في المجال لها وزن صفر وليس لها تاثير على التمهيد.

b) نستخدم انحدار المربعات الصغرى الموزونة ، اذا كان انحدار lowess تستخدم متعددة حدود من الدرجة الاولى ،اذا كان انحدار loess تستخدم متعددة حدود تربيعية

c) قيمة التمهيد تعطى بواسطة انحدار الموزون لقيمة التنبؤ ، اما اذا كان حساب التمهيد يتضمن نفس العدد للجوار الاقرب لنقاط البيانات على جانبي نقاط البيانات الممهدة فأن دالة الوزن متماثلة و أذا كان عدد نقاط بيانات جوارالتغير غير متماثل حول نقاط البيانات الممهدة فأن دالة الوزن هي دالة غير متماثلة.

٤-١ تمهيد Robust Loess:

أن تقنية Losee هي تقنية غير حصينة لانها تعتمد الية المربعات الصغرى التي تتاثر بوجود القيم المتطرفة والتي تؤثر على قيمة التمهيد وتجعلها لاتعكس سلوك الجزء الاكبر من قيم البيانات ، لذا نلجأ الى استخدام الاجراء الحصين robust لمعالجة هذا التطرف في البيانات الذي يضيف متانة الى اوزان المربعات الصغرى في خطوات Loess أضافة الى استخدامها تقنية bisquare والتي تعيد اوزان النقاط الاساسية الى البواقي ، فاذا كانت البواقي كبيرة (انحراف كبير في النموذج) فان اوزان هذه النقاط تكون منخفض وبالعكس. ويمكن وصف هذه الطريقة بعدة خطوات :

- يتم حساب البواقي من الصيغة التالية :

$$\varepsilon_i = y_i - y_i^{\wedge}$$

- حساب الاوزان الحصينة لكل نقطة من نقاط الفترة ، والاوزان تعطى بواسطة دالة bisquar :

$$w_i = \begin{cases} (1-(\frac{\varepsilon_i}{MAD})^2)^2 & |\varepsilon_i| \le MAD \\ 0 & |\varepsilon_i| > MAD \end{cases}$$

حيث ان ε_i تمثل البواقي أو قيمة الخطأ العشوائي لكل نقاط الفترة، MAD هو المتوسط المطلق لانحرافات البواقي الذي هو مقياس لكمية انتشار البواقي

$$MAD = median(|\varepsilon|)$$

ان قيم الاوزان الحصينة تتغير مع تغير قيمة الخطأ العشوائي مقارنتا مع قيمة المتوسط المطلق ، فاذا كانت قيمة الخطأ العشوائي (ε_i) صغيرة

مقارنة مع MAD فأن قيم الاوزان الحصينة تكون قريبة الى الواحد ، اما اذا كانت قيمة الخطأ العشوائي (ε_i) كبيرة فأن الاوزان الحصينة MAD تساوي صفر ونقاط البيانات المرتبطة تستبعد من حساب التمهيد .

- نعيد تمهيد البيانات بأستخدام الاوزان الحصينة حيث ان قيمة التمهيد النهائية تستخدم كلا من الانحدار الموضعي الموزون والاوزان الحصينة
- نكرر الخطوات السابقة للحصول على افضل النتائج .
- تم تطبيق robust بأستخدام تقنيتين هما (bisquare) و Last Absolut Residual، وسوف نرمز لها بالرمز (LAR) حيث أن (LAR) هي طريقة لايجاد منحني يقلل الفروق المطلقة للبواقي بدلا من تقليل مربعات الاختلافات ولذلك النقاط التي تمتلك قيم عالية يكون لها تأثير أقل على التمهيد أما bisquare فهي تقلل مجموع مربعات الاوزان حيث أن الوزن يعطى لكل نقطة من نقاط

البيانات حسب بعدها أو قربها من المنحني وبالتالي فالنقطة التي تمتلك وزن اكبر تأخذ القيمة صفر وبالعكس وبذلك فأن هذه التقنية تسعى في نفس الوقت الى أيجاد منحني مناسب لمعظم نقاط البيانات أضافة الى تقليل تأثير القيم المتطرفة أن وجدت.

٥) التطبيق :

ولتطبيق ما تم ذكره اعلاه استخدمنا برنامج **MATLAB 2015**، حيث قمنا بتوليد المتغير العشوائي x_i و الاخطاء العشوائية ε_i بأستخدام الاوامر المتاحة في البرنامج ، أضافة الى توليد المتغير المعتمد من خلال جمع دوال المتغير التوضيحي مع متجه الاخطاء العشوائية ، واعتمدنا دالة اختبار متعددة الحدود من درجات مختلفة لتوضيح تمهيد البيانات .

في هذا البحث اعتمدنا معيار مجموع مربعات البواقي الحاصل بسبب التمهيد **SSE** (التي تعرف بأنها الاختلافات بين قيمة المشاهدة وقيمة التنبؤ) واحصائية معامل الارتباط R^2 في الطريقتين أعلاه ، ان استخدام **SSE** اداة تشخيص مفيدة من اجل تحديد ماذا كان المنحني ممهد بشكل يشمل اغلبية البيانات حيث ان القيمة الصغيرة له تشير الى ان النموذج يمتلك خطأ عشوائي صغير وبالعكس، اما **R-Square** فهي مقياس لمدى نجاحنا في وصف او بيان العلاقة بين المتغيرات او هي مربع الارتباطات بين متغير الاستجابة ومتغير التنبؤ وهذه الاحصائية تاخذ قيمة بين ٠ و ١ .

$$SSE = \sum (\bar{y}_i - y_i^{\wedge})^2$$

$$R^2 = \frac{\sum (y_i^{\wedge} - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

ووفقا لمعادلة الانحدار التالية والتي يتوزع فيها المتغير x توزيعا طبيعيا بمتوسط ٠.٥ وتباين ٠.٢٩٣ :

$$g(x_i) = \beta_0 + \beta_1 x + \beta_2 x^2$$

فقد كانت النتائج لطريقة **Loess** كالتالي

β_0 =-0.03475

β_1 = 1.36

β_2 = - 0.4992

Goodness of fit:

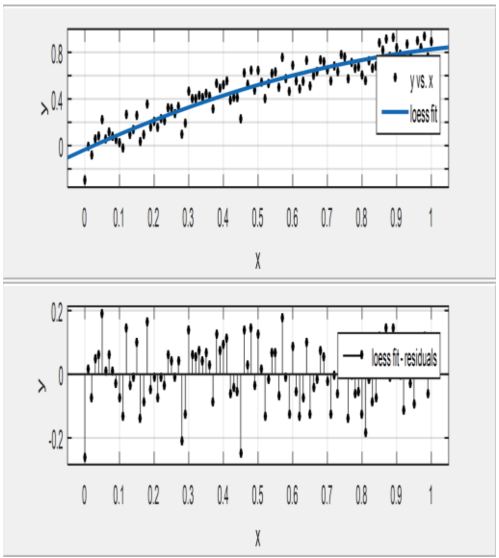
SSE = 0.9101

R- Square = 0.8772

Adjusted R-Square = 0.8747

RMSE=0.09637

والشكل رقم (١) يوضح معادلة الانحدار للنتائج اعلاه.



شكل رقم (١) : معادلة الانحدار لطريقة تمهيد loess

اما لطريقة **robust-loess by bisquar** فقد حصلنا النتائج التالية:

β_0 = 0.9887

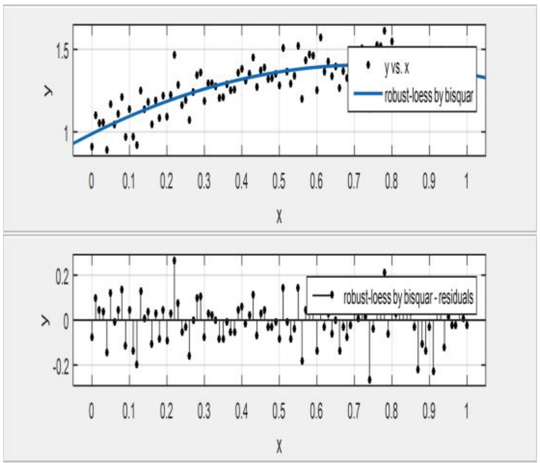
β_1 = 1.138

β_2 = - 0.7756

Goodness of fit:

SSE = 0.9687
R- Square = 0.5908
Adjusted R-Square = 0.5824
RMSE=0.09942

والشكل رقم (٢) يوضح معادلة الانحدار للنتائج اعلاه.



شكل رقم (٢) : معادلة الانحدار لطريقة robust-loess by bisquare

بينما اظهرت طريقة **robust-loess by (LAR)** النتائج التالية:

$\beta_o = 1.028$, $\beta_1 = 0.9745$

$\beta_2 = -0.6433$

Goodness of fit:

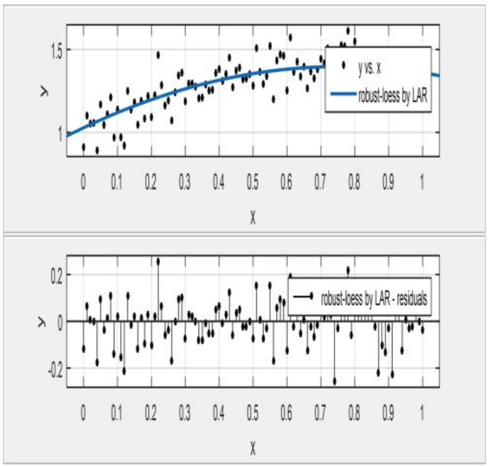
$SSE = 0.9504$

R- Square = 0.5985

Adjusted R-Square = 0.5903

RMSE=0.09848

والشكل رقم (٣) يوضح معادلة الانحدار للنتائج اعلاه.



شكل رقم (٣) : معادلة الانحدار لطريقة robust-loess by (LAR)

٦) الاستنتاج :

١- أن أستخدام **loess** يمثل نهجا مرنا للغاية لمشكلة البيانات أضافة الى أننا لانحتاج الى تحديد اولي لشكل العلاقة بين المتغيرات المستقلة واعتمدة كذلك سهولة ووضوح تركيبها .

٢- تعتمد طريقة **loess** أستخدام المربعات الصغرى في التمهيد لذا نستطيع أعتبارها طريقة عامة لكل من الاسوبين المعلمي واللامعلمي ولكن بنفس الوقت يؤاخذ عليها تأثرها بالقيم المتطرفة لان المربعات الصغرى تتأثر بوجود الشواذ.

٣- يمثل أضافة أستخدام **robust** حلا لمشكلة القيم المتطرفة.

٧) التوصيات :

نوصي باستخدام تقنية robust-loess في التمهيد لما تضيفه من متانة الى طريقة المربعات الصغرى حيث انها تعيد اوزان النقاط الى البواقي وبأستطاعتنا اضافة robust الى الشرائح spline او عندما يكون توزيع الخطأ العشوائي غير طبيعي (ملوث).

المصادر :

١- خلود يوسف خمو "مقارنة اساليب بيز مع طرائق اخرى لتقدير منحني الانحدار" أطروحة دكتوراه في الاحصاء ،جامعة بغداد ، كلية الادارة واقتصاد (٢٠٠٤).

٢- زينب حسن راضي "تمهيد دالة الانحدار اللامعلمي بطرائق تمهيد متنوعة" رسالة ماجستير في الاحصاء الرياضي ،جامعة القادسية ،كلية علوم الحاسوب والرياضيات (٢٠١٥).

٣- Robust locally weighted "William S.Cleveland Journal of regression and smoothing" "American statistical association (1979) issue368 ،vol.74

٤- William S. Cleveland and sudan J. Devlin Locally weighted regression: Approach " Journal of American "analysis by local pp596- ،NO.403 ، vol.83 "statistical association .(1988) ،610

٥- version 6 release 12"Learn Matlab" Matlab .

The Technique of Robust – Loess in Regression Analysis

ZAINB HASSAN RADHY

Al Qadisiya University

College of Computer Science and Information Technology

Abstract :

In this paper the robust-loess method is used to estimate the nonparametric regression function. The Loess is non-robust method and used in case of outliers where it bases on the less squares in regression which affects by presence of outliers. In this paper, the robust-loess has been implemented through applying the last absolute residuals and bi-square techniques which enhanced robustness to the weighted least squares in loess.