تشخيص وتقدير القيمة الشاذة في حالة النموذج الآسي

انكين انترانيك هايك*

1. المقدمة وهدف البحث :

تختلف الطرق والاساليب الاحصائيـــة الخاصة بتحليل البيانـــات مـــع اختلاف طبيعتها وتعتبر دقة وصحة المعلومات جانب مهم في البحث العلمي ، من هنا جاءت اهمية البحث .

ففي حالة وجود قيم شاذة في اي مجموعة من البيانات فمن الطبيعي انعكاس ذلك على المقاييس الاحصائية المستخرجة وبالتالي الوصول الى ارقام ربما لاتعبر عن حقيقة النتيجة المطلوبة ، التي على اساسها يتم اتخاذ القرار حول الظاهرة المدروسة وخصوصاً اذا كانت تلك الظاهرة عن الاستهلاك او الاستثمار او اي بيانات اخرى يتخذ الاقتصادي منها وسيلة لاتخاذ قرار بشأن تقييم حركة الاقتصاد في بلد معين .

يهدف البحث الى تشخيص القيمة الشاذة وطرق معالجتها ومدى تأثير تلك القيمة على عملية التقدير .

2. الجانب النظري

ان نماذج منحنيات النمو هي نماذج غير خطية وتستخدم بشكل واسع في الحياة العملية وباتت مألوفة بالنسبة للعاملين في الاجهزة المسؤولة عن العمل الاحصائي ، وقد استخدمت هذه

* مدرس/ كلية الادارة والاقتصاد/قسم الاحصاء/الجامعة المستنصرية

النماذج كثيراً في البحوث الحياتية وفي دراسة نمو الانتاج الاقتصادي للمنشأة وغيرها ، لان معدل النمو للظاهرة غير ثابت وبالتالي فالعلاقة بين قيم الظاهرة والزمن تكون غير خطية .

هناك مجموعة من النماذج الرياضية غير الخطية التي بدأت تلعب دوراً هاماً في التطبيق العملي عند القيام بعمليات التنبؤ ومن امثلة هذه النماذج .

1. النموذج الاسي Exponential model

يمكن التعبير عن النموذج الاسي بالصيغة الاتية :

$$Y = \gamma e^{\lambda X}.E \qquad \gamma > 0 \ , \quad \lambda > 0 \qquad ...(1)$$

ان γ : عبارة عن ثابت يتحدد بواسطة بعض الشروط الاولية او انها تمثل القيمة الحدية لحجم المجتمع عند بدء المشاهدة .

λ : يمثل نسبة معدل النمو خلال الزمن t $(\lambda > 1)$

باستخدام اللوغاريتمات يمكن تحويل الصيغة رقم (1) الى علاقة خطية ثم اجراء عملية التوفيق عليها باستخدام طريقة المربعات الصغرى .

1.2 القيم الشاذة :

تم دراسة موضوع القيم الشاذة في حالة المتغير الواحد من قبل العديد من الباحثين [1] وان اغلب تلك الدراسات تتعلق بطرق الاحصاء المعلمية (parametric statistics) .

عند فحص بيانات السلسلة الزمنية على الرغم من انها تحتوي على متغيرين الا ان المتغير الاول يمثل سنوات السلسلة وهي قيمة متصاعدة اي لانتوقع ان يظهر فيها قيمة شاذة لكونها

فترات منتظمة . لذا نهمل دور المتغير الاول ونهتم بفحص قيم السلسلة لمعرفة مدى شذوذها ولكن فيما لو كان هناك ظهور واضح في السلسلة للتأثير الدوري فمن غير الممكن اهمال دور المتغير الاول وتتم معالجة مثل تلك الحالة ضمن طرق اخرى لفحص البيانات .

ان الطريقة المعتمدة هنا في فحص القيمة المشكوك فيها بالسلسلة تتعلق بطرق الاحصاء اللامعلمية التي تفترض عدم معرفة التوزيع الحقيقي او التقريبي لمشاهدات مجموعة من البيانات واهم هذه الطرق هي طريقة Tukey (1977) وتسمى طريقة الصندوق والقطع المخططه مع الملخصات الخمسة . Box – and – whisher plots with 5 number summaries

وتتلخص هذه الطريقة في ايجاد :

1. الملخصات الخمسة : اذ تتضمن :

أ. الوسيط median

ان رتبه الوسيط بين القيم المرتبة تصاعدياً او تنازلياً هي [(n+1) / 2] . وبهذه فان قيمة الوسيط هي تلك القيمة الواقعة في وسط القيم المرتبة . اذا كان العدد فردياً ، اما اذا كان العدد زوجيافهي متوسط القيمتين الوسطيتين .

الربيع الادنى Lower Hinge

هو القيمة التي ترتيبها $\frac{n}{4}$ بين القيم المرتبة تصاعدياً .

جـ. الربيع الاعلى upper Hinge

هو القيمة التي ترتيبها $\frac{3n}{4}$ بين القيم المرتبة تصاعديا

2. رسم الصندوق والقطع الخمسة : اذ تتضمن

أ. انتشار الربيعين H – Spread

هو القيمة الناتجة من الفرق بين قيمتي الربيعين اي

$$H - Spread = L_2 - L_1$$

ب. الخطوة Step

هي القيمة الناتجة من حاصل ضرب (1.5) في انتشار الربيعين اي ان

$$Step = (1.5)\ (H - Spread)$$

جـ . السياج الداخلي Inner Fence

يمكن تحديده باحتساب بعد خطوة واحدة عن الربيعين اي ان

الحد الاعلى $L_2 + Step$ = السياج الداخلي

الحد الادنى $L_1 - Step$ =

ء. السياج الخارجي Outer Fence

يمكن تحديده بأحتساب بعد خطوتين عن الربيعين اي ان

الحد الاعلى $L_2 + 2(Step)$ = السياج الخارجي

الحد الادنى $L_1 - 2(Step)$ =

هـ . القيم المجاورة Adjacent Value

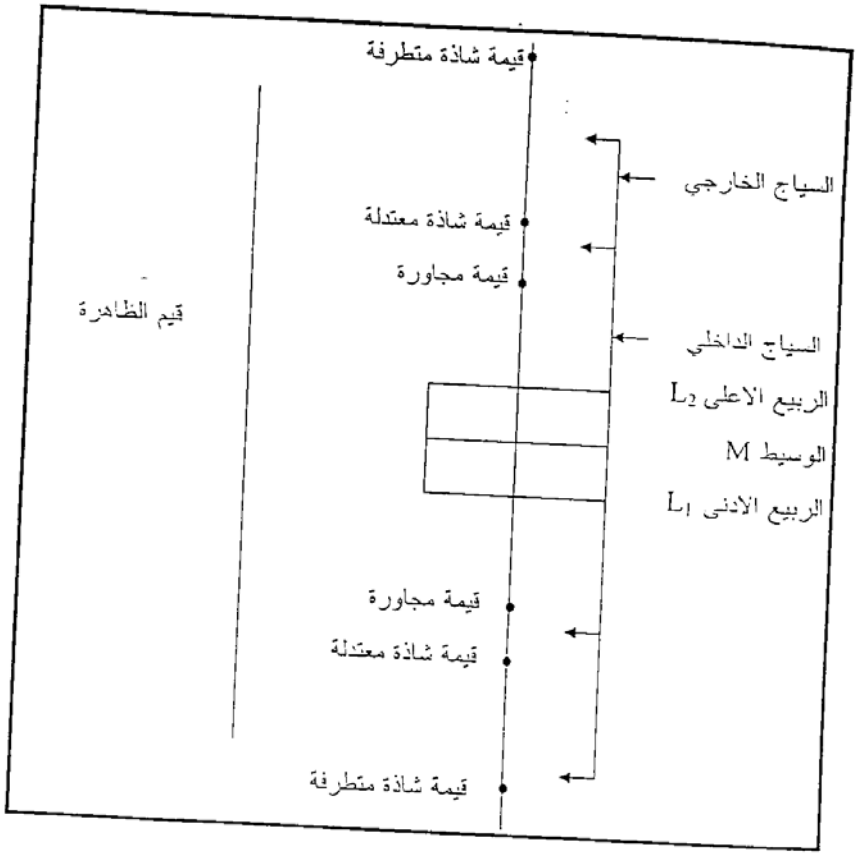
هي تلك القيمة التي تكون قريبة من نهاية السياج الداخلي ولكنها لازالت داخلة .

و. القيم الخارجة Outside (القيم الشاذة المعتدلة) :

هي القيم التي تقع بين السياج الداخلي العلوي (السفلي) والسياج الخارجي العلوي (السفلي) .

ز. القيم البعيدة Farout (القيم الشاذة المتطرفة) :

هي القيمة التي تقع اعلى السياج الخارجي العلوي (او اسفل السياج الخارجي السفلي) .
والشكل رقم (1) يوضح الصندوق والقطع المخططة مع الملخصات الخمسة .



شكل رقم (1)

يوضح الصندوق والقطع المخططة مع الملخصات الخمسة

3. الجانب التطبيقي:

في هذا الجانب تم دراسة حالة القيمة غير المنسجمة والظاهرة في البيانـــات الموضحة فـــي الجدول رقم (1) التي تبين الاستهلاك للقطـــاع الخاص فـــي العراق للسنوات من 1963-1972 والتي اخذت من المجوعة الاحصائية السنوية

International Financial statistics(IFS) / International Monetary Fund (1992)

جدول رقم (1)

يوضح البيانات الاحصائية لاستهلاك القطاع الخاص في العراق للسنوات 1963-1972

السنوات	الاستهلاك
1963	301.3
1964	397.8
1965	455.1
1966	496.8
1967	495.9
1968	547.7
1969	565.1
1970	609.9
1971	633.2
1972	956.1

3_1 التحليل الاحصائي

لتحليل البيانات الاحصائية الموضحة في جدول رقم (1) نجري عملية التقدير على اعتبار ان البيانات صحيحة . وان اتجاه البيانات اقرب الى الدالة الاسية الموجبة لذا فان الدالة المتبعة في التقدير هي :

$$Y = \gamma e^{\lambda X} . E$$

وبأخذ اللوغاريتم للطرفين يكون

$$\log Y = \log\gamma + \lambda X + \log E$$

واذا فاضلنا الطرفين بالنسبة الى X يكون

$$\frac{\partial \log Y}{\partial X} = \lambda$$

وان λ تمثل معدل النمو (Growth Rate) وليكن

$$A = \log\gamma \quad , \quad Z_i = \log Y_i$$

حيث ان $i = 1963 , \dots , 1972$

او $t = 1 , 2 , \dots , 10$

وبهذا يكون جدول البيانات كالاتي :

جدول رقم (2)

t	الاستهلاك
1	301.3
2	397.8
3	455.1
4	496.8
5	495.9
6	547.7
7	565.1
8	609.9
9	633.2
10	956.1

بأستخدام طريقة المربعات الصغرى يكون :

$$\hat{\lambda} = \frac{\sum X_i Z_i - n\overline{XZ}}{\sum X_i^{\,2} - n\overline{X}^2} \qquad \ldots(2)$$

$$\hat{A} = \overline{Z} - \lambda\overline{X} \qquad \ldots(3)$$

وتكون التقديرات الاولية للمعلمتين هي $\lambda_0 = 0.0945$ $\gamma_0 = 311.07$

والتي تستخدم في اولى خطوات كاوس نيوتن ضمن برنامج الانحدار اللاخطي non-linear Regressio في نظام spss وتم التوصل الى المقدرات للمعلمتين $\hat{\gamma}, \hat{\lambda}$ والتي تجعل مجموع مربعات الخطأ اقل ما يمكن وبذلك المعادلة التقديرية تكون

$$\hat{Y} = 301.567 \exp[(0.100)t]$$

وان قيمة الاستهلاك المتوقع للسنوات 1973 ، 1974 ، 1975 ، 1976 هي :

$Y_{(11)} = 905.957$

$Y_{(12)} = 1001.238$

$Y_{(13)} = 1106.539$

$Y_{(14)} = 1222.914$

المطلوب هنا فحص البيانات فيما اذا كانت هناك قيمة او قيم شاذة (outlier values) وحسب طريقة الملخصات الخمسة نجد كل مما يأتي :

1. الوسيط Median

ترتيب الوسيط 5.5 وان Me = 522.3

2. الربيعين Hinges

حيث ترتيبهما على التوالي 2.5 ، 7.5 وان $L_1 = 426.45$ ، $L_2 = 587.5$

3. القيمتين العليا والدنيا هما 301.3 و 956.1

4. انتشار الربيعين (H-Spread)

$$H\text{- Spread} = L_2 - L_1 = 161.05$$

الخطوة Step

$$Step = (1.5)(H\text{-Spread})$$

$$= 241.575$$

السياج الداخلي Inner Fence : يمثل بعد خطوة واحدة عن الربيعين

$L_2 + Step = 829.075$

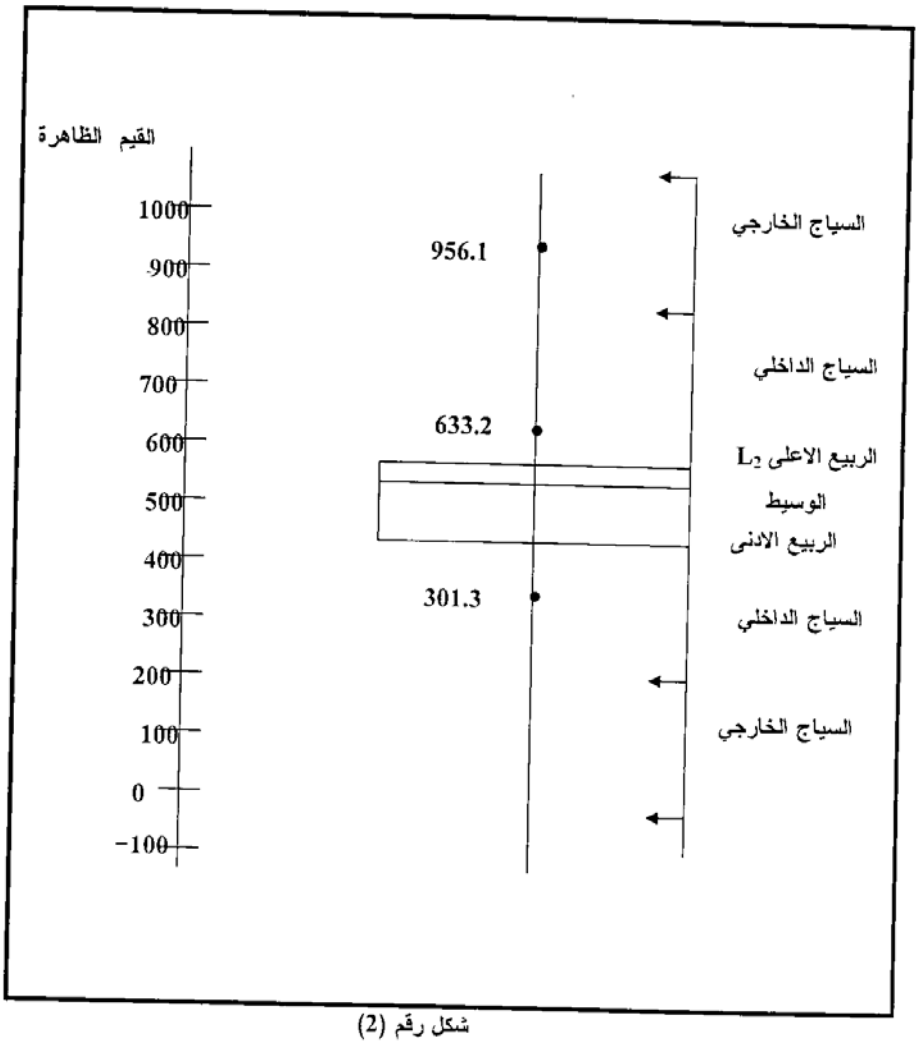
$L_1 - Step = 184.875$

السياج الخارجي Outer Fence : يمثل بعد خطوة واحدة عن السياج الداخلي او بعد خطوتين عن الربيعين

$L_2 + 2$ (Step) = 1070.65

$L_1 - 2$ (Step) = - 56.7

ومن الشكل (2) نجد ان القيمة 956.1 تقع خارج السياج الداخلي وبهذا تكون القيمة 956.1 هي القيمة الشاذة الوحيدة وبالنظر لتأثير هذه القيمة على عملية التقدير ومن غير الممكن حذف القيمة مما يؤدي الى نقص كبير في البيانات خصوصاً فيما لو كان هناك اكثر من قيمة شاذة والذي يؤثر على مستقبل عملية التنبؤ من هنا جاءت فكرة تقدير تلك القيمة او القيم الشاذة .



شكل رقم (2)

يمثل مخطط الصندوق والقطع المخططه

طريقة التقدير المقترحة :

ان يتم التعويض بدل القيمة الشاذة في الكبر بقيمة جديدة تساوي قيمة انتشار الربيعين + (القيمة الشاذة – الوسيط) – 1 ، وفي حالة ظهور اكثر من قيمة في هذا الطرف من القيم الشاذة فنطرح من القيمة [انتشار الربيعين +(القيمة الشاذة – الوسيط)] 1 ، 2 وهكذا . اما اذا كانت القيمة الشاذة هي في الطرف الاصغر فأننا نضيف واحد الى الفرق بين (القيمة الشاذة – الوسيط) وقيمة انتشار الربيعين ونضيف 2 ، 3 وهكذا في حالة وجود اكثر من قيمة شاذة واحدة . وبهذا تكون القيمة لسنة 1972 تساوي

$$H + (\text{Outlier Value} - Me) - 1 = 161.05 + (956.1 - 522.3) - 1 = 593.85$$

ففي هذه الحالة نضمن وقوع تلك القيمة ضمن السياج الداخلي (Inner Fence) . ثم نحاول اعادة تقدير الاستهلاك مرة اخرى .

بأستخدام طريقة المربعات الصغرى والصيغتين (2) ، (3) كانت التقديرات الاولية للمعلمتين $\lambda_0 = 0.0685$ ، $\gamma_0 = 342.15$ والتي تستخدم في اولى خطوات كاوس نيوتن ضمن برنامج الانحدار اللاخطي non – linear Regression في نظام spss وتم التوصل لمقدرات المعلمتين $\hat{\lambda}$, $\hat{\gamma}$ التي تجعل مجموع مربعات الخطأ اقل ما يمكن وبذلك المعادلة التقديرية تكون

$$\hat{Y} = 361.413\exp[(0.060)t]$$

وبهذا تكون قيم التقديرات للسنوات 1973 ، 1974 ، 1975 ، 1976 هي

$Y_{1973} = 699.259$

$Y_{(12)} = Y_{1974} = 742.499$

$Y_{(13)} = Y_{1975} = 788.412$

$Y_{(14)} = Y_{1976} = 837.165$

والجدول رقم (3) يوضح القيم المستخرجة قبل تعديل القيمة الشاذة والقيم المستخرجة بعد تعديل تلك القيمة .

جدول رقم (3)يوضح نتائج التقديرقبل وبعد تعديل القيمة الشاذة

	المعادلة التقديرية	تقدير القيم
قبل التعديل	$\hat{Y} = 301.567\exp[(0.100)t]$	Y_{73} =905.957 Y_{74} = 1001.238 Y_{75} = 1106 . 539
بعد التعديل	$\hat{Y} = 361.413\exp[(0.060)t]$	Y_{73} =699.259 Y_{74} =742.499 Y_{75} = 788 . 412

نلاحظ مدى تأثير القيمة الشاذة على قيم التقدير والذي يؤدي في اغلب الاحوال الى اعطاء صورة غير صحيحة لمتخذ القرار عند دراسة ظاهرة اقتصادية معينة .

4. الاستنتاجات والتوصيات :

1. ان عملية الحصول على بيانات دقيقة وصحيحة تعتبر ضرورية في العمل الاحصائي ونظراً للظروف التي يمر بها قطرنا العزيز فاننا نواجه مشكلة عدم توفر البيانات عن الظواهر الاقتصادية في اغلب الاحيان واذا ما توفرت ربما ينقصها الدقة في احيان كثيرة ومن اجل مواجهة النقص يمكن القيام بعمليات التنبؤ المبنية على اسس علمية احصائية ولتحقيق نتائج تنبؤية سليمة للظواهر الاقتصادية فلابد من تدقيق البيانات والتاكد من صحتها من حيث وجود او عدم وجود قيم شاذة فيها .

2. تبين من الطريقة المقترحة كونها طريقة ملائمة في حل مشكلة القيمة الشاذة اذا ما كانت هناك قيمة او اكثر في البيانات .

3. ان ظهور القيمة الشاذة يؤدي الى نتائج مظللة في عملية التنبؤ مما يؤدي الى تضخم عملية التقدير او التنبؤ فيما لوكانت هناك قيمة شاذة كبيرة او الى صغر حجم التنبؤ في حالة وجود قيمة شاذة صغيرة .

4. يوصي الباحث بضرورة دراسة نماذج نمو اخرى لتحديد فعالية الطريقة المقترحة في حل مشكلة القيم الشاذة مع التطبيق على بيانات حديثة .

المصادر

1. المصادر العربية

1.الجبوري ، شلال حبيب (1990) " اهمية طريقة اكتشاف وتقدير القيمة (القيم الشاذة) في حالة الانحدار الخطي البسيط " مجلة كلية الادارة والاقتصاد " العدد الثاني .

2. المصادر الاجنبية :

[1] Grubbs . F.E. (1950) "Sample criteria for Testing observation " Ann. Math . statist , 21 , 27-58.

[2] Thompson , W.R. (1935) , "on a criterion for the Rejection of observations and the Distribution of the Ratio of the Deviation to the sample standard Deviation " , Ann. Math . statist , 6 , 214-219 .

[3] Tukey , John W. " Exploratory Data Analysis " Addision Wesley publishing compang , 1977