SUST Journal of Natural and Medical Sciences (JNMS) vol. 19 (2) 2018
ISSN (Print): 1858-6805 e-ISSN (Online): 1858-6813

SUST
Journal of Natural and Medical Sciences
Journal home page:
http://Scientific-journal.sustech.edu/

طرق إزالة أثر المشاهدات الشاذة على نموذج الانحدار الخطي البسيط

حامد حسين علي حمد[1]* و احمد محمد عبد الله حمدي[1] و امل السر الخضر عبد الرحيم[2]

1 – جامعة السودان للعلوم والتكنولوجيا – كلية العلوم – قسم الإحصاء

Email : hamidhamad444@gmail.com

تاريخ الاستلام: يونيو 2018 **تاريخ قبول الورقة: اغسطس 2018**

المستخلص

في هذه الورقة العلمية و لإزالة أثر المشاهدات الشاذة على نموذج الانحدار الخطي البسيط تم اقتراح طرق جديدة تتضمن طريقة لاكتشاف المشاهدات الشاذة مصحوبة بمقياس بياني لتحديد عدد المشاهدات الشاذة و طريقة لمعالجة القيم الشاذة من خلال استبدالها بقيم جديدة. تم تطبيق هذه الطرق على 35 نموذج انحدار خطي بسيط بهدف المقارنة بالطرق السابقة و اختبار متوسط الكفاءة النسبية و اختبار كفاءة و منهجية هذه الطرق. بينت النتائج ان الطرق المقترحة اكفأ من الطرق السابقة من حيث رفع قيمة معامل التحديد البسيط R^2 و خفض قيمة متوسط مربعات الخطأ العشوائي MSE و أدت الطرق المقترحة الي تحسن كبير في مؤشرات نماذج الانحدار الخطي البسيط.

الكلمات المفتاحية: ـ الشوارد – الشواذ – الكفاءة النسبية – التشخيص – المعالجة.

Abstract

In this sientific paper to eliminate the outliers impact on the simple linear regression models we proposed a new ways, wich inclowded a way for detecting outliers observation followed by a graphical measures so as to determine it's figuers, and a way to treat the outliers values substitude them with new ones. Application has been done on 35 simple linear models targeting it's compairement by the brevious ways and rational proficiency mean test of the proposed ways. results showed that proposed ways are more perfect than the brevious ones as they increase the R^2 value and decrease MSE value, and lead to a significant development in the indicators of linear regression models.

Keywords: outlying - outliers - rational efficiency - Detect - treatment.

المقدمة

ان وجود المشاهدات الشاذة في مجموعة بيانات الانحدار الخطي البسيط ستجعل النتائج غير واقعية الامر الذي يجعل عملية الانحدار غير نافذة لاغلبية المشاهدات لذا فإن هذه الورقة تركز علي إيضاح مشكلة الشواذ في النماذج الخطية البسيطة من حيث اكتشاف القيم الشاذة و إزالة اثرها من خلال تقدير قيم جديدة لها. و تعد طريقة المربعات الصغري الاعتيادية (ordinary least squares) من افضل الطرق لتقدير معالم نموذج الانحدار الخطي و لكن في ظل وجود المشاهدات الشاذة تكون النتائج باستخدام هذه الطريقة لا يعتمد عليها ، حيث ذكر (Huber, 1973) في مقولته المشهورة أن وجود قيمة شاذة واحدة تهدم المزايا الجيدة لمقدرات المربعات الصغرى. عليه لابد من إيجاد طريقة منهجية

لاكتشاف المشاهدات الشاذة و طريقة منهجية أخرى لتقدير قيم جديدة للمشاهدات الشاذة ، حتى يتسنى لنا الاعتماد علي طريقة المربعات الصغرى في تقدير معالم النموذج. تعتبر طريقة (Srikantant, 1961) من أوائل طرق اكتشاف المشاهدات الشاذة في الانحدار الخطي البسيط و التي اقترح فيها تشخيص مشاهدة واحدة ثم قدمت العديد من الطرق الحديثة و في هذه المقدمة نستعرض المشكلة و الأهمية و الهدف من هذه الورقة العلمية كما يلي: .

تتمثل مشكلة إزالة اثر المشاهدات الشاذة على نموذج الانحدار الخطي المتعدد في محورين أساسيين، المحور الأول هو التشخيص الدقيق و السليم للمشاهدات الشاذة أي يجب تطبيق طرق تشخيص دقيقة لتشخيص كل المشاهدات الشاذة من النموذج اذ ان هناك بعض طرق التشخيص تشخص جزء من المشهدات الشاذة و تترك البعض من دون تشخيص و كذلك التشخيص السليم أي تشخيص المشاهدات الشاذة فقط دون سواها لان هنالك طرق تشخيص تشخص مشاهدات غير شاذة على أساس انها شاذة و هي ليست كذلك، المحور الثاني هو المعالجة من دون تحيز أي عملية تقدير قيم جديدة للمشاهدات الشاذة بالوضع الذي يحد من تأثيرها السلبي على النموذج مع مراعات نسبة تمثيل هذه المشاهدات الي العدد الكلي لمشاهدات النموذج، اضف الى ذلك صعوبة طرق التشخيص و المعالجة و كثرة و ضبابية هذه الطرق في تعاطي تشخيص و معالجة مشكلة الشواذ خصوصا عند تطبيق طريقة المربعات الصغرى في توفيق نماذج الانحدار الخطي البسيط، حيث يتم تشخيص بعض المشاهدات على انها شاذة من خلال بواقي طريقة المربعات الصغري الامر الذي يقود الى تشخيص مشاهدات غير شاذة بانها شاذة.

تاتي أهمية هذه الورقة لمناقشة مشكلة الشواذ في نماذج الانحدار الخطي البسيط من خلال تناول طرق التشخيص المصنفة الي طرق بيانية و طرق تحليلية و بعض طرق المعالجة التي أظهرت درجة من الكفاءة النسبية في معالجة مشكلة الشواذ ، و من اجل اختبار منهجية و متوسط الكفاءة النسبية للطرق المقترحة في حل مشكلة الشواذ تم جمع و تصنيف مجموعة نماذج بلغ عددها 35 نموذج انحدار خطي بسيط تم اخذ معظمها من صفحة (Github, 2018) على الانترنت متاحة على الموقع الالكتروني (https://vincentarelbundock.github.io/*Rdatasets*/) و البعض الآخر من أماكن إقليمية و محلية متفرقة ، منها على سبيل المثال بيانات (Mickey, el at 1967) التي استخدمها العديد من الباحثين لنفس الغرض.

تهدف هذه الورقة العلمية الي الأتي :.

1/ التأكد من وجود المشاهدات الشاذة في نماذج الانحدار الخطي البسيط من خلال تشخيصها باستخدام الطرق المقترحة الجديدة و إمكانية قياس درجة تأثيرها على مؤشرات نموذج الانحدار الخطي البسيط.

2/ قياس الكفاءة النسبية للطرق المقترحة الجديدة مقارنة بالطرق السابقة من حيث تشخيص المشاهدات الشاذة و معالجتها.

3/ اثبات منهجية و كفاءة الطرق المقترحة لازالة اثر المشاهدات الشاذة علي نموذج الانحدار الخطي البسيط من حيث التشخيص و المعالجة و سهولة تطبيقها و مقارنة نتائج النماذج الموفقة بطريقة المربعات الصغري لثلاث أوضاع الأول نتائج البيانات الاصلية و الثاني نتائج البيانات بعد حذف الشواذ و الثالث نتائج البيانات بعد معالجة الشواذ.

تاتي الافتراضات لاختبار مدى فاعلية و كفائة الطرق المقترحة في تشخيص و معالجة الشواذ و

ذلك من خلال مقارنة متوسطات مؤشرات النماذج الخطية البسيطة بعد توفيقها بطريقة المربعات الصغرى وفق التسلسل التالي :

1/ توجد فروق ذات دلالة معنوية بين مؤشرات النماذج للبيانات الاصلية و لنفس البيانات بعد حذف الشواذ.

2/ توجد فروق ذات دلالة معنوية بين مؤشرات النماذج للبيانات الاصلية و لنفس البيانات بعد معالجة الشواذ.

3/ لا توجد فروق ذات دلالة معنوية بين مؤشرات النماذج للبيانات بعد حذف الشواذ و لنفس البيانات بعد معالجة الشواذ.

4/ متوسط الكفائة النسبية لمؤشرات البيانات بعد معالجة الشواذ الى مؤشرات البيانات الاصلية اكبر من الواحد الصحيح.

تم تقسيم هذه الورقة العلمية الى قسمين الأول و يتناول الجانب النظري و الثاني و يتناول الجانب التطبيقي للطرق المقترحة و مدى فعاليتها في تشخيص و معالجة المشاهدات الشاذة في نماذج الانحدار الخطي البسيط.

الجانب النظري

طرق التشخيص و المعالجة

نستعرض في هذا المبحث طرق تشخيص و معالجة المشاهدات الشاذة . و نشير الى ان نماذج الانحدار الخطي البسيط تتاثر بالقيم الشاذة و المتطرفة إن و جدت كغيرها من النماذج الإحصائية الأخرى. و مما لا شك فيه ان طريقة المربعات الصغرى و بسبب القيم الشاذة و المتطرفة تخفق في إعطاء التقديرات الدقيقة لمعالم النموذج لذا فإن هذا الجزء سيركز على طرق تشخيص القيم الشاذة و طرق معالجتها و تحديد مدى تأثير وجود هذه القيم على مقدرات نموذج الانحدار الخطيء البسيط و سنكتفي في هذه الورقة باستعراض بعض الطرق البيانية و الطرق التحليلية.

الطرق البيانية:

هذه الطرق تعتمد في اغلب الأحيان علي شكل انتشار بيانات نموذج الانحدار الخطي و على مهارة الباحث في تحديد القيم الشاذة من خلال رصد القيمة او القيم التي تظهر بصورة معزولة من اغلبية المشاهدات و من ثم استبدالها بقيم تقديرية ، او كما نص (الراوى، 1987) "يمكن اكتشاف الخوارج بيانيا عندما نرسم الرسم البياني لـ e_{is} (Standardized Residual) ضد X_i فالنقاط التي تقع خارج (±2) تعد من الخوارج"، و هنالك العديد من الطرق البيانية مثل طريقة (الجبورى، و آخرون 2002) التي اقترحا فيها معادلة القطع الناقص الذي مركزه متوسطي المتغيرين X و Y و عليه فإن النقطة التي تقع خارج القطع الناقص تعتبر شاذة، بالإضافة لثماني طرق لتقدير قيم جديدة للمشاهدات الشاذة حيث تم التطبيق العملي لهذه الطرق على بيانات Mickey سالفة الذكر و اعطت نتائج جيدة اذ رفعت قيمة معامل التحديد البسيط و خفضت متوسط مربعات الخطأ، اضف الى ذلك طريقة (الباحث 2005) البيانية التي تم تطبيقها علي نفس البيانات و التي أبدت كفاءة اكبر من طرق الجبوري و ناسي، حيث كانت تعتمد في تشخيص المشاهدات الشاذة على المعادلة $Z_i = Y_i - \hat{\beta}_1 X_i$ من خلال توفيق النموذج بطريقة المربعات الصغري و حساب قيم Z_i و ترتيبها تصاعدياً و تمثيلها بيانياً من بعد ذلك تبرز المشاهدات الشاذة و يسهل التعرف عليها و من ثم يتم استبدال القيم الشاذة بقيم جديدة بواسطة طريقة المعالجة المقترحة التي يأتي تفصيلها في سياق هذه الورقة العلمية.

الطرق التحليلية:

في هذه الجزئية نستعرض طرق التشخيص و المعالجة على النحو التالي :.

1/ اكتشاف المشاهدات الشاذة في المتغيرات المستقلة.

"تعرف المشاهدات الشاذة في هذه الحالة بنقاط الجذب (Leverage Points) و هي تعني ان قيمة من قيم x_i شاذة الا ان قيم y_i المقابلة لها تطابق النموذج و هذه النقطة تجذب تقديرات المربعات الصغرى نحوها" (عبد الله ، 2015).

2/ اكتشاف المشاهدات الشاذة في المتغير التابع.

يستخدم لهذا الخصوص بواقي ستيودنت المحذوفة المتحصل عليها بايجاذ القيمة المعيارية للباقي المحذوف الذي يحسب بالصيغة $d_i = y_i - \hat{y}_{i(i)}$ الامر الذي يستدعي بناء عدد n نموذج انحدار خطي و التي طورها (Neter, et al, 1990, p:399) على الصورة $d_i^* = \frac{d_i}{S.e(d_i)}$ ، و لتشخيص المشاهدات الشاذة تتم مقارنة القيمة المطلقة لباقي ستيودنت المحذوف d_i^* بقيمة توزيع t عند درجات حرية $n-p-1$ و مستوى معنوية α فاذا كانت $|d_i^*| > t_{\alpha,n-p-1}$ تعتبر الحالة y_i حالة شاذة لابد من دراستها و تحديد مدى تاثيرها على مقدرات المربعات الصغرى.

3/ المشاهدات الشاذة المؤثرة و طرق الكشف عنها.

توجد عدة مقاييس لتحديد الحالات المؤثرة تعتمد في تحديد اثر الحالة الشاذة على قياس الفرق بين قيم مقدرات المربعات الصغرى باستخدام كل الحالات و باسقاط حالة واحدة $n-1$ و من امثلة هذه المقاييس مقياس COVRATIO الذي طوره (Belsley, et al, 1980) و يقيس هذا المقياس اثر حذف المشاهدة رقم i علي الأخطاء المعيارية.

4/ بعض الحلول المقترحة لمعالجة المشاهدات الشاذة:

بعد التأكد من وجود المشاهدات الشاذة في بيانات المتغير المستقل أو بيانات المتغير التابع أو في كليهما و باعتبار ان المشاهدات الشاذة تمثل بيانات حقيقية فهناك عدد من الحلول حيث ذكر (إسماعيل، 2001)

انه يجب اجراء تحويلات اما للمتغير التابع أو المتغير المستقل مثل تحويلة اللوغريثم و المعكوس و الجذر التربيعي و غيرها، او حذف المشاهدات الشاذة اذا كان حجم العينة كبيراً و إعادة حل النموذج، او جمع بيانات إضافية لزيادة حجم العينة، او تطبيق طريقة متوسط البتر Trimmed Mean (الجبوري، 1998) و تتلخص خطوات اجراء هذه الطريقة بترتيب مشاهدات العمود المشخص بأنه شاذ تصاعديا و تحذف اكبر قيمة و اصغر قيمة في بيانات العمود الشاذ ثم إيجاد الوسط الحسابي للقيم المتبقية و الذي يمثل تقديراً لهذه القيمة.

الطرق المقترحة

من اجل إزالة اثر المشاهدات الشاذة على نموذج الانحدار الخطي البسيط تم اقتراح طرق جديدة تعتمد في تصميمها علي الصيغة $Z_i = Y_i - \hat{\beta}_1 X_i$ و تتضمن طريقة لتشخيص المشاهدات الشاذة مصحوبة بمقياس بياني لتحديد عدد المشاهدات الشاذة و طريقة لمعالجة القيم الشاذة تستبدل هذه المشاهدات الشاذة بقيم جديدة، أي انه لتشخيص القيم الشاذة الموجودة في نماذج الانحدار الخطي البسيط تم تصميم طريقة جديدة لتشخيص المشاهدات الشاذة و لتقدير هذه القيم الشاذة تم تصميم طريقة لتقليل اثر المشاهدات الشاذة (طريقة المعالجة) على نموذج الانحدار الخطي المتعدد، و تعمل طريقة المعالجة وفق معيار نسبى أي انه اذا كان عدد الشواذ قليل

يتم تقليل اثرها بدرجة كبيرة و اذا كان عدد الشواذ كبير يتم تقليل اثرها بدرجة اقل، بعبارة أخرى يفترض اذا تم تطبيق هذه الطرق فأن موشرات النموذج تتحسن بدرجة نسبية عكسية لنسبة المشاهدات الشاذة الي كل المشاهدات و نتناول بالشرح هذه الطرق و كالآتي:.

الطريقة المقترحة لتشخيص القيم الشاذة

تفترض طريقة التشخيص المقترحة ان مؤشرات نموذج الانحدار الخطي البسيط يطرأ عليها تحسن كبير في حال حذف كل المشاهدات الشاذة ، تم تصميم هذه الطريقة بناءاً على الصيغة التالية:.

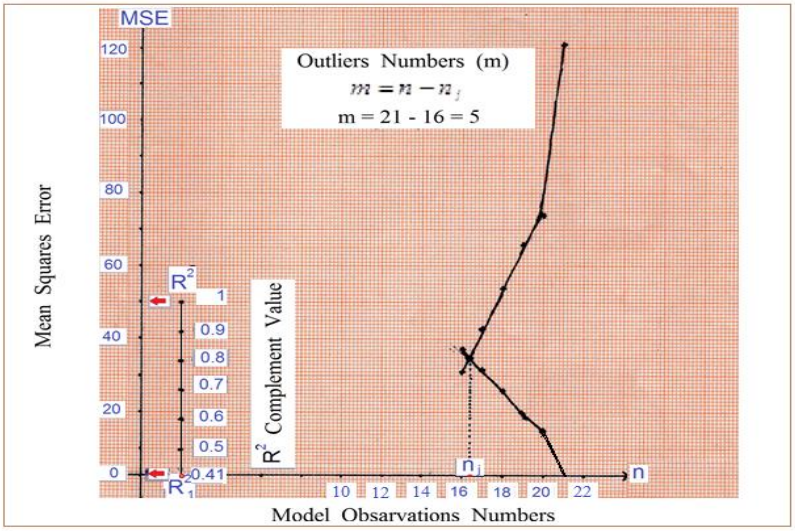
$$Z_{ij} = Y_{ij} - \hat{\beta}_{1j} X_{ij} \rightarrow (1)$$

حيث $Z_{ij} = \hat{\beta}_{0j} + e_{ij}$ و $i = 1,2,\cdots\cdots,n; j = 1,2,\cdots\cdots,m$ حيث تمثل n عدد المشاهدات للنموذج و تمثل m عدد مرات توفيق النموذج، و بناءاً على الصيغة (1) يتم حذف مشاهدة شاذة في كل مرة يتم فيها توفيق النموذج من المشاهدة الشاذة ذات الأثر الأكبر و حتي المشاهدة الشاذة ذات الأثر الأقل، و ذلك من خلال توفيق النموذج بالمشاهدات n_j و تعوض قيمة $\hat{\beta}_{1j}$ في الصيغة (1) و حساب قيمة Z_{ij} و ترتيبها تصاعدياً او تنازلياً و من ثم حذف المشاهدة المناظرة للخطأ المعياري المطلق الأكبر أي ل $|e_{ij}| = |Z_{ij} - \hat{\beta}_{0j}|$ الأكبر، يتكرر هذا الاجراء الى ان يحدد المقياس البياني عدد المشاهدات الشاذة.

المقياس البياني

و بما انه يتم توفيق النموذج في كل محاولة لحذف مشاهدة شاذة من المفترض ان يتناقص متوسط مربع الخطا المعياري MSE_j و يتزايد R_j^2 و ينقص عدد المشاهدات n_j بمعدل مشاهدة واحدة في كل مرة. عليه يمكن تصميم المقياس البياني عن طريق تمثيل نقاط المنحنى (n_j, MSE_j) و نقاط المنحنى (n_j, R_j^2) و نسبة لوجود محوران عموديان محور MSE و محور R^2 و محور افقي مشترك (n)، و لكي ما يؤدي هذا المقياس هدفه المنشود يجب وضع الواحد الصحيح الذي يمثل اعلى قيمة لمحور R^2 بمحاذاة قيمة نتيجة ضرب $MSE_1 * R_1^2$ على محور MSE و وضع قيمة R_1^2 بمحاذاة نقطة الأصل و من ثم تقسيم المسافة $(1 - R_1^2)$ بمقياس الرسم الذي يتم تحديده لمحور MSE اما بالنسبة لمحور (n) المشترك يتم تقسيمه بالوضع الذي يلائم نصف عدد المشاهدات الكلي او اكثر بقليل اذ لا يعقل ان يتعدى عدد المشاهدات الشاذة نصف العدد الكلي للمشاهدات، و من نقطة تقاطع المنحنيان و النزول عمودياً على محور (n) عند القيمة n_j التي تعبر عن عدد المشاهدات الغير شاذة و بعد تقريبها لاقرب عدد صحيح يكون عدد المشاهدات الشاذة مساوياً لعدد مرات توفيق النموذج m أي ان $m = n - n_j$ و الشكل (1) يوضح تصميم المقياس البياني:.

شكل 1: يوضح تصميم المقياس البياني



الصيغة $Z_i = Y_i - \hat{\beta}_1 X_i \rightarrow (2)$ ، و (1) و

كذلك $Z_i = \hat{\beta}_0 + e_i$ أي انه يتم تعويض $\hat{\beta}_0$ للنموذج الموفق بالمشاهدات الغير شاذة n_j ايضاً، و بما ان النموذج يتم توفيقه لمرة واحدة تكتب الصيغة (1) على الصورة التالية:.

و لكن يتم حساب Z_i لكل المشاهدات بما في ذلك المشاهدات الشاذة، و بوضع $d_i = e_i$ تمثل انحرافات Z_i عن $\hat{\beta}_0$ أي ان $d_i = e_i = Z_i - \hat{\beta}_0$ ، و بوضع $\Pr = m/n_j$ و ضربها في d_i نحصل على d_i^* أي ان $d_i(\Pr) = e_i(m/n_j) = e_i^* = Z_i^* - \hat{\beta}_0 = d_i^*$ و بما انه يمكننا حساب $Z_i^* = d_i^* + \hat{\beta}_0$ ، فأنه يمكننا بعد ترتيبها تصاعدياً او تنازليا حساب ازواج المشاهدات الشاذة (x_i^*, y_i^*) كالآتي:.

الطريقة المقترحة لمعالجة المشاهدات الشاذة

لهذه الطريقة يقترح الباحث ترجيح الأخطاء المعيارية e_i لكلٍ من المشاهدات الشاذة و الغير شاذة بنسبتيهما لنحصل علي أخطاء معيارية e_i^* جديدة تقابلها ازواج مشاهدات (x_i^*, y_i^*) جديدة، أي ان $e_i^* = e_i(n_j/n)$ للغير شاذة، $e_i^* = e_i(m/n)$ للشاذة، و هذا الاجراء يغير كل مشاهدات النموذج! و بضرب النسب الترجيحية في مقلوب نسبة الترجيح للمشاهدات الغير شاذة أي ان $e_i^* = e_i(n_j/n)(n/n_j) = e_i$ للغير شاذة تقابلها ازواج المشاهدات (x_i, y_i) للبيانات الاصلية، و $e_i^* = e_i(m/n)(n/n_j) = e_i(m/n_j)$ للشاذة تقابلها ازواج المشاهدات (x_i^*, y_i^*) الجديدة. هذه الطريقة تعتمد على الصيغة (1) كذلك و لكن يتم توفيق النموذج لمرة واحدة بالمشاهدات الغير شاذة n_j فقط و تعوض قيمة $\hat{\beta}_1$ في

$$\forall \ (y_i < y_{i+1})\ y_i^* = y_i + (y_{i+1} - y_i)(Z_i - Z_i^*)/(Z_i - Z_{i+1}) \rightarrow (3)$$

$$\forall \ (y_i > y_{i+1})\ y_i^* = y_i - (y_i - y_{i+1})(Z_i - Z_i^*)/(Z_i - Z_{i+1}) \rightarrow (4)$$

$$\forall \ (x_i < x_{i+1})\ x_i^* = x_i + (x_{i+1} - x_i)(Z_i - Z_i^*)/(Z_i - Z_{i+1}) \rightarrow (5)$$

$$\forall \ (x_i > x_{i+1})\ x_i^* = x_i - (x_i - x_{i+1})(Z_i - Z_i^*)/(Z_i - Z_{i+1}) \rightarrow (6)$$

حيث $i = 1,2,.....n$

نود ان ننبه الى ان طريقة التشخيص قد تسلك منحى مختلف في تشخيص الشواذ خاصة اذا تضمنت مجموعة البيانات مشاهدة او اكثر من الشواذ ذات التأثير الجامح و التي غالباً ما تظهر بصورة معزولة عن اغلب المشاهدات، او انها قد لا تشخص مثل هذا النوع من المشاهدات اذا كان يقع علي خط الانحدار الامر الذي يجعل مؤشرات النموذج تبدو جيدة ظاهريا، لذا يجب الانتباه لمثل هذه المشاهدات من خلال شكل انتشار البيانات (Scatar Plot) قبل بداية التشخيص و اعطاءها قيم تجعلها قريبة من اغلب المشاهدات و اذا ظهرت بعد المعالجة يجب تصنيفها من ضمن الشواذ و معالجتها.

الجانب التطبيقي

في هذا القسم سنبين قدرة الطرق المقترحة في تشخيص و معالجة المشاهدات الشاذة في نموذج الانحدار الخطي البسيط و ذلك في جانبين، الجانب الأول مقارنتها بالطرق السابقة و الجانب الثاني يتعلق باختبار كفاءتها و منهجيتها.

الجانب التطبيقي الأول: مقارنة الطرق المقترحة بالطرق السابقة

تم تطبيق طرق التشخيص و المعالجة على بيانات (Mickey, et al 1967) كما وردت عند (الجبوري، و آخرون 2002) و التي استخدمها (الشميري، و آخرون 2014). عليه تتم مقارنة الطرق المقترحة بطرق الجبوري و التي استخدم فيها للتشخيص طريقة القطع الناقص و عدد من طرق المعالجة، و طريقة الشميري التي استخدم فيها للتشخيص حد الخطأ العشوائي e_i و للمعالجة موجة حد الخطأ المقدر كما نشير في جانب تطبيق الطرق المقترحة الي انه تم تطبيق خطوات طريقة التشخيص من خلال برنامج التحليل الاحصائي SPSS و المقياس البياني باستخدام ورق الرسم البياني انظر (الشكل (1)) هو نفسه المقياس البياني الذي تم استخدامه لتحديد عدد المشاهدات الشاذة و بالإضافة لبرنامج الجداول الالكترونية Exel الذي استوعب كل الصيغ الرياضية لطريقتي التشخيص و المعالجة.

الجدول 1: يبين نتائج المقارنة لقيم المؤشرات ($R^2; MSE$) و الكفائة النسبية للطرق من حيث رفع قيمة R^2 و تقليل قيمة MSE للبيانات بعد التشخيص و المعالجة

عدد الشواذ	ارقام المشاهدات الشاذة	الكفاءة النسبية من حيث		متوسط مربعات الخطأ MSE	معامل التحديد البسيط R^2	اسم طريقة التشخيص	توضيح
		رفع R^2	تقليل MSE				
0	–	1.000	1.000	121.505	0.41	Or.	البيانات العادية
7	3,13,11,14,18,19,20	1.378	2.071	58.674	0.565	NSR-1	افضل طرق الجبوري و ناسي (2002)
1	19	1.417	1.721	70.605	0.581	NSR-4	
2	18,19	1.076	1.673	72.642	0.441	NSR-6	
4	3,13,14,19	1.815	3.122	38.924	0.744	البيانية	طريقة الباحث 2005 غير داخلة في المقارنة
5	3,13,14,19,18	1.446	2.532	47.998	0.605	**البيانية المعدلة**	طريقة الباحث 2005 بعد التعديل
1	19	1.366	1.648	73.7	0.56	الخطأ المقدر	طريقة الشميري (2014)
6	19,13,3,14,20,18	1.734	3.284	36.995	0.711	**الطرق المقترحة**	طرق الباحث الحالية

الجدول (1) يبين نتائج المقارنة للطرق المقترحة مع الطرق السابقة من خلال استخدام مقياس الكفاءة النسبية من حيث رفع قيمة معامل التحديد البسيط R^2 و خفض قيمة متوسط مربعات الخطأ العشوائي MSE، كما نشير الى ان مقياس الكفائة النسبية لرفع قيمة R^2 تم حسابه بقسمة قيمة R^2 للبيانات بعد المعالجة بايٍ من الطرق على قيمة R^2 للبيانات الاصلية و بالنسبة لتقليل قيمة MSE تم حسابه بقسمة قيمة MSE للبيانات الاصلية على قيمة MSE للبيانات بعد المعالجة بايٍ من الطرق لاحظ ان مقياس الكفاءة النسبية للبيانات الاصلية هو الواحد الصحيح لكلا القيمتين. الطريقة البيانية (2005) تم عليها تعديل نظراً لعدم مراعاتها المشاهدات التي تظهر بصورة معزولة من اغلب البيانات و في نفس الوقت تقع قريبة من خط انحدار البيانات مما يجعل النموذج يظهر بمؤشرات جيدة ظاهريا مثل المشاهدة 18 التي اعتبرها الباحث من الشواذ في الدراسة الحالية، عليه تستبعد هذه الطريقة من المقارنة و تعتمد بعد التعديل (بعد ضم المشاهدة 18 لمشاهداتها الشاذة إعادة المعالجة).

من الجدول (1) نلاحظ ان افضل طريقة من بين الطرق السابقة قللت قيمة MSE بكفاءة نسبية قدرها 2.532 و هي طريقة الباحث البيانية المعدلة بينما نجد ان الطرق المقترحة الحالية قللت قيمة MSE بكفاءة نسبية قدرها 3.284 الامر الذي يبين كفاءة الطرق المقترحة الحالية من حيث تقليل قيمة MSE، كما نلاحظ ان افضل طريقة من بين الطرق السابقة رفعت قيمة R^2 بكفاءة نسبية قدرها 1.446 و هي طريقة الباحث البيانية المعدلة بينما نجد ان الطرق المقتحة الحالية رفعت قيمة R^2 بكفاءة نسبية قدرها 1.734 الامر الذي يبين كفاءة الطرق المقترحة الحالية من حيث رفع قيمة R^2. عليه نستنتج ان الطرق المقترحة اكفأ من الطرق السابقة.

الجانب التطبيقي الثاني: اختبار كفاءة و منهجية الطرق المقترحة

لهذا الغرض تم تطبيق الطرق المقترحة على 35 نموذج انحدار خطي بسيط علماً ان هذه النماذج تم اختيارها مصادفةً بما فيها البيانات السابقة، و ذلك من اجل استخلاص ممؤشرات هذه النماذج بالنسبة للثلاث أوضاع السالفة الذكر و بهدف اختبار الفروض المطروحة بالنسبة لهذه الورقة العلمية فكانت المؤشرات المستخلصة كالتالي:.

1/ الوضع الأول:

تم توفيق النماذج بطريقة المربعات الصغرى علي البيانات الاصلية (Orignal Data) و الجدول (2) يوضح مؤشرات النماذج:.

الجدول 2: يبين قيم المؤشرات $(R_2; F; t_{\beta_0}; t_{\beta_1}; MSE)$ للبيانات الاصلية (orignal)

رقم ملف البيانات Data File No.	رقم مسلسل No.	عدد مشاهدات النموذج (n)	معامل التحديد (R^2)	قيمة F المحسوبة (F)	قيمة t المحسوبة ($\lvert t_{\beta_0} \rvert$)	قيمة t المحسوبة ($\lvert t_{\beta_1} \rvert$)	متوسط مربعات الخطأ (MSE)
book001 02 012	1	12	0.746	29.301	7.756	5.413	11.619
book002 02 022	2	22	0.778	70.227	2.594	8.38	10814.373
book003 02 005	3	5	0.877	21.472	0.463	4.634	9543.214
book004 02 050	4	50	0.651	89.567	2.601	9.464	236.532

book005 02 097	5	97	0.629	160.993	9.366	12.688	0.082
book006 02 007	6	7	0.799	19.906	1.679	4.462	242.971
book007 02 012	7	12	0.576	13.56	0.586	3.682	7.226
book008 02 051	8	51	0.836	250.259	10	15.82	0.392
book009 02 025	9	25	0.54	26.954	1.367	5.192	45.783
book010 02 056	10	56	0.716	135.934	12.888	11.659	126.241
book011 02 026	11	26	0.812	103.615	0.906	10.179	0.211
book012 02 053	12	53	0.537	59.206	0.543	7.695	18.727
book013 02 013	13	13	0.915	117.795	1.73	10.853	4.826
book014 02 020	14	20	0.462	22.608	19.366	4.755	0.000
book015 02 011	15	11	0.235	2.759	3.817	1.661	909.885
book016 02 033	16	33	0.415	22.013	12.153	4.692	4.879
book017 02 039	17	39	0.313	16.875	0.552	4.108	1202.677
book018 02 067	18	67	0.863	409.836	0.877	20.244	7500
book019 02 020	19	20	0.995	500	55.835	25	1.513
book020 02 051	20	51	0.448	39.822	4.777	4.448	6500
book021 02 010	21	10	0.644	14.503	0.439	3.808	45.367
book022 02 005	22	5	0.972	105.257	67.351	10.259	0.062
book023 02 047	23	47	0.499	44.842	27.876	6.696	0.009
book024 02 030	24	30	0.677	58.573	0.327	7.653	885.685
book025 02 027	25	27	0.857	150.038	2.248	12.249	0.011
book026 02 061	26	61	0.429	44.296	57.217	6.656	3500
book027 02 053	27	53	0.352	27.646	8.985	5.258	157.017
book028 02 069	28	69	0.203	17.018	3.509	4.125	18.5
book029 02 021	29	21	0.41	13.202	21.681	3.633	121.505
book030 02 015	30	15	0.825	61.097	1.236	7.816	120.641
book031 02 020	31	20	0.872	122.208	14.511	11.055	2471.864
book032 02 022	32	22	0.245	6.507	0.339	2.551	36.551
book033 02 024	33	24	0.409	15.209	1.526	3.9	71.321
book034 02 022	34	22	0.874	139.212	0.918	11.799	0.385
book035 02 025	35	25	0.462	19.782	0.124	4.448	380.085

بالنظر للجدول (2) أعلاه نلاحظ ان قيمة R^2 اقل من 0.75 لعدد 22 نموذج انحدار خطي بسيط من اصل 35 نموذج انحدار خطي بسيط، الامر الذي يدل على انخفاض قيمة R^2 بالنسبة لنماذج الانحدار الخطي البسيط .

2/ الوضع الثاني:

تم توفيق النماذج بطريقة المربعات الصغرى علي البيانات بعد تشخيص المشاهدات الشاذة و حذفها (Deleted Outliers) و الجدول (3) يوضح مؤشرات النماذج:

الجدول 3: يبين قيم المؤشرات $(R_2; F; t_{\beta_0}; t_{\beta_1}; MSE)$ للبيانات بعد التشخيص و الحذف (detected)

رقم ملف البيانات Data File No.	رقم مسلسل No.	عدد مشاهدات النموذج (n_j)	عدد المشاهدات الشاذة (m)	معامل التحديد (R^2)	قيمة F المحسوبة (F)	قيمة t المحسوبة ($\lvert t_{\beta_0} \rvert$)	قيمة t المحسوبة ($\lvert t_{\beta_0} \rvert$)	متوسط مربعات الخطأ (MSE)
book001 02 012	1	10	2	0.888	63.728	11.924	7.983	5.194
book002 02 022	2	20	2	0.899	160.079	2.447	12.652	4768.627
book003 02 005	3	4	1	0.923	24.075	0.724	4.907	8696.778
book004 02 050	4	40	10	0.805	157.118	3.017	12.535	70.17
book005 02 097	5	74	23	0.83	352.596	16.095	18.778	0.029
book006 02 007	6	6	1	0.916	43.418	1.612	6.589	59.25
book007 02 012	7	9	3	0.868	46.111	1.895	6.791	2.603
book008 02 051	8	43	8	0.904	385.131	8	19.625	0.138
book009 02 025	9	19	6	0.783	61.275	3.301	7.828	12.338
book010 02 056	10	40	16	0.827	181.095	15.793	13.457	42.094
book011 02 026	11	16	10	0.892	115.221	1.313	10.734	0.064
book012 02 053	12	38	15	0.774	123.642	1.974	11.119	5.259
book013 02 013	13	11	2	0.968	273.893	2.858	16.55	1.878
book014 02 020	14	13	7	0.733	30.131	34.577	5.489	0
book015 02 011	15	6	5	0.937	59.442	12.098	7.71	128.392
book016 02 033	16	20	13	0.832	89.408	20.877	9.456	1.393
book017 02 039	17	22	17	0.565	26.007	5.373	5.1	120.667
book018 02 067	18	61	6	0.923	709.346	1.693	26.634	2800
book019 02 020	19	16	4	0.997	1000	75.337	35	0.81

book020 02 051	20	35	16	0.75	95.749	13.625	9.785	1800
book021 02 010	21	7	3	0.899	44.605	1.919	6.679	16.162
book022 02 005	22	4	1	0.992	256.889	111.375	16.028	0.009
book023 02 047	23	32	15	0.729	80.652	46.882	8.981	0.002
book024 02 030	24	26	4	0.867	155.949	1.396	12.488	341.052
book025 02 027	25	21	6	0.919	214.834	5.868	14.657	0.004
book026 02 061	26	38	23	0.798	142.565	83.537	11.94	1200
book027 02 053	27	27	26	0.354	13.687	22.288	3.7	16.388
book028 02 069	28	35	34	0.358	18.385	23.455	4.288	0.784
book029 02 021	29	15	6	0.834	70.157	39.676	8.376	33.007
book030 02 015	30	14	1	0.971	404.701	1.625	20.117	10.644
book031 02 020	31	19	1	0.931	230.737	16.154	15.19	1233.225
book032 02 022	32	14	8	0.863	75.539	0.339	8.691	6.886
book033 02 024	33	17	7	0.767	49.414	1.526	7.03	16.681
book034 02 022	34	19	3	0.956	367.889	0.918	19.18	0.07
book035 02 025	35	16	9	0.734	38.524	0.124	6.207	137.612

بالنظر للجدول (3) أعلاه و بعد حذف المشاهدات الشاذة من نفس النماذج السابقة نلاحظ ان قيمة R^2 أصبحت اقل من 0.75 لعدد 6 نماذج انحدار خطي بسيط فقط من اصل 35 نموذج انحدار خطي بسيط مقارنة مع الجدول (2) نلاحظ تحسن كبير في قيمة R^2 ، الامر الذي يدل على تأثر نماذج الانحدار الخطي البسيط بالمشاهدات الشاذة.

3/ الوضع الثالث:

تم توفيق النماذج بطريقة المربعات الصغرى علي البيانات بعد تشخيص المشاهدات الشاذة و معالجتها (Detected Treated Outliers) و الجدول (4) يوضح مؤشرات النماذج:.

الجدول 4: يبين قيم المؤشرات $(R_2; F; t_{\beta_0}; t_{\beta_1}; MSE)$ للبيانات بعد التشخيص و المعالجة (modulated)

رقم ملف البيانات Data File No.	رقم مسلسل No.	عدد مشاهدات النموذج (n)	عدد المشاهدات الشاذة (m)	قيمة الترجيح الاحتمالية W Pr	معامل التحديد (R^2)	قيمة F المحسوبة (F)	قيمة t المحسوبة ($\lvert t_{\beta_0} \rvert$)	قيمة t المحسوبة ($\lvert t_{\beta_0} \rvert$)	متوسط مربعات الخطأ (MSE)
book001 02 012	1	12	2	0.2	0.899	89.081	13.3	9.438	4.424
book002 02 022	2	22	2	0.1	0.922	235.019	2.743	15.33	4356.054
book003 02 005	3	5	1	0.25	0.961	74.618	1.079	8.638	5903.076

book004 02 050	4	50	10	0.25	0.827	229.27	3.318	15.142	66.829
book005 02 097	5	97	23	0.311	0.844	513.642	20.048	22.664	0.027
book006 02 007	6	7	1	0.167	0.915	53.928	1.902	7.344	53.526
book007 02 012	7	12	3	0.333	0.871	67.817	2.499	8.235	2.466
book008 02 051	8	51	8	0.186	0.93	655.701	15	25.607	0.127
book009 02 025	9	25	6	0.316	0.789	86.207	3.038	9.285	12.196
book010 02 056	10	56	16	0.4	0.839	280.937	19.591	16.761	45.696
book011 02 026	11	26	10	0.625	0.986	1743.307	0.784	11.332	0.119
book012 02 053	12	53	15	0.395	0.703	120.625	2.497	10.983	6.511
book013 02 013	13	13	2	0.182	0.973	390.58	3.154	19.763	1.659
book014 02 020	14	20	7	0.539	0.669	36.359	34.165	6.03	0
book015 02 011	15	11	5	0.833	0.624	14.949	6.296	3.866	543.642
book016 02 033	16	33	13	0.65	0.757	96.372	21.195	9.817	2.284
book017 02 039	17	39	17	0.773	0.319	17.35	5.266	4.165	798.61
book018 02 067	18	67	6	0.098	0.931	883.03	2.113	29.716	3500
book019 02 020	19	20	4	0.25	0.998	2000	109.23	40	0.666
book020 02 051	20	51	16	0.457	0.74	139.665	16.818	11.818	2500
book021 02 010	21	10	3	0.429	0.888	63.442	3.332	7.965	15.243
book022 02 005	22	5	1	0.25	0.982	161.706	88.479	12.716	0.015
book023 02 047	23	47	15	0.469	0.759	141.653	58.27	11.902	0.003
book024 02 030	24	30	4	0.154	0.885	214.662	1.729	14.651	301.892
book025 02 027	25	27	6	0.286	0.908	245.94	6.773	15.682	0.003
book026 02 061	26	61	23	0.605	0.717	149.589	89.019	12.231	2000
book027 02 053	27	53	26	1	0.352	27.646	8.985	5.258	157.017
book028 02 069	28	69	34	1	0.203	17.018	3.509	4.125	18.5
book029 02 021	29	21	6	0.4	0.711	46.803	35.456	6.841	36.995
book030 02 015	30	15	1	0.071	0.97	426.727	1.648	20.657	10.47
book031 02 020	31	20	1	0.053	0.934	254.371	16.733	15.949	1170.016
book032 02 022	32	22	8	0.571	0.74	56.812	4.879	7.537	20.405
book033 02 024	33	24	7	0.412	0.758	68.859	3.505	8.298	21.912
book034 02 022	34	22	3	0.159	0.954	415.457	2.426	20.383	0.066
book035 02 025	35	25	9	0.563	0.624	38.137	1.016	6.176	154.594

بالنظر للجدول (4) أعلاه و بعد معالجة قيم المشاهدات الشاذة من نفس النماذج السابقة نلاحظ ان قيمة R^2 أصبحت اقل من 0.75 لعدد 11 نموذج انحدار خطي بسيط من اصل 35 نموذج انحدار خطي بسيط و بالمقارنة مع الجدول (2) الذي اظهر ان قيمة R^2 اقل من 0.75 لعدد 22 نموذج انحدار خطي بسيط من اصل 35 نموذج انحدار خطي بسيط أي المقارنة بين 11 التي تمثل وسط المدى من 1 الى 22 ، نلاحظ انه قد تم معالجة قيم المشاهدات الشاذة بدرجة متوسطة مع مراعات نسبة وجودها الممثلة بنسبة الترجيح (W Pr)، الامر الذي يدل على وسطية المعالجة لنماذج الانحدار الخطي البسيط.

و من اجل اختبار متوسط الكفاءة النسبية تم قسمة مؤشرات النماذج للبيانات الموفقة بعد التشخيص و المعالجة علي مؤشرات النماذج للبيانات الموفقة الاصلية باستثناء مؤشر MSE الذي تم اخذ مقلوبه بعد القسمة بصفته المؤشر الوحيد الذي تقل قيمته بعد التشخيص او بعد التشخيص و المعالجة و ذلك حتي نضمن تحسن مؤشرات الكفاءة النسبية بالزيادة (او بالاختلاف الإيجابي عن الواحد الصحيح) و الجدول (5) يوضح قيم الكفاءة النسبية للمؤشرات:.

الجدول 5: يبين قيم مؤشرات الكفاءة النسبية ($R_2; F; t_{\beta_0}; t_{\beta_1}; MSE$) للبيانات بعد التشخيص و المعالجة الى البيانات الاصلية

رقم ملف البيانات Data File No.	رقم مسلسل No.	عدد مشاهدات النموذج (n)	عدد المشاهدات الشاذة (m)	قيمة الترجيح الاحتمالية W Pr	معامل التحديد (R^2)	قيمة F المحسوبة (F)	قيمة t المحسوبة ($\lvert t_{\beta_0} \rvert$)	قيمة t المحسوبة ($\lvert t_{\beta_0} \rvert$)	متوسط مربعات الخطأ (MSE)
book001 02 012	1	12	2	0.2	1.205	3.040	1.715	1.744	2.626
book002 02 022	2	22	2	0.1	1.185	3.347	1.057	1.829	2.48
book003 02 005	3	5	1	0.25	1.096	3.475	2.331	1.864	1.617
book004 02 050	4	50	10	0.25	1.270	2.56	1.276	1.6	3.539
book005 02 097	5	97	23	0.311	1.342	3.191	2.141	1.786	3.037
book006 02 007	6	7	1	0.167	1.145	2.709	1.133	1.646	4.539
book007 02 012	7	12	3	0.333	1.512	5.001	4.265	2.237	2.930
book008 02 051	8	51	8	0.186	1.112	2.620	1.5	1.619	3.087
book009 02 025	9	25	6	0.316	1.461	3.198	2.222	1.788	3.754
book010 02 056	10	56	16	0.4	1.171	2.067	1.520	1.438	2.763
book011 02 026	11	26	10	0.625	1.214	16.825	0.865	1.113	1.773
book012 02 053	12	53	15	0.395	1.309	2.037	4.599	1.427	2.876
book013 02 013	13	13	2	0.182	1.063	3.316	1.823	1.821	2.909
book014 02 020	14	20	7	0.539	1.447	1.608	1.764	1.268	1

book015 02 011	15	11	5	0.833	2.655	5.418	1.65	2.328	1.674
book016 02 033	16	33	13	0.65	1.824	4.378	1.744	2.092	2.136
book017 02 039	17	39	17	0.773	1.019	1.028	9.54	1.014	1.506
book018 02 067	18	67	6	0.098	1.079	2.155	2.409	1.468	2.143
book019 02 020	19	20	4	0.25	1.003	4	1.956	1.6	2.272
book020 02 051	20	51	16	0.457	1.652	3.507	3.521	2.657	2.6
book021 02 010	21	10	3	0.429	1.379	4.374	7.59	2.092	2.976
book022 02 005	22	5	1	0.25	1.0103	1.536	1.314	1.24	4.133
book023 02 047	23	47	15	0.469	1.521	3.159	2.090	1.778	3
book024 02 030	24	30	4	0.154	1.307	3.665	5.288	1.914	2.934
book025 02 027	25	27	6	0.286	1.06	1.639	3.013	1.280	3.667
book026 02 061	26	61	23	0.605	1.671	3.377	1.556	1.838	1.75
book027 02 053	27	53	26	1	1	1	1	1	1
book028 02 069	28	69	34	1	1	1	1	1	1
book029 02 021	29	21	6	0.4	1.734	3.545	1.635	1.883	3.284
book030 02 015	30	15	1	0.071	1.176	6.984	1.333	2.643	11.523
book031 02 020	31	20	1	0.053	1.071	2.082	1.153	1.443	2.113
book032 02 022	32	22	8	0.571	3.021	8.731	14.392	2.955	1.791
book033 02 024	33	24	7	0.412	1.853	4.528	2.297	2.128	71.321
book034 02 022	34	22	3	0.159	1.092	2.984	2.643	1.728	0.385
book035 02 025	35	25	9	0.563	1.351	1.928	8.194	1.389	380.085

بالنظر للجدول (5) أعلاه نلاحظ ان جميع قيم مؤشرات الكفاءة النسبية $(R_2; F; t_{\beta_0}; t_{\beta_1}; MSE)$ تختلف عن الواحد الصحيح باستثناء النماذج التي يتساوى فيها عدد المشاهدات الشاذة مع نصف او اقل من النصف بمشاهدة واحدة لعدد المشاهدات الكلي.

اختبار الفرضيات

بالنسبة للثلاث فروض التالية يتم تطبيق اختبار t لمقارنة متوسطات العينات المرتبطة باستخدام برنامج التحليل الاحصائي SPSS و كالاتي:.

1/ توجد فروق ذات دلالة معنوية بين مؤشرات النماذج للبيانات الاصلية و لنفس البيانات بعد حذف الشواذ. أي اختبار فرض العدم $H_0: \mu_{i_{original}} - \mu_{i_{detected}} = 0$ على مستوى دلالة $\alpha = 0.025$ مقابل الفرض البديل $H_1: \mu_{i_{original}} - \mu_{i_{detected}} \neq 0$ بعد تطبيق الاختبار كانت النتائج كما في الجدول (6) و كالآتي:.

جدول 6: يبين الفروق بين متوسطات المؤشرات للبيانات الاصلية (orignal) و للبيانات بعد حذف الشواذ (detected)

Paired Samples Test									
		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	97.5% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	R^2 _orignal - R^2 _detected	-.203	.163	.028	-.268	-.138	-7.357	34	.000
Pair 2	F _orignal - F f_detected	-91.71	108.32	18.31	-134.65	-48.77	-5.009	34	.000
Pair 3	t_{β_0} _orignal - t_{β_0} _detected	-6.671	9.845	1.664	-10.57	-2.768	-4.008	34	.000
Pair 4	t_{β_1} _orignal - t_{β_1} _detected	-3.867	2.786	.471	-4.971	-2.763	-8.212	34	.000
Pair 5	MSE _orignal - MSE _detected	670.06	1483.3	250.73	82.083	1258.03	2.672	34	.011

و من النتائج في الجدول (6) و بما ان $Sig.(2-tailed) \le 0.011$ لكل المؤشرات المختبرة اقل من $\alpha = 0.025$ عليه توجد فروق ذات دلالة معنوية بين مؤشرات النماذج للبيانات الاصلية و لنفس البيانات بعد حذف الشواذ، الامر الذي يثبت مدي فعالية الطريقة الشمولية و المقياس البياني في تشخيص المشاهدات الشاذة بالضبط، بعبارة أخرى ان المشاهدات الغير شاذة تشكل نموذجا ذو مؤشرات افضل من مؤشرات نموذج البيانات الاصلية، وعلي الرقم من ذلك لا يمكننا الاعتماد على نتائج هذا النموذج خاصة اذا كان عدد المشاهدات الشاذة المشخصة كبير، مما يعني ان معالجة المشاهدات الشاذة و اعادتها للنموذج ضرورة حتمية .

2/ توجد فروق ذات دلالة معنوية بين مؤشرات النماذج للبيانات الاصلية و لنفس البيانات بعد معالجة الشواذ. أي اختبار فرض العدم $H_0 : \mu_{i_{original}} - \mu_{i_{\text{modulated}}} = 0$ على مستوى دلالة $\alpha = 0.025$ مقابل الفرض البديل $H_1 : \mu_{i_{original}} - \mu_{i_{\text{modulated}}} \neq 0$ بعد تطبيق الاختبار كانت النتائج كما في الجدول (7) و كالآتي:

جدول 7: يبين الفروق بين متوسطات المؤشرات للبيانات الاصلية (orignal) و للبيانات بعد معالجة الشواذ (modulated)

Paired Samples Test									
		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	97.5% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	R^2 _orignal - R^2 _modulated	-.172	.123	.021	-.221	-.123	-8.24	34	.000
Pair 2	F _orignal - F _modulated	-203.0	362.9	61.35	-346.9	-59.14	-3.31	34	.002
Pair 3	t_{β_0} _orignal - t_{β_0} _modulated	-7.190	11.35	1.919	-11.69	-2.691	-3.75	34	.001
Pair 4	t_{β_1} _orignal - t_{β_1} _modulated	-5.125	3.499	.592	-6.512	-3.738	-8.67	34	.000
Pair 5	MSE _orignal - MSE _modulated	665.0	1497	253.1	71.544	1258.5	2.63	34	.013

و من النتائج في الجدول (7) و بما ان $Sig.(2-tailed) \le 0.013$ لكل المؤشرات المختبرة اقل من $\alpha = 0.025$ فأنه يمكننا رفض فرض العدم و قبول الفرض البديل القائل بأنه توجد فروق ذات دلالة معنوية بين مؤشرات النماذج للبيانات الاصلية و لنفس البيانات بعد معالجة الشواذ، الامر الذي يشير الي كفاءة و فاعلية الطرق المقترحة في ازاله اثر المشاهدات الشاذة علي نموذج الانحدار الخطي البسيط، علي الرقم من انه تم مراعات عدد المشاهدات الشاذة أي انه اذا كان عددها قليل يتم معالجتها بالوضع الذي يزيل اثرها بدرجة كبيرة، و اذا كان عددها كبير تتم معالجتها بالوضع الذي يزيل اثرها بدرجة اقل الي ان تترك من دون إزالة اثرها اذا ساوى عددها نصف عدد المشاهدات الكلي او نقص بمفردة واحدة، ما يدل على عدم تحيز طريقة المعالجة لاعتبارها نسبة و جود المشاهدات الشاذة الى المشاهدات الكلية، أي انه اذا كانت نسبة المشاهدات الشاذة صغيرة لا تستحق التأثير في النموذج بدرجة كبيرة و اذا كانت نسبتها كبيرة فيجب تخفيف اثرها على النموذج بدرجة نسبية الامر الذي يجعل طريقة المعالجة تتصف بالمنهجية العلمية.

3/ لا توجد فروق ذات دلالة معنوية بين مؤشرات النماذج للبيانات بعد حذف الشواذ و لنفس البيانات بعد معالجة الشواذ. اي اختبار فرض العدم $H_0 : \mu_{i_{detected}} - \mu_{i_{modulated}} = 0$ على مستوى دلالة $\alpha = 0.025$ مقابل الفرض البديل $H_1 : \mu_{i_{detected}} - \mu_{i_{modulated}} \neq 0$ بعد تطبيق الاختبار كانت النتائج كما في الجدول (8) و كالآتي:.

جدول 8: يبين الفروق بين متوسطات المؤشرات للبيانات بعد حذف الشواذ (detected) و للبيانات بعد معالجة الشواذ (modulated)

Paired Samples Test									
		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	97.5% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	R^2 _detected - R^2 _modulated	.032	.082	.014	-.001	.064	2.29	34	.028
Pair 2	F _detected - F _modulated	-111.3	316.8	53.54	-236.86	14.28	-2.08	34	.045
Pair 3	t_{β_0} _detected - t_{β_0} _modulated	-92.57	380.75	64.36	-243.5	58.35	-1.438	34	.159
Pair 4	t_{β_1} _detected - t_{β_1} _modulated	1.172	4.741	.801	-.708	3.051	1.462	34	.153
Pair 5	MSE _detected - MSE _modulated	-5.052	549.13	92.82	-222.72	212.6	-.054	34	.957

و من النتائج في الجدول (8) و بما ان $Sig.(2-tailed) \geq 0.028$ لكل المؤشرات المختبرة اكبر من $\alpha = 0.025$ فأنه يمكننا قبول فرض العدم القائل بأنه لا توجد فروق ذات دلالة معنوية بين مؤشرات النماذج للبيانات بعد حذف الشواذ و لنفس البيانات بعد معالجة الشواذ، الامر الذي يعني ان الطرق المقترحة تقلل من اثر الشواذ بدرجة كبيرة اذا كان عددها قليل و العكس صحيح اذا كان عددها كبير و تترك الشواذ من دون تقليل لاثرها اذا تساوى او نقص عدد الشواذ بمفردة واحدة من نصف عدد المشاهدات الكلي للنموذج الامر الذي يبرهن منهجية هذه الطرق و يجعلها توافق المنطق، لان الشواذ في هذه الحالة جديرة بهذا الأثر علي النموذج.

4/ متوسط الكفائة النسبية لمؤشرات البيانات بعد معالجة الشواذ الى مؤشرات البيانات الاصلية اكبر من الواحد الصحيح. أي انه من خلال تطبيق اختبار t لمتوسط العينة الواحدة نختبر فرض العدم $H_0 : \mu_i = 1$ على مستوى دلالة $\alpha = 0.025$ مقابل الفرض البديل $H_1 : \mu_i > 1$ بعد تطبيق الاختبار كانت النتائج كما في الجدول (9) و الجدول (10) و كالآتي:

جدول 9: يبين متوسطات الكفاءة النسبية للمؤشرات

One-Sample Statistics				
المؤشرات	N	Mean	Std. Deviation	Std. Error Mean
R^2	35	1.3717661	.44232677	.07476687
F	35	3.6003545	2.81089946	.47512873
t_{β_0}	35	2.9579113	2.88529274	.48770349
t_{β_1}	35	1.7327529	.46821258	.07914237
MSE	35	2.8851770	1.85103194	.31288150

جدول 10: يبين اختلاف متوسطات الكفاءة النسبية للمؤشرات عن الواحد الصحيح

One-Sample Test						
	Test Value = 1					
	t	df	Sig. (2-tailed)	Mean Difference	97.5% Confidence Interval of the Difference	
					Lower	Upper
R^2	4.972	34	.000	.37176611	.1964336	.5470986
F	5.473	34	.000	2.60035454	1.4861510	3.7145581
t_{β_0}	4.015	34	.000	1.95791132	.8142193	3.1016034
t_{β_1}	9.259	34	.000	.73275292	.5471596	.9183462
MSE	6.025	34	.000	1.88517703	1.1514523	2.6189017

من النتائج في الجدول (9) نستنتج ان متوسطات الكفائة النسبية لمؤشرات البيانات بعد معالجة الشواذ الى مؤشرات البيانات الاصلية هي

$$(R_2 = 1.372; F = 3.6; t_{\beta_0} = 2.958; t_{\beta_1} = 1.733; MSE = 2.885)$$،

الامر الذي يشير الى تحسن مؤشرات نموذج الانحدار الخطي البسيط بنسبة 37.2% لمؤشر R^2 و 260% لمؤشر F و 195.8% لمؤشر t_{β_0} و 73.3% لمؤشر t_{β_1} و 188.5% لمؤشر MSE . و من الجدول (10) و بما ان $Sig.(2-tailed) = 0.000$ لكل المؤشرات المختبرة اقل من $\alpha = 0.025$ فأنه يمكننا رفض فرض العدم و قبول الفرض البديل القائل بأن متوسط الكفائة النسبية لمؤشرات البيانات بعد معالجة الشواذ الى مؤشرات البيانات الاصلية اكبر من الواحد الصحيح، مما يدل علي ثبات كفاءة الطرق

المقترحة بالنسبة لتشخيص و معالجة مشكلة الشواذ في نماذج الانحدار الخطي البسيط.

منقاشة النتائج :

بالرجوع للجدول (1) و علي الرقم من تطبيق هذه الطرق علي بيانات نموذج واحد تلاحظ ان عدد المشاهدات المشخصة للطرق السابقة تراوح بين 1 الي 7 مشاهدات شاذة الامر الذي يشير الي عدم توفر معيار ثابت بالنسبة لهذه الطرق في تحديد عدد المشاهدات الشاذة بينما نجد ان الطريقة الشمولية و من خلال استخدام المقياس البياني حددت العدد 6 للمشاهدات الشاذة قبل التعديل و حددت العدد 5 للمشاهدات الشاذة بعد التعديل، أي انه بالنظر الى تصميم المقياس البياني الشكل (1) نلاحظ انه صمم على مبدأ العلاقة العكسية بين مؤشر R^2 و مؤشر MSE و مبدأ النسبية لنفس المؤشرين لذا نلاحظ دقة هذا المقياس في تحديد عدد المشاهدات الشاذة. و بالنظر لبيانات الجدول (2) و (3) و (4) تمثل قيم المؤشرات لبيانات 35 نموذج انحدار خطي بسيط تم استخلاصها لثلاث أوضاع من خلال توفيق النماذج بطريقة المربعات الصغري و تطبيق طرق التشخيص و المعالجة باستخدام برنامج التحليل الاحصائي SPSS و برنامج الجداول الالكتونية Exel مما يشير الي سهولة تطبيق هذه الطرق. و بمقارنة قيم المؤشرات للجدول (3) و الجدول (4) نلاحظ ان قيم المؤشرات للجدول (3) جميعها تتحسن بالمقارنة مع مؤشرات البيانات الاصلية في الجدول (2)، و لكن تتحسن قيم الموشرات بالنسبة للجدول (4) اكثر منه في الجدول (3) اذا قل عدد المشاهدات الشاذة عن ربع عدد المشاهدات الكلي، اما اذا زاد عدد المشاهدات الشاذة عن ربع المشاهدات فأن مؤشرات النماذج في الجدول (3) تكون احسن منه كما في الجدول (4) الى ان تصل لنفس مستوى مؤشرات البيانات الاصلية في الجدول (2) اذا اقترب عدد المشاهدات الشاذة من نصف العدد الكلي تابع النموذج رقم 27 و النموذج رقم 28 بالنسبة للثلاث جداول (2) (3) (4).

النتائج :

في ضوء ما تم عرضه في هذه الورقة العلمية نخلص الي النتائج التالية:

1/ سهولة إمكانية تطبيق الطرق المقترحة من خلال برنامج الجداول الالكترونية Exel و برنامج التحليل الاحصائي SPSS .

2/ اثبتت الطريقة المقترحة للتشخيص و المقياس البياني كفاءة عالية في تشخيص المشاهدات الشاذة.

3/ اثبتت الطريقة المقترحة للمعالجة كفاءة عالية في معالجة المشاهدات الشاذة.

4/ أظهرت الطرق المقترحة للتشخيص و المعالجة تفوقاً كبيراً على طرق التشخيص و المعالجة السابقة.

5/ تعاملت الطرق المقترحة مع المشاهدات الشاذة بمنطق التأثير حسب نسبة وجودها في النموذج.

6/ أظهر تطبيق الطرق المقترحة تحسناً كبيراً في مؤشرات نماذج الانحدار الخطي البسيط ، الامر الذي يدل على انها كانت متاثرة بالمشاهدات الشاذة بدرجة كبيرة.

7/ اتضح ان الطرق المقترحة بإزالتها لاثر المشاهدات الشاذة قد مهدت الطريق امام طريقة المربعات الصغري العادية في عملية توفيق نماذج الانحدار الخطي البسيط.

التوصيات

على ضوء ما تم التوصل اليه من نتائج في هذه الورقة العلمية نوصي بالآتي:

1/ التأكد من وجود المشاهدات الشاذة في بيانات نموذج الانحدار الخطي البسيط و معالجتها حتى اذا أظهرت نتائج توفيق النموذج مؤشرات جيدة لانه بعد الكشف و المعالجة للمشاهدات الشاذة تحصل على مؤشرات للنموذج جيدة بدرجة اكبر.

2/ تطبيق الطرق المقترحة في عمليات تشخيص و معالجة المشاهدات الشاذة لتفوقها على الطرق السابقة.

3/ البحث في إمكانية تضمين هذه الطرق المقترحة في النظام الاحصائي SPSS في جزئية أسلوب تحليل الانحدار الخطي لتصبح بمثابة تحديث للبرنامج الاحصائي SPSS.

المراجع

1- إسماعيل، محمد عبد الرحمن، (2001) *"تحليل الانحدار الخطي"*، معهد الإدارة العامة، المملكة العربية السعودية، مركز البحوث، ص:ص140-142.

2- الجبورى، شلال حبيب، و آخرون (2002)، "اكتشاف و تقدير المشاهدات الشاذة باستخدام معادلة القطع الناقص Ellipse في حالة الانحدار الخطي البسيط"، *المجلة العراقية للعلوم الإحصائية*، العدد 4، المجلد 2، ص:ص 171-186.

3- الجبوري، منى حسين، (1998) *"دراسة تحليل للقيم الشاذة و القيم المفقودة لتصميم المربع اللاتيني و تصميم تام في حالة تكرار مشاهدات العينة"*، رسالة ماجستير مقدمة الي مجلس كلية الإدارة و الاقتصاد، الجامعة المستنصرة، بغداد، العراق.

4- الراوى، خاشع محمود، (1987)، *المدخل إلى تحليل الانحدار*، مديرية دار الكتب للطباعة و النشر، جامعة الموصل، العراق.

5- الشميري، خالد سعد سلطان، و آخرون (2014) "اكتشاف القيم الشاذة و تقديرها في الانحدار الخطي بالتطبيق على بيانات النمو وفقاً لاعمار الأطفال في العام 1987" *مجلة العلوم الطبيعية و الطبية*، جامعة السودان للعلوم و التكنلوجيا، العدد 15، المجلد 2، ص:ص 114-123.

6- عبد الله ، عصام الدين يوسف، (2015)، "تأثير القيم الشاذة في معلمات نموذج الانحدار الخطي المتعدد"، رسالة دكتوراة، جامعة السودان للعلوم و التكنلوجيا، الخرطوم، السودان.

7- Belsley, D. et al, (1980), "*Regression Diogonistics: Identifying Influential Data and Sources of Collinearity*", Wiley, New York, pp:22-24.

8- Github, (2018), "Rdatasets : An archive of datasets distributed with R", available at: *https://vincentarelbundock.github.io/Rdatasets/* (accessed 20 May 2018).

9- Huber, P. J., (1973) "Robust Regression a Symptotics Conjectures Monte Carlo", *Ann. Statist*., 19(2), pp. 136-140.

10- Mickey, O.J. Dunn, et al, "Note on the use of stepwise regression in detecting outliers", *Computers and Biomedical Research*, 1 (1967), pp. 105-111.

11- Neter, J. et al, (1990) "*Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental of Designs*". (3rd edition). I rwin, Homewood, IL 60430, Boston, MA 02116, p:399.

12- Srikantant, K. S. 1961. "Testing for the single outliers in a regression model", *Sankhya-Series* A 23:251-260.