



جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة القادسية
كلية علوم الحاسوب والرياضيات
قسم الإحصاء والمعلوماتية

اقترح ترجيح مركب لتحسين طريقة الـ **BOOTSTRAP** لتقدير معالم نموذج

الانحدار الخطي المتعدد

بحث مقدم

الى مجلس كلية علوم الحاسوب والرياضيات/ قسم الاحصاء والمعلوماتية وهو
جزء من متطلبات نيل درجة البكالوريوس في علوم الاحصاء

أعداد

حيدر عبد الحي ناصر

سرى حسن سامي

إشراف

أ.م جبران عبد الامير اللبان

1438 2017

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

(هُوَ الَّذِي جَعَلَ الشَّمْسُ ضِيَاءً وَالْقَمَرَ نُورًا وَقَدَرَهُ مَنَازِلَ لِتَعْلَمُوا
عَدَدَ السِّنِينَ وَالْحِسَابَ ۚ مَا خَلَقَ اللَّهُ ذَلِكَ إِلَّا بِالْحَقِّ ۚ يُفَصِّلُ
الآيَاتِ لِقَوْمٍ يَعْلَمُونَ)

صدق الله العلي العظيم

الاهداء

الى ..

من غمرني بالفضل واختصني بالنصح وتفضل علي بقبول الاشراف على بحثي استاذي ومعلمي ..

أ.م جبران عبد الامير اللبان ...

الى

من منحنا وقته ووهبنا من علمه فكان عطائه لنا كالنهر الجاري وعاش يتطلع لي مستقبلا مشرقا ويتمني لي نجاحا باهرا الحمد لله الذي مد في عمرك كي ترى هذا النجاح ... ابي العزيز

الى

من سهرت عينها لتغني عيني وبذعائها استجاب ربي دعواي قد حاز الوقت لترى ثمرة صبرك على شجرة النجاح ... امي الغالية

الى ...

كل من شجعني وساعدني لاقدم هذا الجهد المتواضع لاسيما اخوتي واخواتي

واخر دعوانا ان الحمد لله رب العالمين

الشكر والتقدير

لا بد لنا ونحن نخطو خطواتنا الأخيرة في الحياة الجامعية من وقفه نعود الى اعوام قضيناها
في رحاب الجامعة مع اساتذتنا الكرام الذين قدموا لنا الكثير باذلين بذلك جهودا كبيرة في بناء
جيل الغد لتبعث الامه من جديد وقبل ان نمضي تقدم اسمى ايات الشكر والامتنان
والتقدير والمحبة الى الذين عملوا اقدس رساله في الحياة الى الذين مهدوا لنا
طريق العلم والمعرفة الى جميع اساتذتنا الافاضل

"كن عالما فان لم تستطع فكن متعلما ، فان لم تستطع فأحب العلماء ، فان لم تستطع فلا
تبغضهم"

واخص بالشكر والتقدير واتقدم بخالص الشكر الجزيل والعرفان بالجميل والاحترام والتقدير لمن
غمرني بالفضل واختصني بالنصح وتقض علي بقبول الاشراف على بحثي استاذي
ومعلمي أ.م جبران عبد الامير اللبان ..

المستخلص

المستخلص

مقترح هذا البحث بهدف الى جعل طريقة الـ Bootstrap المقترحة من قبل (Efron) 1979 التي سنطلق عليها (BOLS) طريقة حصينة عند ظهور القيم الشاذة في كلا الاتجاهين (x, y) الذي ينتج عنه تقديرات حصينه لنموذج الانحدار الخطي المتعدد ، و ذلك باستخدام دالة هاميل لترجيح بواقي الانحدار المستحصل من طريقة المربعات الصغرى وذلك لاختزال تأثير القيم الشاذة و من ثم تخصيص احتمال لقيمة قيمة من قيم البواقي عن طريقة قسمة وزنها الترجيحي على مجموع الاوزان الترجيحية لغرض استخدامها مع البوتستراب الذي سوف لا يعطي فرصة للباقي ذو الاحتمال القليل بالظهور في عيناته الفرعية. لكن قبل تشغيل خوارزمية البوتستراب اقترحنا معالجة مشكلة نقاط القوة الرافعة الرديئة في بيانات المتغيرات التوضيحية، اذا تم استخدام مسافة مهانلوبيز الحصينة لتشخيص هذه النقاط و بناء وزن مرجح اخر للمتبعي من بيانات المتغيرات التوضيحية. المقصود بالمتبعي هنا هو ما تبقى من هذه البيانات بعد حذف الصفوف المناظرة لمواقع القيم الشاذة المشخصة سابقا بالاوزان الترجيحية لدالة هاميل. لقد اظهرت الطريقة المقترحة كفاءة عالية جدا عند مقارنتها مع طريقة BOLS باستخدام بيانات حقيقية، اذ تم اعتماد الجذر التربيعي للـ MSE لمعاملات المقدرة كمعيار للمقارنة. من النتائج تبين ان اداء طريقتنا المقترحة اكفى و افضل من الطريقة التي تمت المقارنة بها و ان المعاملات المقدرة اكثر دقة. لذلك نوصي باستخدام الطريقة المقترحة عند وجود مثل هذا النوع من المشاهدات

محتويات الفهرست

الفصل الأول		
المقدمة		
الصفحة	الموضوع	الرقم
1	المقدمة	1-1
2	الدراسات السابقة	1-2
3	مشكلة البحث	1-3
3	هدف البحث	1-4

الفصل الثاني

الجانب النظري

الرقم	الموضوع	الصفحة
2-1	مقدمة عن الانحدار الخطي	4
2-2	Concept of Outlier مفهوم القيم الشاذة	6
2-3	اسباب ظهور القيم الشاذة	8
2-4	الأسلوب المقترح	8

الفصل الثالث

الجانب التطبيقي

الرقم	الموضوع	الصفحة
3-1	وصف البيانات	14
3-2	تشخيص القيم الشاذة بالرسم	14
3-3	النتائج	18

الفصل الرابع

الاستنتاجات

الرقم	الموضوع	الصفحة
4-1	الاستنتاجات والتوصيات	19

المستخلص

المستخلص

مقترح هذا البحث بهدف الى جعل طريقة الـ Bootstrap المقترحة من قبل (Efron) 1979 التي سنطلق عليها (BOLS) طريقة حصينة عند ظهور القيم الشاذة في كلا الاتجاهين (x, y) الذي ينتج عنه تقديرات حصينه لنموذج الانحدار الخطي المتعدد ، و ذلك باستخدام دالة هاميل لترجيح بواقي الانحدار المستحصل من طريقة المربعات الصغرى وذلك لاختزال تأثير القيم الشاذة و من ثم تخصيص احتمال لقيمة قيمة من قيم البواقي عن طريقة قسمة وزنها الترجيحي على مجموع الاوزان الترجيحية لغرض استخدامها مع البوتستراب الذي سوف لا يعطي فرصة للباقي ذو الاحتمال القليل بالظهور في عيناته الفرعية. لكن قبل تشغيل خوارزمية البوتستراب اقترحنا معالجة مشكلة نقاط القوة الرافعة الرديئة في بيانات المتغيرات التوضيحية، اذا تم استخدام مسافة مهانلوبيز الحصينة لتشخيص هذه النقاط و بناء وزن مرجح اخر للمتبعي من بيانات المتغيرات التوضيحية. المقصود بالمتبعي هنا هو ما تبقى من هذه البيانات بعد حذف الصفوف المناظرة لمواقع القيم الشاذة المشخصة سابقا بالاوزان الترجيحية لدالة هاميل. لقد اظهرت الطريقة المقترحة كفاءة عالية جدا عند مقارنتها مع طريقة BOLS باستخدام بيانات حقيقية، اذ تم اعتماد الجذر التربيعي للـ MSE لمعاملات المقدرة كمعيار للمقارنة. من النتائج تبين ان اداء طريقتنا المقترحة اكفى و افضل من الطريقة التي تمت المقارنة بها و ان المعاملات المقدرة اكثر دقة. لذلك نوصي باستخدام الطريقة المقترحة عند وجود مثل هذا النوع من المشاهدات

الفصل الأول

1-1 المقدمة

1-2 الدراسات السابقة

1-3 مشكلة البحث

1-4 هدف البحث

1-1 المقدمة

ان Bootstrap هو أسلوب إعادة معاينة و يعتبر حالة خاصة من أسلوب المحاكاة بالمونت كارلو المستندة الى التقدير اللامعلمي لتقدير التوزيع لدالة المشاهدات ، هذا التوزيع يتم ايجاده بأبدال التوزيعات غير المعروفة بالتوزيع التجريبي للبيانات في تعريف الدالة الإحصائية ثم إعادة معاينة البيانات لايجاد توزيع المتغيرات العشوائية الناتجة . لقد برز أسلوب Bootstrap نتيجة المشاكل التي تعرض لها أسلوب Jackknife الذي قدمه (Quenouille –Tukey 1956-1958) لتقليل التحيز وتقدير الفترة ثقة عالية الدقة اثناء تطبيقه على نموذج الانحدار . اذ اعطى Miller اول تفصيل (1979) لاستخدام Jackknife بنموذج الانحدار وذلك بشطب مشاهدة واحدة عشوائية وإعادة المعاينة من دون ارجاع الا ان هذه الطريقة تعرضت لمشكلتين الا و هما اهمال الطبيعة غير المتزنة لبيانات الانحدار و تقييد اختبار حجم إعادة العينة.

واستمر التفكير بأسقاط عدد من المشاهدات في كل مرة الا انها ما زالت تتضمن عيوب اسقاط مشاهدة واحدة لذلك تبني Efron (1979) فكرة الـ Bootstrap بأسلوب اكثر دقة تمثل باستخدام عينة عشوائية بسيطة مع الارجاع لإعادة معاينة البيانات الذي يعود أساسه النظري الى المحاضرة التي القاها Efron (1977) في جامعة stanford وناقش خلالها المشكلة الاتية $X = (x_1, \dots, x_n)$ بحجم n مسحوبة من مجتمع له توزيع احتمالي غير معروف F ثم بين Efron إمكانية تطبيقه على نموذج الانحدار وبتطبيق الطرائق التقليدية ذاتها من خلال أسلوب إعادة المعاينة مع الارجاع (المتمثل بأعادة معاينة البواقي او إعادة معاينة ازواج المشاهدات) .

المتعارف عليه ان طريقة المربعات الصغرى الاعتيادية OLS الاكثر شيوعا لتقدير معلمات نموذج الانحدار الخطي عند تحقق فرضياتها و اهم ميزاتها انها تقدم افضل تقديرات خطية غير متحيزة. من هذه الفرضيات ان الخطأ العشوائي يتوزع طبيعيا بمتوسط قيمته صفرا و تباين ثابت. و من المفترض توفر هذه الفرضيات عند تطبيق البوتستراب. و لحل مشكلة عدم التجانس (عدم ثبات التباين في بواقي الانحدار) مثلا اقترح Wu عام 1986 طريقة إعادة معاينة البواقي البوتسترايبية و قدم هذا المقترح بعد اجراء تعديل عليه من قبل Liu (1988) واطلق عليه Wild Bootstrap .

للاسف قد لا تتحقق فرضية التوزيع الطبيعي للاخطاء العشوائية او بواقي الانحدار عند وجود قيمة شاذة واحدة على الاقل في بيانات عينة الدراسة Hampel (1973) التي من اسباب ظهورها ان توزيعها مختلف عن توزيع بقية الاخطاء العشوائية. و قد يكون توزيع هذه الشواذ طبيعيا لكن بمتوسط قد لا يساوي صفرا او تباين كبير جدا يختلف عن تباين نظيراتها و يعود ظهورها الى عدة اسباب منها انها دخلت بسبب اخطاء القياس او خطأ بشرياً عند تسجيل هذه البيانات .

اعادة المعاينة مع الارجاع ستسبب مشكلة جدية لتقدير معلمات الانحدار باستخدام الـ Bootstrap اذا ما وجدت قيمة شاذة واحدة على الاقل في بيانات عينة الدراسة ، وذلك لان عملية السحب و الارجاع العشوائي قد يعطي فرصة ظهور القيمة الشاذة اكثر من مرة في اي عينة من عينات الـ Bootstrap حيث ان احتمال سحب المشاهدات من العينة متساو و لذلك ربما تتجاوز نسبة وجود الشواذ في عينة الـ Bootstrap قيمة الـ Breakdown Point (BP) لطريقة التقدير المستخدمة. للتغلب على هذه المشكلة قدمت عدد من المقترحات التي اثرت الادبيات الاحصائية.

1-2 الدراسات السابقة

لقد اهتم الباحثون في مجال العلوم الاحصائية بمعالجة مشكلة الشواذ و تحسين عينات الـ Bootstrap (Barrera and Zamar,2002) اقترحا تقدير تحسين المعلمات المقدرة، ثم عاد (Barrera,2003) ليقدم مقترح لزيادة سرعة الخوارزمية و استقرارها. وليام عام 2005 اقترح تطبيق البوتستراب مع طريقة المربعات الصغرى المشدبة (LTS) المقترحة من قبل Rousseeuw و Leroy عام 1987 انظر (Willems,2005). (Imon & Ali,2005) اقترحا خوارزمية تدعى (Diagnostic-Before Bootstrap) لاحتساب معالم الانحدار الخطي مفادها هو تشخيص القيم الشاذة وحذفها قبل اجراءات اعادة المعاينة على المشاهدات المتبقية بدون قيم شاذة. اجراءات تشخيص القيم الشاذة اعتمدت على طريقة Re-weighted Least Squares (RLS) لبواقي الانحدار المقترحة من قبل Rousseeuw and Leroy (1987) باستخدام دالة الوزن الاتية:

$$W_i = \begin{cases} 0, & \text{if } \text{abs}(r_i / s) > 2.5 \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

حيث r_i تمثل بواقي الانحدار اما S تمثل تقدير قياس حصين مكافيء الى

$$s = 1.4826(1 + 5/(n - p))\sqrt{\text{med } r_i^2(\hat{\theta})}$$

حيث n حجم العينة و p عدد معاملات الانحدار.

تقنية اخرى اقترحت من قبل (Uraibi et al,2009) تتمثل بخوارزمية ديناميكية حصينة لها القدرة على تشخيص القيم الشاذة داخل عينات الـ Bootstrap من جانب وتشذيبها من جانب اخر. اذ تتميز هذه التقنية برفض كل العينات المعادة التي تتضمن قيم شاذة تتجاوز قيمة BP ، و امتياز اخر يتمثل بانها تشذب القيم الشاذة حصرا و لا تكون هناك اية خسارة في المشاهدات غير الشاذة، و ذلك من خلال استخدام تقنية LTS داخل الدورات التكرارية للـ Bootstrap على ان نسبة التشذيب تعتمد على التشخيص الديناميكي.

(Ramli,2008) اقترحت تقنية Weighted Bootstrap with Probability (WBP) بالاعتماد على خوارزمية Bootstrapping with Re-sampling Probabilities (Tille,2006) التي تقترح تخصيص احتمال لكل عينة من العينات الممكنة الظهور في عينات الـ Bootstrap ومن ثم اجراء عملية محاكاة تبني على قيمة عشوائية تتبع توزيع منتظم (0,1). هذه القيم العشوائية تشكل نظام تحكم في اختيار عينات دون سواها وفق قاعدة محددة.

اقترحت (Ramli,2008) الاستفادة ايضا من اقتراح (Venables and Ripley,1999) في تشخيص القيم الشاذة من خلال بواقي طريقة وسيط المربعات الصغرى LMS، و من ثم رجحت هذه البواقي بدالة (Hampel,1971) ، لغرض تحديد احتمال ظهور كل مشاهدة من خلال معرفة نسبة الوزن الترجيحي للمشاهدة الى مجموع كل الاوزان الترجيحية. يتضح مما تقدم انها تمكنت من تحويل خوارزمية (Tille,2006) لتخصيص احتمال مرجح للمشاهدة بدلا من العينة، وبالتاكيد ان دالة (Hampel,1971) ستخصص وزن ترجيحي مساو للصفر للقيم الشاذة مما يؤدي الى اخذها للاحتمال صفر الذي سيحرمها من فرصة الظهور عند اعادة المعاينة مرة واحدة او اكثر. هذا يعني ان مشاهدات عينات الـ Bootstrap ستكون نظيفة جدا و غير ملوثة، مما يمكنها من اجراء التقدير باستخدام طريقة المربعات الصغرى.

1-3 مشكلة البحث

ركزت الدراسات السابقة على حصانة تقدير معلمات الانحدار بوجود القيم الشاذة و لم تتطرق الى نوع اخر من الشواذ الذي يظهر في مصفوفة المتغيرات التوضيحية و الذي يسمى نقاط القوة الرافعة (Leverage point). بناء على ذلك مشكلة هذه الدراسة ان طريقة البوتستراب لا تعطي تقديرات حصينة بوجود نقاط القوة الرافعة او القيم الشاذة.

1-4 هدف البحث

اقتراح اسلوب لتحسين طريقة البوتستراب عند ظهور كافة انواع القيم الشاذة و نقاط الرفع الرديئة.

الفصل الثاني

الجانِب النظري

2-1 مقدمة عن الانحدار الخطي

يعد تحليل الانحدار الخطي واحداً من أهم الأساليب الإحصائية الواسعة التطبيق في مختلف الأبحاث العلمية و الإنسانية. يفترض هذا النوع من التحليل نموذجين الأول يدعى بنموذج الانحدار الخطي البسيط عندما يحتوي النموذج على متغير مستقل واحد يؤثر على متغير معتمد واحد، أما النموذج الثاني بنموذج الانحدار الخطي المتعدد الذي يكون فيه مجموعة من المتغيرات المستقلة تؤثر على متغير معتمد واحد. نموذج الانحدار الخطي عموماً يمكن أن يكتب بالصيغة الجبرية للمصفوفات كالآتي:

$$y = x\beta + \varepsilon \quad (1)$$

عندما y هو متجه يمثل مشاهدة للمتغير المعتمد، X مصفوفة $p \times n$ من متغيرات المستقلة مضافاً له الحد الثابت، β هو متجه لمعاملات الانحدار و هو متجه من الأخطاء العشوائية المفترض أن تكون مستقلة و تتوزع طبيعياً بمتوسط قيمته صفر و تباين ثابت لعينة بحجم n من المشاهدات. فرضياته :-

- 1- أن العلاقة بين المتغيرات المستقلة (x) والمتغير المعتمد y Dependent هي علاقة خطية.
- 2- أن يتبع المتغير المعتمد (y) التوزيع الطبيعي أما المتغيرات المستقلة (x) تكون ثابتة.
- 3- أن يتبع الخطأ العشوائي (ε) التوزيع الطبيعي Normal Distribution بوسط حسابي Mean يساوي (0) وتباين ثابت Variance يساوي (σ^2) عند كل قيمة من قيم المتغير المستقل (x).
- 4- أن يكون الخطأ العشوائي غير مرتبط بالمتغير المستقل.
- 5- أن يكون تباين الخطأ العشوائي متجانس Homoscedasticity عند كل قيمة من قيم المتغير المستقل (x).
- 6- أن لا يوجد ارتباط ذاتي Autocorrelation بين الأخطاء العشوائية Random Error

عند تحقق الفرضيات أعلاه فإن التقدير بطريقه المربعات الصغرى يعد افضل تقدير خطي غير متحيز، حيث يمكن الحصول على تقدير متجه معاملات الانحدار كالتالي :

$$\hat{\beta} = (X'X)^{-1}X'y \quad (2)$$

اما القيم التقديرية للتنبؤ التي تعتمد على دقة تقدير المعلمات من خلال تحقق الفرضيات المذكوره سابقا فيمكن الحصول عليها كالآتي :

$$\hat{y} = X'\hat{\beta} \quad (3)$$

عندئذ يمكن حساب بواقي الانحدار من خلال الفرق ما بين القيم التقديرية والقيم الحقيقية للمتغير المعتمد وكالآتي :

$$\hat{\varepsilon} = y - \hat{y} \quad (4)$$

عليه فإن تباين بواقي (الاحطاء العشوائية) الانحدار تكون كالآتي :

$$\sigma_{\varepsilon}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p-1} \quad (5)$$

اما متوسط مربعات الخطأ العشوائي للمعلمات المقدرة لنموذج الانحدار الخطي المعلمي فيمكن الحصول عليه من خلال المعادلة الآتية :

$$MSE(\hat{\beta}) = Var(\hat{\beta}) + Bias(\hat{\beta})^2 \quad (6)$$

حيث ان

$$Var(\hat{\beta}) = (X'X)^{-1}\sigma_{\varepsilon}^2$$

$$Bias(\hat{\beta}) = \beta - \hat{\beta}$$

من الواضح جدا في العلاقة الخطيه بين متوسط مربعات معلمات الانحدار المقدرة وتباين هذه المعالم ومربع تحيزها عن معلمات الانحدار الحقيقيه في مجتمع الدراسة ، وان $MSE(\hat{\beta})$ عباره عن مركبه خطية من جزئين كل جزء منهما يمثل دالة تربيعية ويتأثر بكليهما او باحدهما ، اذ ان التقدير الجيد للمعلمات يقلل من قيمة التحيز والتوصيف الدقيق لمشكله الدراسه وجمع البيانات يزيد من تجانس الاحطاء العشوائية وبالتالي يؤدي الى تقليل تباين الاحطاء الذي بدوره يقلل من تباين المعلمات المقدرة .

لكن تحقيق هذه الحالة غير واقعي في التعامل مع البيانات الحقيقة من حيث حجم العينة التي اذا كانت صغيرة ستؤدي الى تقديرات غير دقيقة قياسا بالعينات الكبيره حسب النظرية الاحصائية مما يخل في تحقيق عملية التوازن ما بين مركبة التباين ومركبة التحيز المؤثرتان في $MSE(\hat{\beta})$.

بمعنى اخر انه قد يحصل تحيز عالي ولكن تباين قليل والعكس صحيح وبالتالي الاعتماد على قيمة $MSE(\hat{\beta})$ في التحليل او الاستنتاج او الاستدلال يكون غير واقعي لاتخاذ القرار الاحصائي حول الظاهرة المدروسة مالم تأخذ تحقيق كامل فرضيات المربعات الصغرى في نظر الاعتبار او ايجاد طرق بديلة

2-2 مفهوم القيم الشاذة : Concept of Outlier

للقيم الشاذة تعريفات عدة نذكر منها ما يلي :

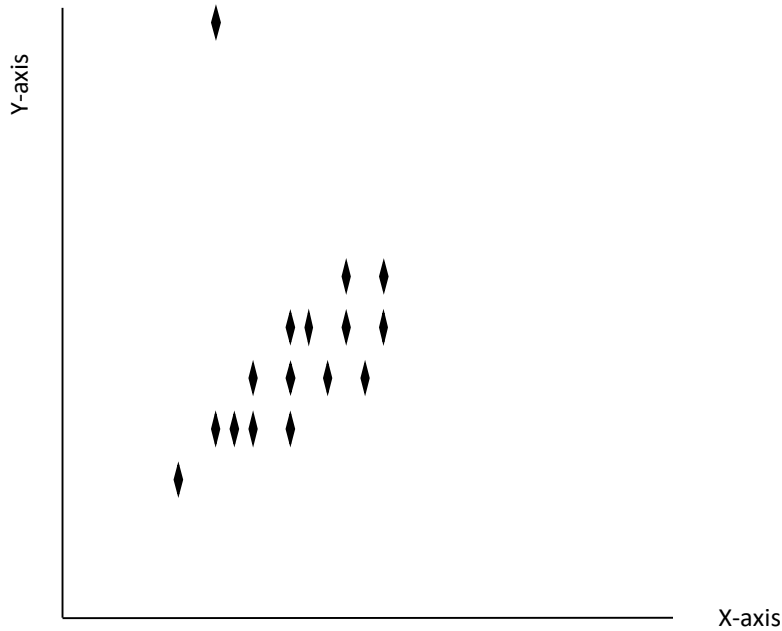
عرف Bross المشاهدة الشاذة بأنها تلك المشاهدة التي تظهر منحرفة بشكل كبير عن سائر مكونات العينة التي وجدت فيها تلك العينة .

أما Freeman فقد عرف القيمة الشاذة بأنها أية مشاهدة لم تتولد بالطريقة العامة التي ولدت الأغلبية العظمى من مشاهدات البيانات.

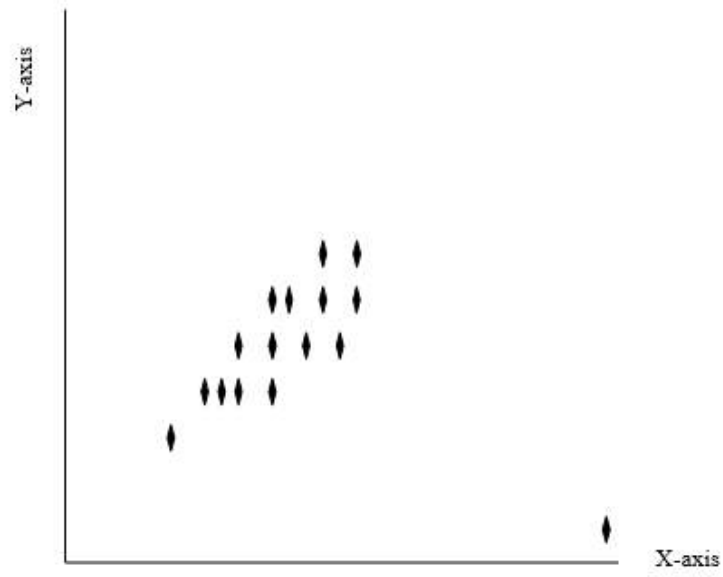
أما Barnett فقد عرفت المشاهدة الشاذة في مجموعة البيانات هي تلك المشاهدة التي تبدو غير منطقية إذا قورنت بسائر مجموعة البيانات .

أما Keller عرف المشاهدات الشاذة بأنها المشاهدات التي تقع بعيدة عن معادلة الانحدار ويكون لها خطأ كبير مقارنة ببقية المشاهدات الطبيعية الأخرى في مجموعة البيانات و ذلك سيكون لها تأثير في النموذج الخطي و تقديراته .

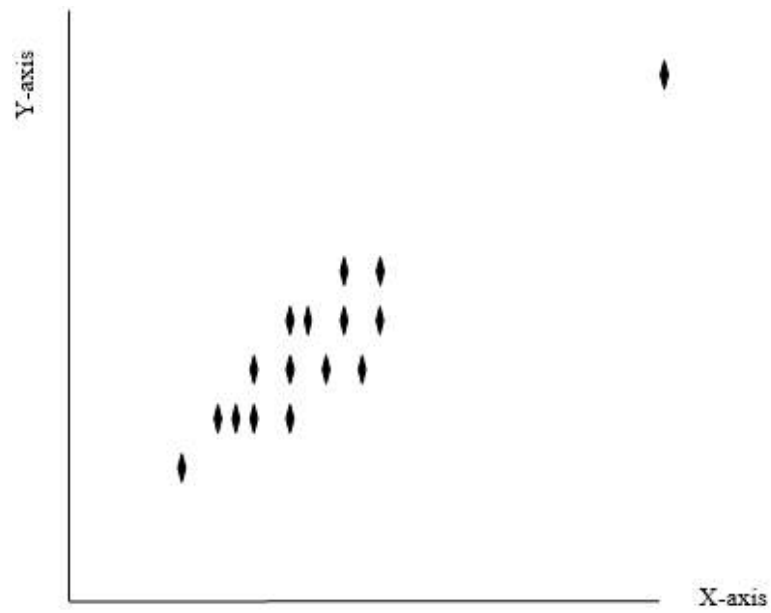
الشواذ التي تظهر في المتغير العشوائي Y سنطلق عليها شاذة عموديا و التي يالتاكيد ستظهر في بواقي الانحدار، اما تلك التي تظهر في مصفوفة المتغيرات التوضيحية تدعى نقاط القوة الرافعة Leverage points و منها يضر بالتقدير و اخر مفيد يحسن من دقة التقدير. الشكل 1،2 و 3 يوضح انواع الشواذ.



الشكل 1- القيمة الشاذة عموديا



الشكل -2- نقطة قوة رافعة مضرة



الشكل -3- نقطة قوة رافعة جيدة

2-3 اسباب ظهور القيم الشاذة

قد تظهر القيم الشاذة في مجموعة البيانات نتيجة لأسباب عدة منها ما يلي :

1- إن البيانات تعود إلى توزيعات غير متماثلة أي يكون فيها التواء عال نحو اليمين أو نحو اليسار، ولصياغة نموذج لهذه التوزيعات قدم Green,1976 تصنيفاً لعوائل هذه التوزيعات توزيعات عرضة للقيم الشاذة وهي التي تكون لها نهايات تؤول إلى الصفر ببطء وهذه التوزيعات تكون عرضة للقيم الشاذة بصورة مطلقة. وتوزيعات تكون مقاومة للقيم الشاذة التي تؤول نهاياتها إلى الصفر بشكل أسرع من سابقتها وتكون مقاومة للقيم الشاذة بصورة مطلقة

2- تأتي البيانات من نوعين من التوزيعات أحدهما التوزيع الأساسي Basic Distribution والذي يولد مشاهدات جيدة بينما الآخر يسمى التوزيع الملوث Contaminating distribution والذي يولد قيماً شاذة .

3- قد تحدث القيم الشاذة نتيجة لأسباب أخرى منها أخطاء يقع فيها الباحث عند رصد القياسات أو نتيجة إلى وجود خلل في أجهزة القياسات وخاصة في التجارب المختبرية ، أو نتيجة أخطاء في الحسابات مما يؤدي إلى ظهور القيم الشاذة .

2-4 الاسلوب المقترح

كما اسلفنا ان مشكلة طريقة البوتستراب هي احدى طرق اعادة المعاينة مع الارجاع و التي يمكن وصفها بالاعتماد على بواقي الانحدار كالاتي

1. تقدير معلمات الانحدار باستخدام طريقة المربعات الصغرى
2. احتساب بواقي الانحدار $\hat{r}_i = Y - \hat{Y}_i$
3. سحب عينة عشوائية مع الارجاع من هذه البواقي بحيث يكون حجم العينة المسحوبة يساوي حجم العينة الاصلية. $\hat{Y}_i^b = f(X\hat{\beta}) + \hat{r}_i^b$
4. نقدر معلمات البوتستراب للعينة المسحوبة كالاتي $\hat{\beta}^b = (X'X)^{-1} \hat{Y}_i^b$
5. تعاد الخطوات 3 و 4 لـ B من المرات
6. المعلمات المقدرة لـ B من عينات البوتستراب و تبين هذه المعلمات المقدرة هو كالاتي

$$\overline{\hat{\beta}_{boot}} = \frac{1}{B} \sum_{b=1}^B \hat{\beta}^b \quad b = 1, 2, \dots, B$$
$$Cov(\overline{\hat{\beta}_{boot}}) = \frac{1}{B-1} \sum_{b=1}^B (\hat{\beta}^b - \overline{\hat{\beta}_{boot}})(\hat{\beta}^b - \overline{\hat{\beta}_{boot}})'$$

من الواضح جدا ان BOLS تعتمد على بواقي الانحدار حصرا، لذلك هي تفشل في اعطاء تقديرات حصينة بوجود الشواذ و نقاط الرفع الضارة في بيانات العينة، لذلك اقترحنا معالجة هذه المشكلة و كالاتي :

أ. تقدير معلمات حصينة لنموذج الانحدار باستخدام طريقة وسيط المربعات الصغرى Least Squares (LMS) Median of Squares المقترحة من قبل Rousseeuw (1984) ، التي تفترض اختيار كل العينات الممكنة بحجم $h = \frac{n}{2} + \frac{(p+1)}{2}$ من العينة الاصلية ، ثم يتم تقدير معلمة الانحدار لكل عينة فرعية باستخدام طريقة المربعات الصغرى و احتساب مربعات بواقي الانحدار لكل عينة فرعية. العينة الفرعية التي تحقق $\min_{\beta} \text{median}(r_i^2)$ تعد مقدراتها مقدرات حصينة.

ب. احتساب بواقي الانحدار لكامل العينة بالاعتماد على المقدرات الحصينة و من ثم تقييسها وفقا للمعادلة (7)

$$u_i = \frac{r_i}{\text{median}(r_i)} \quad (7)$$

حيث r_i تمثل بواقي LMS و $\text{median}(r_i)$ يمثل وسيط هذه البواقي.
ج. ترجيح البواقي المقيسة باستخدام دالة (Hampel,1971) الموزونة، و كالاتي

$$W(u) = \frac{\psi_{Hampel}(u)}{u} \quad (8)$$

حيث ان

$$\psi_{Hampel}(u) = \begin{cases} u & , 0 \leq \text{abs}(u) \leq a \\ a \text{ sign}(u), & , a \leq \text{abs}(u) \leq b \\ a(c - \text{abs}(u))/(c - b) \text{sign}(u) & , b \leq \text{abs}(u) \leq c \\ 0 & , c \leq \text{abs}(u) \end{cases} \quad (9)$$

حيث تأخذ القيم الافتراضية a, b, c على التوالي 2، 4، 8. من الواضح ان القيم الشاذة في بواقي الانحدار ستحصل على الوزن صفر و بالتالي الاحتمال المخصص وفقا للمعادلة (8) سيكون صفرا ايضا.

د. نحسب الاحتمال المخصص لكل باقي من البواقي وفقا للمعادلة 10

$$p_i = \frac{W_i}{\sum_{j=1}^n W_j} \quad 0 \leq p_i \leq 1 \quad \text{and} \quad i = 1, 2, \dots, n. \quad (10)$$

هـ. الان نقوم بتجزئة مصفوفة المتغيرات التوضيحية و متجه المتغير المعتمد الى جزئين بحث الصف i الذي حصل على احتمال $P_i = 0$ يكون في الجزء D اما بقية الصفوف فستدخل في الجزء R بحيث يمكن تشكيل التجزئة الاتية:

$$Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix} \quad X = \begin{bmatrix} X_R \\ X_D \end{bmatrix}$$

حيث R هي مجموعة المشاهدات الباقية (البيانات النظيفة) في التحليل و D تمثل مجموعة المشاهدات المحذوفة (الشاذة).

لكن هذا الاجراء لايعالج مشكلة نقاط قوة الرفع في المتغيرات التوضيحية، عليه اقترحنا تشخيص نقاط قوة الرفع باستخدام Mahalanobis Distance (MD) التي اقترحها Mahalanobis عام (1936). هذه الطريقة تقيس انحرافات نقاط البيانات عن مركزها. المشاهدة i في مصفوفة المتغيرات التوضيحية يمكن ان يعبر عنها كمتجه صفي $(1, X_i) = (1, x_{i1}, x_{i2}, \dots, x_{ik})$ عندما $X_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ متجه صفي طوله k . لكن هذا الاجراء غير حصين ايضا بسبب تأثر الوسط الحسابي و التباين بالقيم الشاذة لذلك استخدمنا مصفوفة التباين و التباين المشترك الحصينة Fast-Minimum Covariance Determinant (FAST-MCD) المقترحة من قبل Rousseeuw and Driessen عام 1999 و التي بإمكانها ان تزودنا بمراكز حصينة لكل متغير توضيحي $T(X)$.

$$FAST - MCD(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - T(X))'(X_i - T(X)) \quad (11)$$

بناء على ما تقدم يمكن الحصول على MD الحصينة للمشاهدة i كالآتي :

$$MD_i = \sqrt{(X_i - T(X))' FAST - MCD(X)^{-1} (X_i - T(X))} \quad i = 1, 2, \dots, n \quad (12)$$

و لتشخيص نقاط قوة الرفع LP اخترنا العتبة لتكون الجذر التربيعي لمربع كاي $\sqrt{\chi^2_{(0.975,p)}}$ بحيث
المشاهدة i التي تملك $MD_i < \sqrt{\chi^2_{(0.975,p)}}$ تعتبر نقطة قوة رافعة لذلك ترجح X_R بالدالة الترجيح
الآتية:

$$W_R = \begin{cases} MD_i / \sqrt{\chi^2_{(0.975,p)}} & \text{if } MD_i > \sqrt{\chi^2_{(0.975,p)}} \\ 1 & \text{Otherwise} \end{cases} \quad (13)$$

للحصول على تقديرات حصينة لنموذج الانحدار يمكن استخدام طريقة المربعات الصغرى للمشاهدات
المتبقية في الجزء R ، بمعنى فقط Y_R و X_R .

تقدير معلمات الانحدار الحصينة باستخدام المربعات الصغرى يكون وفقا للمعادلة (2) كالآتي

$$\hat{\beta}^{(-D)} = (X_R' W_R X_R)^{-1} Y_R' X_R' W_R \quad (14)$$

اما البواقي سوف لن تكون خالية من القيم الشاذة و ان تم استخدام مقدر حصين لاحتساب القيم التقديرية
للمتغير المعتمد $\hat{Y}_i = f(X, \hat{\beta}^{(-D)})$ و ذلك بسبب ان الشواذ عموديا مازالت حاضرة في قيم المتغير
المعتمد الحقيقية

$$\hat{r}_i = Y - \hat{Y}_i^{(-D)} \quad (15)$$

بالإضافة الى ان نقاط قوة الرفع ستكون ايضا حاضرة في مصفوفة المتغيرات التوضيحية X ، بناء على المعطيات المتوفرة التي فان تطبيق تقنية البوتستراب المرجحة بالاحتمال يصبح كالآتي:

سحب اول عينة Bootstrap عشوائيا مع الارجاع من بواقي الانحدار $\hat{r}_i$ و بحجم يكافيء طول المتجه Y_R للحصول على $\hat{r}_i^{b(-D)}$ مع مراعات اخذ الاحتمالات بنظر الاعتبار يبحث الباقي الذي احتماله صفرا لا يحظى بفرصة للظهور في عينات الـ Bootstrap.

تحسب القيم التقديرية للعينة المسحوبة في (1) وفقا لهذه المعادلة

$$\hat{Y}_i^{b(-D)} = f(X_R W_R, \hat{\beta}^{(-D)}) + \hat{r}_i^{b(-D)}$$

و من ثم نقدر معاملات الانحدار باستخدام المربعات الصغرى الاعتيادية

$$\hat{\beta}^{b(-D)} = (X_R' W_R X_R)^{-1} W_R \hat{Y}_i^{b(-D)}.$$

اعادة الخطوات اعلاه لـ B من المرات (حيث B عدد طبيعي اكبر من 2) و ان $b = 1, 2, \dots, B$.

عندئذ يمكن احتساب معاملات الانحدار التقديرية للطريقة المقترحة (MWBP) كالآتي :

$$\overline{\hat{\beta}_{boot}} = \frac{1}{B} \sum_{b=1}^B \hat{\beta}^{b(-D)} \quad b = 1, 2, \dots, B \quad (16)$$

و لغرض احتساب MSE للمعالم المقدرة في المعادلة (6) ، يجب احتساب التحيز ما بين المعالم المقدرة بـ BOLS مع المعلمات التي تم تقديرها مسبقا قبل تطبيق BOLS اذا لم تتوفر لدينا معلمات الانحدار لمجتمع الدراسة،

$$bias_{boot} = \overline{\hat{\beta}_{boot}} - \hat{\beta}^{(-D)} \quad (17)$$

اما التباين فيحسب كالآتي:

$$Cov(\hat{\beta}_{boot}) = \frac{1}{B-1} \sum_{b=1}^B \left(\hat{\beta}^{b(-D)} - \overline{\hat{\beta}_{boot}} \right) \left(\hat{\beta}^{b(-D)} - \overline{\hat{\beta}_{boot}} \right)' \quad (18)$$

إذا الـ MSE للمعالم المقدرة باستخدام MWBP هي:

$$MSE(\hat{\beta}_{boot}) = (\overline{bias_{boot}})^2 + Cov(\hat{\beta}_{boot}) \quad (19)$$

عليه يكون الجذر التربيعي للـ MSE كالآتي:

$$RMSE(\hat{\beta}_{boot}) = \sqrt{MSE(\hat{\beta}_{boot})} \quad (20)$$

بناء على ما تقدم ستكون قيمة RMSE معيار للمفاضلة بين طريقتي التقدير BOLS و MWBP على التوالي.

الفصل الثالث

الجانب التطبيقي

3-1 وصف البيانات

Stack loss Data

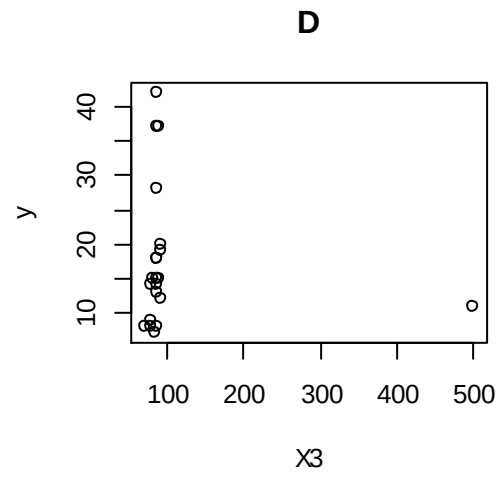
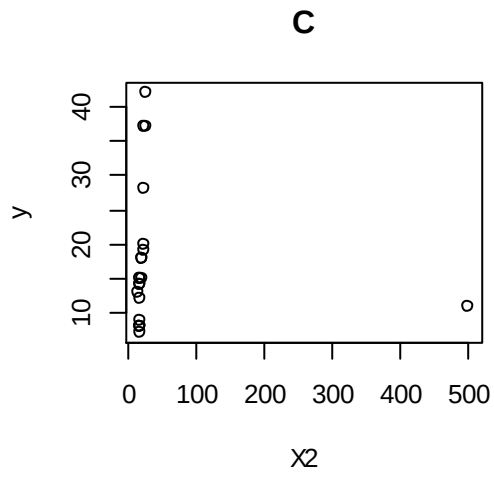
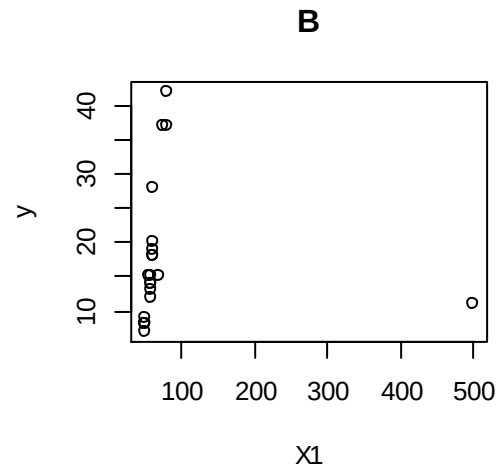
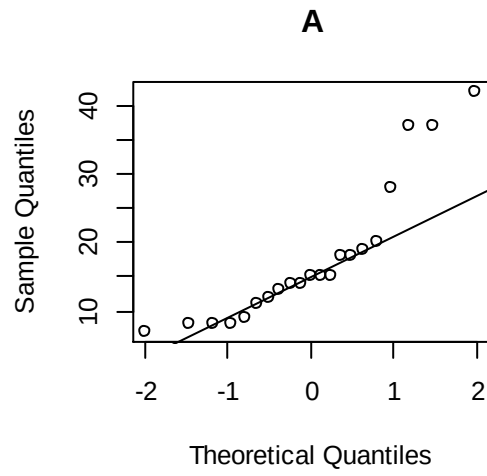
قدمت هذه البيانات من خلال برونلي 1965 و التي تصف العملية النباتية لأكسدة الامونيا و تحويلها الى حامض النتريك . تم تسجيل 21 مشاهدة لاربعة متغيرات ، و هي المتغير المعتمد y (الخسارة في التكديس) و ثلاثة متغيرات توضيحية (معدل عملية التحويل x_1) و (درجة حرارة مدخل المياه المبردة x_2) و تركيز الحامض x_3 . اثبتت العديد من الدراسات ان تحتوي 4 قيم شاذة عموديا الا ان ميدي و ايمون قامو بتعديل هذه البيانات لتتضمن نقطة قوة رافعة رديئة في كل متغير من المتغيرات التوضيحية باستبدال المشاهدات رقم 13 بالقيمة 500. استخدمنا البيانات المعدلة لاختبار اداء الطريقة المقترحة و مقارنة نتائجها بالطريقة السابقة.

3-2 تشخيص القيم الشاذة بالرسم

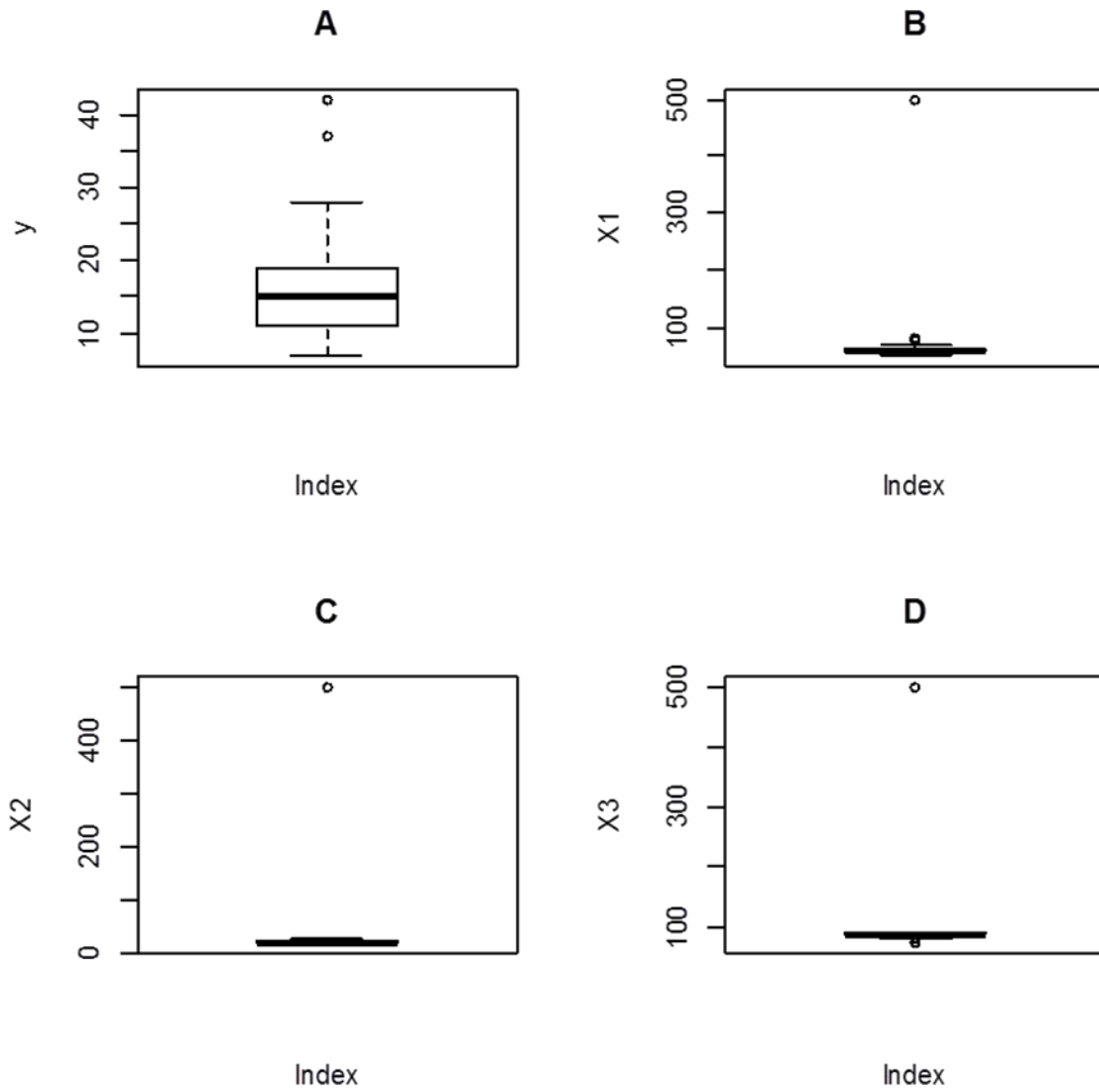
باستخدام لغة R تم اختبار فيما اذا كانت بواقي الانحدار تتوزع توزيعا طبيعيا بمتوسط قيمته صفر وتباين ثابت . الملاحظ من خلال الشكل (1-A) ان هناك بالتحديد 4 قيم شاذة قد ابتعدت عن خط التوزيع الطبيعي و على العكس من معظم البواقي المتبقية التي التفت على خط التوزيع الامر الذي يؤكد ان هذه البواقي تتوزع توزيعا مختلطا بسبب وجود القيم الشاذة بمعنى اخر ان القيم الشاذة تتبع توزيعا اخر يختلف عن توزيع معظم بواقي الانحدار.

اما الشكل (1-B) الذي يوضح العلاقة ب x_1, y يظهر جليا ان هناك نقطة قوة رفع رديئة عند القيمة 500 بالمحور السيني بالاضافه الى 4 قيم شاذة عند المحور الصادي y قد اثرت على اتجاه خط الانحدار وتكررت هذه الحالة في الشكلين (1-C) و (1-D) . و لتأكيد وجود القيم الشاذة ونقاط الرفع الرديئة قام الباحث باستخدام الرسم الصندوقي لكل متغير على حدة . من الواضح جدا في الشكل (2-A) ان هناك عدد من القيم الشاذة ظهرت خارج حدود تقسيمات الرسم الصندوقي ، اما الاشكال (2-B)، (2-C) و (2-D) التي بينت الرسم الصندوقي للمتغيرات x_1, x_2, x_3 على التوالي قد اظهرت بشكل لايقبل الشك ان هناك نقطة قوة رافعة بعيدة جدا قد اثرت على بقية القيم مما ادت الى ضغط الصندوق وجعله كانه خطأ واحدا .

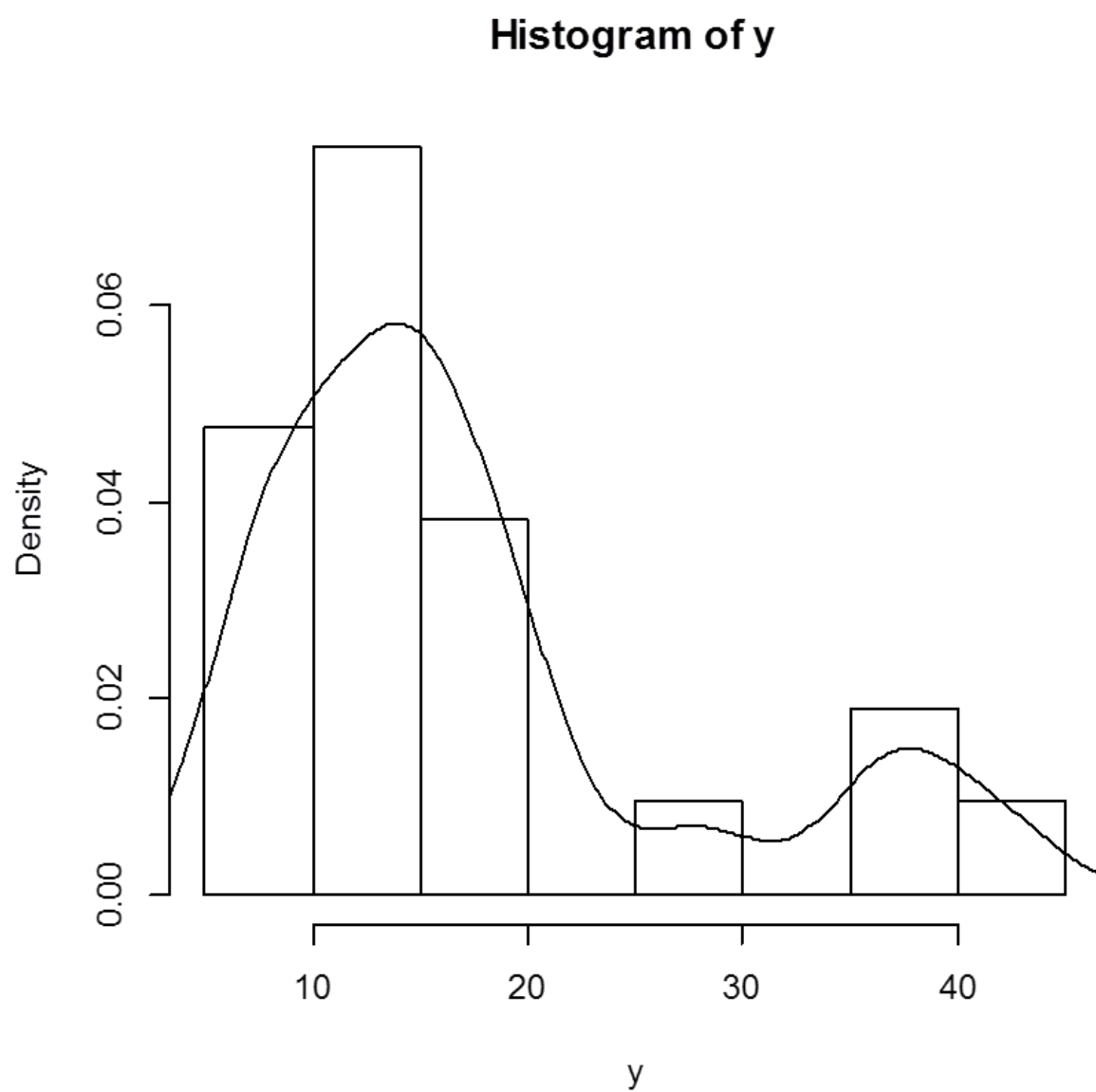
الشكل (3) الذي يتضمن رسم المدرج التكراري ومنحني التوزيع للمتغير المعتمد y فإنه يؤكد بما لا يقبل الشك ان التوزيع هو ليس طبيعيا لا من خلال المنحني او المدرج التكراري اللذان تضما ذيل طويلا الى الجانب الايمن للمنحني التوزيع



الشكل (1) مخطط توضيحي بأربعة اشكال فرعية توضح وجود القيم الشاذة ونقاط الرفع الرديئة في البيانات



الشكل (2) الرسم الصندوقي لتأثير القيم الشاذة ونقاط قوة الرفع الرديئة على البيانات .



الشكل (3) المدرج التكراري ومنحنى دالة الكتلة الاحتمالية لتوزيع المتغير العشوائي y .

3-3 النتائج

الجدول (1) ادناه يستعرض تقدير معلمات الانحدار باستخدام طريقتي MWBP و BOLS و القيم المطلقة لتحيز هذه القيم عن معلمات الانحدار بالإضافة الى الجذر التربيعي لمتوسط مربعات الخطأ لكل معلمة مقدرة من هذه المعالم. الملاحظ من النتائج تأثر طريقة BOLS حيث ان القيمة المطلقة لتحيز المعلمة المقدرة لثابت الانحدار $|bias(\hat{\beta}_{boot0})|$ باستخدام BOLS يساوي 21.61 بينما نظيرتها المحسوبة باستخدام MWBP فقط 0.62. اما النسبة للقيم المطلقة لعدم تحيز معلمات الانحدار المقدرة من خلال طريقة MWBP كانت اقرب الى معلمات الانحدار من تلك المقدرة باستخدام BOLS حيث نلاحظ في الجدول (1) ان الطريقة المقترحة حققت 0.005، 0.008 و 0.004 لكل من $|bias(\hat{\beta}_{boot1})|$ ، $|bias(\hat{\beta}_{boot2})|$ و $|bias(\hat{\beta}_{boot3})|$ على التوالي بالمقابل كانت قيم عدم التحيز المطلقة لطريقة BOLS هي 0.08 ، 0.02 و 0.11 على التوالي. من الواضح جدا ان المسافة بين المعلمات المقدرة من خلال الطريقة المقترحة هي الاقرب للمعلمات الحقيقية. هذا التقدير الدقيق و الكفوء للطريقة انعكس على قيم $RMSE(\hat{\beta}_{boot})$ لكل المعلمات المقدرة من حيث ان الطريقة المقترحة حققت اقل $RMSE(\hat{\beta}_{boot})$ من نظيرتها BOLS و بفارق يمكن ملاحظته ببساطة.

الجدول -1- المعلمات المقدرة و القيمة المطلقة للتحيز و الجذر التربيعي لمتوسط مربعات الخطأ لكل معلمة مقدرة للطريقة المقترحة MWBP و طريقة BOLS

	BOLS	MWBP
$\hat{\beta}_{boot0}$	-28.48	-7.49
$\hat{\beta}_{boot1}$	0.82	0.90
$\hat{\beta}_{boot2}$	0.27	0.30
$\hat{\beta}_{boot3}$	0.30	-0.42
$ bias(\hat{\beta}_{boot0}) $	21.61	0.62
$ bias(\hat{\beta}_{boot1}) $	0.08	0.005
$ bias(\hat{\beta}_{boot2}) $	0.02	0.008
$ bias(\hat{\beta}_{boot3}) $	0.11	0.004
$RMSE(\hat{\beta}_{boot0})$	254.22	5.08
$RMSE(\hat{\beta}_{boot1})$	2.59	0.18
$RMSE(\hat{\beta}_{boot2})$	3.62	0.30
$RMSE(\hat{\beta}_{boot3})$	3.62	0.08

الفصل الرابع

الاستنتاجات

4-1 الاستنتاجات و التوصيات

طريقة البوتستراب المقترحة من قبل ايفرون لتقدير معلمات الانحدار الخطي اعتمدت الى طريقة المربعات الصغرى التي هي حساسة جدا عند ظهور القيم الشاذة و نقاط القوة الرافعة في بيانات العينة. كما نلاحظ من النتائج ان الطريقة قد تاثرت بشدة لظهور مثل هذه القيم في بيانات الـ Stackloss. على العكس من ذلك كانت الطريقة المقترحة اكثر قوة لمقاومة هذه الشواذ و تقديم مقدرات قريبة جدا من معلمات توزيع بيانات الانحدار و هذا يؤشر ان الطريقة المقترحة تتميز بدقة التقدير و بأقل خطأ ممكن. لذلك يوصي الباحث الى اعتماد هذه الطريقة عند وجود نقاط قوة رفع رديئة و قيم شاذة في بواقي الانحدار.

المصادر والمراجع

1. Barrera, M. S. , M.& Zamar, R. H. (2002) Botstrapping Robust Estimates of Regression, *J. [Ann. Statist.](#)*, 30, pp.556-582.
2. Barrera, M. S. M. (2003) . Fast and stable bootstrap methods for robust estimates. *Computing Science and Statistics*, **34**, 346-359.
3. Barrera, M. S. (2006). Bootstrapping MM-estimators for linear regression with fixed designs, *Statistics & Probability Letters* 76, 1287–1297.
4. Barrera, M. S., M., Aelst, S. V. And **Willems, G.**(2008) Fast and Robust Bootstrap, *Stat. Meth. & Appl* ,17:41–71.
5. Barnett, V. & Lewis, T. (1978), "Outliers in Statistical Data", Jon Wiley and Sons, New York
6. Bross, I.D.J. (1961), "Outliers in Patternend Experiments: strategic Re-Appraisal". *Technometrics* . 3,19-102
7. Efron B.& Robert J. Tibshiriani., (1993) An Introduction to the Bootstrap,6th ed.,CRC press, USA.
8. Freeman, P.R.(1980), "On the Number of Outliers in Data from a linear Model" In *Bayesian Statistics*, Ed .J.M.
9. Hampel, F.R. 1971., A general Qualitative Definition of robustness, *The Annals Mathematical of Statistics.*, 42,1887-1896.

10. Imon A.H.M.R, & Ali.M.M, (2005) Bootstrapping regression residuals, *J.Korean Data Inform.Sci.Soc.*,16:665-682.
11. Keller, G. & Brian Warrack, (2000),"Statistic For Management and Economics, 5th edition, Duxbury, Thomson Learning, U.S.A
12. Midi B. Habshah, Uraibi S. Hassan & Tlib A. Bashar, (2009) Dynamic Robust Bootstrap Method Based LTS Estimator. *European Journal for Scientific Research.*, 32: 277-287.
13. Ramli, N. M. (2008) Weighted maximum median Likelihood Estimation for Parameters in multiple linear regression model, Ph.D thesis, University Putra Malaysia (UPM), Malaysia.
14. Rousseeuw P. J. (1984) Least median of squares regression, *J. Am. Stat. Assoc.*, 79, 871-880.
15. Rousseeuw P. J. & Van Driessen K.(1999) Computing LTS regression for Large Data Sets, *Data Mining Knowledge Discovery* 12, 29-45.
16. Tille, Y. (2006) Sampling Algorithms. Springer, Verlag New York, LLC.
17. Uraibi S. Hassan, Midi H. , Bashar A. Al-Talib, & Jabar H. Yousif (2009). Linear Regression Model Selection Based on Robust Bootstrapping Technique, *American Journal of Applied Sciences* 6 (6): 1191-1198.
18. Venables, W. N.& Ripley, B. D. (1999) *Modern Applied Statistics with S-PLUS*, Third Edition, New York : Springer.