بسم الله الرحمن الرحيم

(وَأَنزَلَ اللَّهُ عَلَيْكَ الْكِتَابَ وَالْحِكْمَةَ وَعَلَّمَكَ مَا لَمْ تَكُنْ تَعْلَمُ وَكَانَ فَضْلُ اللَّهِ عَلَيْكَ عَظِيمًا) (سورة النساء الاية 113)

شكر وتقدير

بعد الشكر والحمد لله العلى القدير أود أن أتوجه بالشكر والتقدير لأستاذى الفاضل ا د/أسامة عبد العزيزحسين لإشرافه على الرسالة وعلى ماقدمه لى من عون ونصائح قيمة ،فلولا توفيق الله ثم دعمه المستمر لى ما كان يتسنى لى أن أنتهى من الرسالة فجزاه الله عنى خير الجزاء .

وأود أن أعبر عن عميق شكرى وتقديرى لأستاذى الفاضل ا د/محمد على محمد أحمد والذى كان له فضل السبق فى توجيهى لموضوع الرسالة من البداية ،وكذلك تفضله بالإشتراك فى لجنة الحكم على الرسالة رغم مشاغله الكثيرة مما يفيد الباحث من مناقشاته العلمية المتميزة .

كما أتقدم بجزيل الشكر والتقدير والإمتنان للأستاذ د/ عبدالرؤوف عبد الرحمن لتكبده عناء قراءة الرسالة وتفضله بقبول الإشتراك فى لجنة الحكم رغم مشاغله الكثيرة فله منى كل الشكر والتقدير

كما يشكر الباحث ا.د. على هادى على تعاونه فى توفير البرامج الاحصائية التنفيذية التى تم إستخدامها فى التطبيق العملى للدراسة .

كما أود أن أشكر جميع أساتذتى الأجلاء فى قسم الإحصاء لما تعلمته منهم من خلق وعلم ،وأخص بالشكر ا د/ امتثال محمد حسن لدعمها المعنوى المستمر لى طوال فترة إعداد الرسالة ،كما أشكر أساتذتى الذين نهلت من علمهم وأخلاقهم الكثير ا د/عادل حلاوة ، و ا د/عبدالمرضى حامدعزام ، ولا يفوتنى أن أذكر الراحل د/ يحيى سعد زغلول رحمه الله وأسكنه فسيح جناته.

وأخيراً أقدم عميق شكرى إلى زملائى فى قسم الاحصاء على تشجيعهم وحسن تعاونهم فجزى الله الجميع عنى كل الخير.

إهداء

إلى روح أبى رحمه الله وإلى أمى رمز العطاء

براً وإحساناً

إلى زوجتى رمز التضحية

مودة ورحمة

إلى أبنائى يوسف وبلال وجنى

رعايةً وحباً

إلى أساتذتى رموز العلم والأخلاق

تقديراً وعرفاناً

لجنة المناقشة والحكم على الرسالة :

ا . د / عبد الرؤوف عبد الحمن عبد الواحد

أستاذ الاحصاء بكلية التجارة – جامعة طنطا

رئيساً

ا . د / محمد على محمد أحمد

أستاذ الاحصاء ووكيل الكلية لشئون التعليم والطلاب – كلية التجارة – جامعة الأسكندرية

عضواً

د / أسامة عبد العزيز حسين

أستاذ الاحصاء المساعد

كلية التجارة – جامعة الأسكندرية

مشرفاً وعضواً

دراسة مقارنة لبعض الطرق المتينة للكشف عن المشاهدات الشاذة فى نماذج الإنحدار الخطى

A Comparative Study of Some Robust Outliers Diagnostics in Linear Regression Models

لنيل درجة الماجستير فى الإحصاء التطبيقى

مقدمة من الطالب / أحمد صدقى محمد الديب
المعيد بقسم الإحصاء و الرياضة والتأمين

إشراف

د/ أسامة عبد العزيز حسين
أستاذ م بقسم الإحصاء والرياضة والتأمين
كلية التجارة - جامعة الأسكندرية

د/محمد توفيق قفشير
مدرس بقسم الإحصاء والرياضة والتأمين
كلية التجارة - جامعة الأسكندرية

قائمة محتويات الدراسة

دراسة مقارنة لبعض الطرق المتينة للكشف عن المشاهدات الشاذة فى نماذج الإنحدار الخطى

	رقم الصفحة
الفصل الاول : مقدمة	
١-١ تمهيد	١
١-٢ مشكلة الدراسة	٢
١-٣ هدف الدراسة	3
١- ٤ أهمية الدراسة	4
١- ٥ خطة الدراسة	5
الفصل الثانى : إكتشاف المشاهدات البعيدة عن مركز البيانات	
مقدمة	8
٢-١ هيكل وإفتراضات النموذج	8
2-٢ أنواع المشاهدات	10
٢-٣ طرق إكتشاف المشاهدات البعيدة عن مركز البيانات	12
٢-٣-١ المقاييس التقليدية لإكتشاف المشاهدات البعيدة عن مركز البيانات	12
٢-٣-١-١ العناصر القطرية للمصفوفة المقدرة (Diagonal Elements of Hat Matrix)	12
٢-٣-١-٢ مسافة ماهالانوبيس (Mahalanobis Distance)	14
٢-٣-٢ المفاهيم والمعايير المستخدمة في المقارنة بين أساليب الانحدار المتين	16
١- نقطة الإنكسار (Breakdown Point)	16
٢- مفهوم الكفاءة النسبية لمقدر الانحدار المتين (Relative Efficiency)	17
٣-مفهوم التأثير المحدود (Bounded Influence Concept)	17

4-مفهوم إمكانية التحويل (Equivariance)	19
٥-القدرة الحسابية(Computability)	20
٢-٤ المقدرات المتينة للموقع والمقياس(Robust Estimators for Location and Scale)	21
٢-4-١ مقدرات M للموقع والمقياس (M-estimators of Location and Scale)	21
٢-٤-٢ مقدرات الحجم الأدنى للقطع الناقص(Minimum Volume Ellipsoid,MVE) للموقع والمقياس	22
٢-٤-٣ مقدرات أدنى محدد للتغاير(Minimum Covariance Determinant,MCD) للموقع والمقياس	25
٢-٤-٤ المسافات المتينة بإستخدام الخوارزم السريع لطريقة أدنى محدد للتغاير(Fast-MCD Algorithm)	26
٢-٤-٥ مسافة باكون (BACON Distance , BD)	28
٢-٤-٦ المسافات المتينة 2006 (Robust Distance 2006 , RD06)	32
٢- ٥ مقارنة بين طرق إكتشاف المشاهدات البعيدة عن مركز البيانات	35
٢-٥-1 القابلية للحساب عملياً (Comulational Practicality)	35
٢-٥-2 إمكانية التحويل (Equivariance)	36
الفصل الثالث : إكتشاف المشاهدات الشاذة والمشاهدات المؤثرة	37
مقدمة	
3-1 المقاييس التقليدية لتحديد المشاهدات الشاذة	37
٣-٢ الصعوبات التي تواجه الطرق التقليدية لإكتشاف المشاهدات الشاذة	39
٣-٣ الطرق غير المباشرة لإكتشاف المشاهدات الشاذة	40
3-٣-1 طرق الإنحدارذات قيمة الإنكسار العالية	41
١) طريقة أدني قيمة مطلقة (Least Absolute Values,LAV)	41

٢) طريقة وسيط المربعات الأدنى (Least Median of Squares,LMS)	42
٣) طريقة المربعات المشذبة الصغري (Least Trimmed of Squares,LTS)	43
٣-٣-2 الطرق ذات الكفاءة العالية (High Efficiency Methods)	44
١) طريقة مقدرات M (M-estimators)	45
٢) طريقة تقدير M العامة (GM)	46
٣) طريقة تقدير MM (MM-Estimation)	47
٣-٤ الطرق المباشرة لإكتشاف المشاهدات الشاذة	51
٣-٤-1 الأعمال المبكرة لإكتشاف المشاهدات الشاذة المتعددة	51
٣-٤-2 طريقة (Hadi and Simonoff,HS93&HS97)	52
3-4-٣ طريقة (Chatterjee and Machler1997 , CM97)	53
3-4-4 طريقة المربعات الصغرى المعاد ترجيحها تكراريا ًا2001 (Iteratively Reweighted Least Squares Method ,IRWLS01)	55
٣-٤-5 طريقة المربعات الصغرى المعاد ترجيحها تكراريا ًا٢٠٠٦ (Iteratively Reweighted Least Squares Method ,IRWLS06)	57
٣-٥ إكتشاف المشاهدات المؤثرة (Detection of Influential Observations)	60
٣-٦ مقاييس إكتشاف المشاهدات المؤثرة	62
3-٦-1 مسافة كوك (Cook's Distance)	62
3-٦-2 مسافة كوك المعدلة (Modified Cook`s Distance)	62
3-٦-3 الفروق فى القيم الموفقة (DFFITS)	63
3-٦-4 التأثير على معامل إنحدار واحد (DFBETAS)	63
3-٦-5 نسبة التغاير (COVRATIO)	64
الفصل الرابع : الدراسة التطبيقية لإكتشاف المشاهدات البعيدة عن المركز	66
٤-١ مقارنة أداء طرق إكتشاف المشاهدات البعيدة عن مركز البيانات فى مجموعات البيانات	

١- Air data	69
٢- Aircraft data	70
٣- Buxton data	71
٤- Delivery data	72
٥- Education data	73
٦- Hawkins data	75
٧- Salinity data	76
٨- Stackloss data	77
٩- Stars data	78
١٠- Two Systems data	79
الفصل الخامس :الدراسة التطبيقية لإكتشاف المشاهدات الشاذة و المشاهدات المؤثرة	81
٥- ١ معايير تقويم أداء طرق إكتشاف المشاهدات الشاذة	82
٥-١-١ معدل الخطأ فى تحديد المشاهدات الشاذة (Outliers Error Rate , OER)	82
٥-١-٢ معدل الخطأ فى تحديد المشاهدات الجيدة (Inliers Error Rate , IER)	82
٥-٢ مقارنة أداء طرق الانحدار المتين فى إكتشاف المشاهدات الشاذة من حيث	
• القدرة على إكتشاف المشاهدات الشاذة (Detection Capability,DC)	
• معدل الخطأ فى تحديد المشاهدات الشاذة (OER)	
• معدل الخطأ فى تحديد المشاهدات الجيدة (IER)	
• معدل الإنذار الخطأ (False Alarm Rate,FR)	
• الأشكال البيانية التشخيصية (Diagnostic Plots)	
٥-٣ طرق التحليل ذو المرحلتين (البواقى - المسافة)	84
1- طريقة LMS - MVE 2- طريقة LTS – MCD	

3- طريقة IRWLS01 – BACON	4- طريقة IRWLS06 – RD06
١- Stackloss data	88
٢- Aircraft data	92
٣- Delivery data	96
٤- Salinity data	100
٥- Two Systems data	104
٦- Stars data	108
٧- Education data	112
٨- Hawkins data	116
٩- Buxton data	120
١٠- Air data	124
٤-٥ مقارنة أداء طرق إكتشاف المشاهدات المؤثرة قبل وبعد الترجيح:	126
• مسافة كوك (Cook's Distance)	126
• فروق القيم الموفقة (DFFITS)	126
• مسافة كوك المعاد ترجيحها (Reweighted Cook's Distance)	128
• فروق القيم الموفقة المعاد ترجيحها (Reweighted DFFITS)	130
▪ مقارنة أداء مقياس الفروق فى القيم الموفقة قبل وبعد الترجيح	132
▪ مقارنة أداء مسافة كوك قبل وبعد الترجيح	134
الفصل السادس : الإستنتاجات العامة والتوصيات	
٦-١ الإستنتاجات العامة	135
٦-٢ التوصيات	144
جدول النتائج الكلية	145
مراجع الدراسة	146
ملاحق الدراسة	
ملحق (١) الأشكال البيانية لمقاييس إكتشاف المشاهدات البعيدة عن مركز البيانات	155
ملحق (٢) الأشكال البيانية للطرق غير المباشرة لإكتشاف المشاهدات الشاذة	173
ملحق (٣) الأشكال البيانية للطرق المباشرة لإكتشاف المشاهدات الشاذة	190

قائمة جداول الدراسة

رقم الجدول	عنوان الجدول	رقم الصفحة
(1 – 4)	جدول نقاط القطع(Cutoff Points) المستخدمة فى طرق إكتشاف المشاهدات البعيدة عن مركز البيانات	67
(2 – 4)	جدول المقارنة بين طرق اكتشاف المشاهدات البعيدة عن مركز البيانات لمجموعات البيانات المستخدمة فى الدراسة	68
(1 – 5)	جدول نسب المشاهدات الشاذة الحقيقية ونسب المشاهدات العادية مرتبة وفقاً لعدد المشاهدات الكلى لمجموعات البيانات	86
(2 – 5)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Stackloss)	٨٧ ، ٨٨
(3 – 5)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (StackLoss) وفقاً لطرق التحليل ذو المرحلتين (البواقى - المسافة)	90
(4 – 5)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Aircraft)	91 ، 92
(5 – 5)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (Aircraft) وفقاً لطرق التحليل ذو المرحلتين (البواقى - المسافة)	94
(6 – 5)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Delivery)	95-97
(7 – 5)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (Delivery) وفقاً لطرق التحليل ذو المرحلتين (البواقى - المسافة)	98
(8 – 5)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Salinity)	99-101
(9 – 5)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (Salinity) وفقاً لطرق التحليل ذو المرحلتين (البواقى - المسافة)	102
(10 – 5)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Two Systems)	103-105

رقم الجدول	عنوان الجدول	رقم الصفحة
(5 – 11)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (Two systems) وفقاً لطرق التحليل ذو المرحلتين (البواقى-المسافة)	107
(5 – 12)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Stars)	107-109
(5 – 13)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (Stars) وفقاً لطرق التحليل ذو المرحلتين(البواقى - المسافة)	110
(5 – 14)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Education)	111-113
(5 – 15)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (Education) وفقاً لطرق التحليل ذو المرحلتين (البواقى - المسافة)	114
(5 – 16)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Hawkins)	115,116
(5 – 17)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (Hawkins) وفقاً لطرق التحليل ذو المرحلتين (البواقى - المسافة)	811
(5 – 18)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Buxton)	119-121
(5 – 19)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (Buxton) وفقاً لطرق التحليل ذو المرحلتين (البواقى - المسافة)	122
(5 – 20)	جدول معدلات الأخطاء فى تحديد المشاهدات الشاذة (OER) ومعدلات الأخطاء فى تحديد المشاهدات الجيدة (IER) فى مجموعة بيانات (Air)	123-125
(5 – 21)	تصنيف المشاهدات الموجودة فى مجموعة بيانات (Air Quality) وفقاً لطرق التحليل ذو المرحلتين (البواقى -المسافة)	612
(5 – 22)	جدول المشاهدات المؤثرة قبل وبعد الترجيح بالأوزان فى مجموعات البيانات	812
(6 – 1)	جدول ترتيب طرق إكتشاف المشاهدات الشاذة وفقاً لمعايير (OER , IER)	137,138
(6 – 2)	جدول النتائج الكلية لطرق إكتشاف المشاهدات الشاذة	141,142

قائمة الأشكال البيانية المستخدمة فى مقاييس إكتشاف المشاهدات البعيدة عن مركز البيانات (ملحق رقم ١)

رقم الشكل البيانى	عنوان الشكل البيانى	رقم الصفحة
شكل (٤-١)	Air quality:Index plot of leverage values h_{ii}	١٥٥
شكل (٤-٢)	Index plot of Mahalanobis Distance MD	١٥٥
شكل (٤-٣)	Index plot of Robust Distance RD (using MCD)	١٥٥
شكل (٤-٤)	Scatter plot of Mahalanobis Distance versus Robust Distance	١٥٥
شكل (٤-٥)	Index plot of BACON Distance BD	١٥٦
شكل (٤-٦)	Index plot of Robust Distance(RD06)	١٥٦
شكل (٤-٧)	Aircraft :Index plot of leverage values h_{ii}	١٥٦
شكل (٤-٨)	Index plot of Mahalanobis Distance MD	١٥٦
شكل (٤-٩)	Index plot of Robust Distance RD (using MCD)	١٥٧
شكل (٤-١٠)	Scatter plot of Mahalanobis Distance versus Robust Distance	١٥٧
شكل (٤-١١)	Index plot of BACON Distance BD	١٥٧
شكل (٤-١٢)	Index plot of Robust Distance(RD06)	١٥٧
شكل (٤-١٣)	D-D plot for Air quality	١٥٨
شكل (٤-١٤)	D-D plot for Aircraft	١٥٨
شكل (٤-١٥)	Buxton:Index plot of leverage values h_{ii}	١٥٨

رقم الشكل البيانى	عنوان الشكل البيانى	رقم الصفحا
شكل (٤-١٦)	Index plot of Mahalanobis Distance MD	١٥٨
شكل (٤-١٧)	Index plot of Robust Distance RD (using MCD)	١٥٩
شكل (٤-١٨)	Scatter plot of Mahalanobis Distance versus Robust Distance	١٥٩
شكل (٤-١٩)	Index plot of BACON Distance BD	١٥٩
شكل (٤-٢٠)	Index plot of Robust Distance(RD06)	١٥٩
شكل (٤-٢١)	Delivery :Index plot of leverage values h_{ii}	١٦٠
شكل (٤-٢٢)	Index plot of Mahalanobis Distance MD	١٦٠
شكل (٤-٢٣)	Index plot of Robust Distance RD (using MCD)	١٦٠
شكل (٤-٢٤)	Scatter plot of Mahalanobis Distance versus Robust Distance	١٦٠
شكل (٤-٢٥)	Index plot of BACON Distance BD	١٦١
شكل (٤-٢٦)	Index plot of Robust Distance(RD06)	١٦١
شكل (٤-٢٧)	D-D plot for Buxton	١٦١
شكل (٤-٢٨)	D-D plot for Delivery	١٦١
شكل (٤-٢٩)	Education: Index plot of leverage values h_{ii}	١٦٢
شكل (٤-٣٠)	Index plot of Mahalanobis Distance MD	١٦٢
شكل (٤-٣١)	Index plot of Robust Distance RD (using MCD)	١٦٢
شكل (٤-٣٢)	Scatter plot of Mahalanobis Distance versus Robust Distance	١٦٢

رقم الشكل البيانى	عنوان الشكل البيانى	رقم الصفحا
شكل (٤-٣٣)	Index plot of BACON Distance BD	١٦٣
شكل (٤-٣٤)	Index plot of Robust Distance(RD06)	١٦٣
شكل (٤-٣٥)	Hawkins:Index plot of leverage values h_{ii}	١٦٣
شكل (٤-٣٦)	Index plot of Mahalanobis Distance MD	١٦٣
شكل (٤-٣٧)	Index plot of Robust Distance RD (using MCD)	١٦٤
شكل (٤-٣٨)	Scatter plot of Mahalanobis Distance versus Robust Distance	١٦٤
شكل (٤-٣٩)	Index plot of BACON Distance BD	١٦٤
شكل (٤-٤٠)	Index plot of Robust Distance(RD06)	١٦٤
شكل (٤-٤١)	D-D plot for Education	١٦٥
شكل (٤-٤٢)	D-D plot for Hawkins	١٦٥
شكل (٤-٤٣)	Salinity: Index plot of leverage values h_{ii}	١٦٥
شكل (٤-٤٤)	Index plot of Mahalanobis Distance MD	١٦٥
شكل (٤-٤٥)	Index plot of Robust Distance RD (using MCD)	١٦٦
شكل (٤-٤٦)	Scatter plot of Mahalanobis Distance versus Robust Distance	١٦٦
شكل (٤-٤٧)	Index plot of BACON Distance BD	١٦٦
شكل (٤-٤٨)	Index plot of Robust Distance(RD06)	١٦٦
شكل (٤-٤٩)	Stackloss:Index plot of leverage values h_{ii}	١٦٧