تقدير معلمات الانحدار الخطي المتعدد باستخدام الطرائق الحصينة (دراسة مقارنة)

أ. م. جبران عبد الأمير خطار
قسم الإحصاء والمعلوماتية
كلية علوم الحاسوب والرياضيات
جامعة القادسية/القادسية/العراق
jubalaa98@yahoo.com

م. د. گورگيس شهيد محمد
قسم الرياضيات
كلية التربية
جامعة القادسية/القادسية/العراق
dunya9000@yahoo.com

محمد سالم إسماعيل
قسم الرياضيات
كلية علوم الحاسوب والرياضيات
جامعة القادسية/القادسية/العراق
mohammad.salim1991@yahoo.com

الخلاصة

تم في هذه البحث استخدام عدة طرائق حصينة – قليلة الحساسية بالقيم الشاذة – لتقدير مصفوفة المعلمات منها طريقة مقترحة من قبل الباحث، وقد تمت مقارنة هذه الطرائق باستخدام معيار متوسط مربع الخطأ العشوائي وذلك بالاعتماد على أسلوب المحاكاة للحصول على بيانات تحاكي الواقع العملي والتي على أساسها تمت المقارنة بين الطرائق، تم توليد نوعين من المتغيرات وهي: متغيرات طبيعية ومتغيرات طبيعية ملوثة من جهتين. تم التوصل من خلال المقارنة بين نتائج ثلاث طرائق حصينة إلى أنه في حالة البيانات غير الملوثة تتطابق نتائج الطرائق الحصينة الثلاث، أما في حالة البيانات الملوثة فإن الطريقة المقترحة – التي تم تعديل قيمة ثابت القطع لها – تعطي غالباً نتائج أفضل من طريقتي مقدر أصغر محدد مصفوفة تباين مشترك (MCD) و(MCD) معادة الوزن، وتمت المقارنة باستخدام معيار معدل متوسطات مربعات الأخطاء.

الكلمات المفتاحية: الانحدار الخطي المتعدد، القيم الشاذة، متوسط مربع الخطأ، الطرائق الحصينة، تلوث البيانات.

---

1. المقدمة:

يعد الانحدار الخطي من أقدم وأهم الأدوات الإحصائية في دراسة العلاقة بين المتغيرات واستُخدم بشكل واسع في بحوث علوم الحياة والبحوث البيئية وغيرها من العلوم التطبيقية، كما أن موضوع تحليل النماذج الخطية متعددة المتغيرات من المواضيع المهمة في الإحصاء الرياضي لأنه يصف العلاقة بين المتغيرات على هيئة معادلة خطية، كما إن له استعمالات كثيرة منها التنبؤ (Prediction)، حيث يتم به تقدير الاستجابة (Response) والتنبؤ بها الذي كثيراً ما يفيد في التخطيط (Planning) واتخاذ القرارات (Decision Making)، ويستخدم أيضاً لأغراض السيطرة (Control) على قيم المتغيرات المعتمدة (Dependent Variables) بتغيير قيم المتغيرات التوضيحية (Explanatory Variables).

وللحصول على تنبؤ جيد حول ظاهرة معينة يجب معرفة قيم المعلمات (Parameters) الموجودة في النموذج الخطي متعدد المتغيرات، ولصعوبة معرفة القيم الحقيقة لمعلمات النموذج تم اللجوء إلى تقدير هذه المعلمات، ومن أكثر وأقدم طرائق التقدير هي طريقة المربعات الصغرى الاعتيادية (Ordinary Least Squares Method) عندما تتوزع البيانات توزيعاً طبيعياً (Normal Distribution)، وبقيت هذه الطريقة إحدى أساليب تقدير معلمات الانحدار الخطي لمدة زمنية طويلة وقد تم ذكرها أيضاً في أغلب الكتب الإحصائية لما امتازت به مقدراتها من صفات جيدة، إلا أن الباحثين واجهوا مشكلة ابتعاد البيانات عن التوزيع الطبيعي عند احتواءها على القيم الشاذة (Outliers) والتي نتجت عن أخطاء في وصف البيانات أو تسجيل مشاهدات خاطئة قد يكون لها تأثير على مقدرات المربعات الصغرى حيث أنها قد تسحب إليها خط التوفيق للمربعات الصغرى.

في العقود الخمسة السابقة حظيت مشكلة القيم الشاذة في البيانات باهتمام كبير بسبب إدراك الكثير من الباحثين لخطورة استخدام الأساليب التقليدية في تقدير المعلمات عند ظهور هذه المشكلة، والذي أدى إلى البحث عن طرائق بديلة عن طريقة المربعات الصغرى الاعتيادية والتي تدعى بالطرائق الحصينة (Robust Methods) حيث أنها ذات مقدرات قليلة الحساسية والتأثر تجاه

القيم الشاذة عند وجودها، حيث أن مقدرات هذه الطرائق تمتلك كفاءة مساوية إلى طريقة المربعات الصغرى في حالة عدم وجود القيم الشاذة وذات كفاءة أعلى بكثير من المربعات الصغرى في حالة وجود القيم الشاذة، كما أنها تمتلك صفات جيدة سواء كان التوزيع طبيعياً أو غير طبيعي.

2. هدف البحث:

يهدف هذا البحث إلى معالجة مشكلة ابتعاد مصفوفة المعلمات المقدرة عن المعلمات الحقيقية عندما تبتعد مصفوفة الأخطاء العشوائية عن التوزيع الطبيعي متعدد المتغيرات لاحتوائها على قيم ملوثة، وذلك باعتماد عدة طرائق حصينة يمكن استعمالها لتقدير معلمات مجتمع متعدد المتغيرات هي: طريقة مقدر أصغر محدد تباين مشترك (MCD)، طريقة أصغر محدد تباين مشترك معاد الأوزان (RMCD) واستخدم فيه الوزن (0,1) وطريقة (RFMCD) التي تم استبدال الوزن فيها بدالة وزن مقترحة بالاعتماد على دالة (Huber) بعد تغيير قيمة ثابت القطع (C) المستخدم بقيمة ثابت قطع مقترحة تم بها تحقيق نتائج أفضل، وسيتم استعمال أسلوب المحاكاة (Simulation Procedure) لتوليد البيانات والحصول على نتائج تبين الفرق بين هذه الطرائق من حيث سلوكها ودرجة استجابتها لتأثير بعض العوامل منها وجود قيم شاذة في البيانات، لذلك سيتم افتراض التوزيع الطبيعي متعدد المتغيرات الملوث (Contamination) للتعبير عن حالة وجود قيم شاذة من جانبين وكذلك حجم العينة، إذ شملت الدراسة عينتين مختلفتي الحجم وستتم المقارنة بين تلك الطرائق من خلال معيار متوسط مربع الخطأ العشوائي (MSE).

3. نموذج الانحدار الخطي متعدد المتغيرات [1]:

يعرف النموذج الخطي متعدد المتغيرات كما يلي:

$$Y = XB + E \qquad (1)$$

حيث أن:

Y: تمثل مصفوفة قيم المشاهدات لمتغيرات الاستجابة وهي من الدرجة $n \times p$.

X: تمثل مصفوفة قيم المتغيرات التوضيحية (التفسيرية) وهي من الدرجة $n \times q$.

B: مصفوفة المعلمات غير المعلومة من الدرجة $q \times p$ والمطلوب تقديرها.

E: مصفوفة الأخطاء العشوائية في النموذج من الدرجة $n \times p$ وتتكون من n من المتجهات E_i الصفية المستقلة. وهذه المتجهات تخضع لعدة فرضيات منها:

1. أنها تتبع التوزيع الطبيعي متعدد المتغيرات بوسط حسابي $\underline{0}_{1\times p}$ ومصفوفة تباين – تباين مشترك $S_{p\times p}$ وتكون موجبة التعريف تماماً (positive definite) أي أن:

$$E = [\underline{E}_1, \underline{E}_2, \dots, \underline{E}_n]$$

$$E_i = [\underline{E}_{i1}, \underline{E}_{i2}, \dots, \underline{E}_{in}]$$

وعليه فإن

$$E_i \sim N_p\left(\underline{0}_{1\times p}, {\sigma^2}_{p\times p}\right) \qquad (2)$$

2. وكذلك

$$E(E_i E_j) = \begin{cases} \sigma, & if \quad i = j \\ 0, & other\ wise \end{cases} \qquad (3)$$

والتي تمثل تساوي التباينات لجميع الأخطاء العشوائية والاستقلالية بين الأخطاء العشوائية المختلفة لكل $i = j$ حيث أن $i, j = 1, 2, \dots, n$.

3. الاستقلالية بين المتغيرات التوضيحية (التفسيرية) والأخطاء العشوائية أي أن:

$$E(X_{ij} E_{ij}) = 0$$

فإن لم تتحقق فرضية واحدة على الأقل من هذه الفرضيات فإنّ هذا سيؤدي بالنتيجة إلى أن المقدرات تكون غير جيدة وهذا يسبب مشاكل كثيرة للباحث.

ولتقدير المعلمات غير المعلومة B للنموذج (1) فإن في كثير من الأحيان تم استعمال طريقة المربعات الصغرى

(Least Square Method)، إن ميزة هذه الطريقة هي أنها لا تعتمد على معرفة التوزيع الاحتمالي للأخطاء العشوائية وهي كثيرة الاستعمال وتسمى بطريقة المربعات الصغرى الاعتيادية (Ordinary Least Square) (O.L.S) إذا تحققت الفرضيات الخاصة بها. تعتمد عملية تقدير الطريقة على إيجاد مقدر B الذي يجعل مجموع مربعات الأخطاء أقل ما يمكن أي أن يكون المقدار $E'E$ أقل ما يمكن حيث أن:

$$E = (Y - X\hat{B}) \qquad (4)$$

وبما أن Y و X مصفوفتان بقيم ثابتة (معلومة، متولدة) لذا فإن مجموع مربعات الأخطاء العشوائية هو دالة بالمعلمات B فقط وإن تقدير هذه المعلمات $\hat{B}$ الذي يجعل مجموع مربعات الأخطاء العشوائية أقل ما يمكن يسمى بتقدير المربعات الصغرى، ولإيجاد هذا التقدير نوجد الاشتقاق الجزئي إلى المقدار $E'E$ بالنسبة إلى B ثم نجعل الناتج مساوياً للصفر وبالتالي ينتج p من المعادلات الطبيعية وتكتب بالصيغة الآتية:

$$X'X\hat{B} = X'Y \qquad (5)$$

وبحل هذه المعادلات آنياً نحصل على قيم $\hat{B}$ علماً أن قيم $\hat{B}$ التي تحقق المعادلة (5) يجب أن تحقق الشرط الثاني للتصغير وهو أن تكون المشتقة الثانية للمقدار $E'E$ بالنسبة للمعلمات B يجب أن تكون دائماً أكبر من الصفر، أي أن:

$$\left|\frac{\partial^2 E'E}{\partial B \partial B'}\right| > 0 \qquad (6)$$

4. متوسط مربع الخطأ (MSE) [2]:

متوسط مربع الخطأ العشوائي هو مقياس للدقة يتم حسابه بتربيع الخطأ لكل مشاهدة في مجموعة البيانات، ومن ثم إيجاد المعدل أو متوسط القيم لمجموع هذه المربعات، حيث أن الأخطاء يتم تربيعها قبل أخذ مجموعها ويمكن توضيح هذا المعيار كما يأتي:

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n - p - 1}$$

وباستخدام المصفوفات:

$$MSE = \frac{e'e}{n - p - 1}$$

$$e'e = y'y - \hat{B}x'y$$

$y'y$: يمثل حاصل ضرب مصفوفة قيم المتغير التابع ومدورها.

$\hat{B}$: يمثل مصفوفة معلمات الانحدار المقدرة.

$x'y$: يمثل حاصل ضرب مدور مصفوفة قيم المتغيرات المستقلة ومصفوفة قيم المتغير التابع.

n: حجم العينة، p: عدد المتغيرات الاستجابة.

5. القيم الشاذة (Outliers) [1]، [4]:

يعتبر مصطلح القيم الشاذة من المصطلحات التي لا يوجد إجماع على تعريفها بشكل محدد وقد أطلقت عليها عدة تسميات منها: القيم الشاردة أو المشاهدات غير المنطقية أو القيم المتطرفة أو الخوارج أو القيم الفعّالة أو الملوِّثات ... الخ ويكون موقعها في إحدى أو كلتا نهايتي منحني التوزيع الاحتمالي ويسمى التوزيع ذي القيم الشاذة من جهة واحدة بالتوزيع الملوّث من جهة الذي تتم فيه إزاحة معلمة الوسط، بينما يسمى التوزيع ذي القيم الشاذة من طرفين بالتوزيع الملوّث من جهتين الذي تتم فيه إزاحة التباين.

فقد عرف Grubbs [5] المشاهدات الشاذة بأنها تلك "التي تبدو منحرفة بشكل ملحوظ عن بقية مشاهدات العينة.

وقد أوضح Huber [5] تأثير القيم الشاذة على مقدرات المربعات الصغرى من خلال جملته المشهورة "إن وجود نقطة شاذة واحدة قدم يهدم المزايا الجيدة لمقدرات المربعات الصغرى، كما وأنها قد تسحب إليها خط التوفيق للمربعات الصغرى".

كما أوضح Rousseeuw [6] بالأمثلة تأثير القيم الشاذة على مقدرات المربعات الصغرى، وأوضح بالرسم كيف أن المشاهدة الشاذة الواحدة تغير من اتجاه خط المربعات الصغرى.

إن لوجود القيم الشاذة من بين البيانات أسباباً عديدة منها: أخطاء القياس، أخطاء التسجيل، أخطاء المعاينة وغيرها، وغالباً ما تنشأ من توزيعات متينة الذيل (Heavy-tailed Distributions)) أو توزيعات مختلطة (Mixture Distributions).

6. الحصانة (Robustness) [7]:

لقد وردت عدة تعاريف بشأن الحصانة ويعد Box [8] أول من أشار إليها حيث أن الطريقة الإحصائية تسمى حصينة (robust) إذا كان الاستدلال الإحصائي لا يتأثر بشكل ملحوظ نتيجة لاختراق أي من شروطها الأساسية، وقد ورد رأيان حول مفهوم الحصانة بالنسبة للاختبار الإحصائي وهما:

الأول [9]: يعد الاختبار الإحصائي حصيناً لاختراق أي شرط من الشروط اللازمة له إذا لم يحصل أي تغير ملحوظ في أي من احتمالي الخطأ من النوع الأول والخطأ من النوع الثاني.

الثاني [8]: ينطبق مفهوم الحصانة على مفهوم الخطأ من النوع الأول لأن الأمر يكون أكثر تعقيداً إذا تم توسيعه ليشمل الخطأ من النوع الثاني ويعود سبب ذلك إلى أن قوة الاختبار قد لا تتأثر كثيراً نتيجة الابتعاد عن بعض الشروط لذلك الاختبار، بينما يحدث عكس ذلك لقيمة الخطأ من النوع الأول. وبكلام آخر يمكن تعريف حصانة الاختبار الإحصائي في: أن الخطأ من النوع الأول للاختبار يكون غير حساس لاختراق أي من شروط ذلك الاختبار.

وعلى العموم فإن الحصانة هي إهمال القيم الشاذة أو تقليل تأثيرها على البيانات، وإن التقدير الحصين هو التقدير الذي يكون قليل التأثير تجاه القيم الشاذة وذا كفاءة تعادل كفاءة مقدرات المربعات الصغرى في حالة عدم وجود القيم الشاذة.

7. طرائق التقدير الحصينة:

(1-7) طريقة مقدر أصغر محدد تباين مشترك (MCD) [10]:

إن طريقة (MCD) مبينة في الخوارزمية الآتية:

1- في أول خطوة يتم تحديد مصفوفة Y التي تحتوي على n من الصفوف للمشاهدات و p من الأعمدة(المتغيرات).

2- يتم إيجاد متجه الأوساط الحسابية الاعتيادي μ_{usual} ومصفوفة التباين – التباين المشترك S_{usual} بالاعتماد على طريقة المربعات الصغرى الاعتيادية (الكلاسيكية) كقيم أولية، حيث أن:

$$\mu_{j_{usual}} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}$$

$$S_{j_{usual}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_{ij} - \mu_{j_{usual}}\right)'\left(Y_{ij} - \mu_{j_{usual}}\right)$$

حيث أن: $i = 1, \dots, n, j = 1, \dots, p$.

ومن الجدير بالذكر أننا استبدلنا متجه الأوساط الحسابية بمتجه الوسيطات في جميع المعادلات لأن الوسيط – القيمة الوسطى بين القيم بعد ترتيبها تصاعدياً أو تنازلياً – من المقاييس الحصينة قليلة الحساسية تجاه القيم الشاذة وأفضل من الوسط الحسابي وتم استخدامه في جميع الطرائق، ويكون متجه الوسيطات بالصيغة:

$$Median_{j_{usual}} = \begin{cases} \dfrac{Y_{ij_{(\frac{n}{2})}} + Y_{ij_{(\frac{n}{2}+1)}}}{2}, & n\ is\ even\ number \\ Y_{ij_{(\frac{n}{2}+1)}}, & n\ is\ odd\ number \end{cases}$$

3- يتم بعد ذلك حساب مسافات مهالونوبيس (Mahalanobis Distances) التربيعية حسب الصيغة الآتية:

$$d_i^2 = (Y_{ij} - \mu_{usual})S_{usual}^{-1}(Y_{ij} - \mu_{usual})'$$

4- بعدها يتم ترتيب المسافات التربيعية تصاعدياً وكالآتي:

$$(d_i^2)_{1:n} \le (d_i^2)_{2:n} \le \cdots \le (d_i^2)_{n:n}$$

5- يتم اختيار عدد ثابت h والذي يمثل عدد الصفوف المرتبة حسب المسافات التربيعية التي لها أقل محدد مصفوفة تباين مشترك ويحسب كالآتي:

$$h = int\left(\frac{n+p+1}{2}\right)$$

6- يتم تكرار الخطوات (1-5) r من المرات وذلك لضمان استقرار النموذج وفي كل مرة يتم حساب $\mu_{MCD,r}$ و $S_{MCD,r}$.

7- بعد ذلك يتم حساب تقدير مصفوفة المعلمات (B) ومتجه المعلمات (A) وكالأتي:

$$\hat{B}_{(q-1\times p)} = S_{XX}^{-1}{}_{(q-1\times q-1)} S_{XY_{(q-1\times p)}},$$

$$\hat{A}_{(1\times p)} = \mu_{Y_{(1\times p)}} - \mu_{X_{(1\times q-1)}}\hat{B}_{(q-1\times p)}$$

حيث أن:

$$S_{XX_{j_{(q\times q)}}} = \frac{1}{h-1}\sum_{i=1}^{n}(X_{ij} - \mu_X)'(X_{ij} - \mu_X)$$

$$S_{XY_{j\,(q\times p)}} = \frac{1}{h-1}\sum_{i=1}^{n}(X_{ij}-\mu_X)'(Y_{ij}-\mu_Y)$$

8- بعد ذلك يتم أخذ معدلات لـ r من المرات وبذلك يتم تقدير مصفوفتي المعلمات بطريقة MCD.

وقد تم إجراء بعض التعديلات على خوارزمية طريقة (MCD) لتكون بدقة عالية ووقت أقل للتنفيذ.

(2-7) طريقة حساب المقدر لأصغر محدد تباين مشترك موزون (RMCD):

يمثل ثابت القطع (C) درجة الحصانة [4] الذي له تأثير كبير جداً على حصانة التقديرات، فإن اختيار القيمة الجيدة لثابت القطع يؤدي إلى زيادة حصانة المقدرات وتحليل التباين، لذلك اهتم بعض الباحثين بتحديد قيم ثابت القطع، فمثلاً لدالة Huber، اقترحت [11] أفضل القيم وهي (1.5) أو (1.7) أو (2)، أو القيم (1.5) أو (1.345) التي تعطي كفاءة تقاربية بنسبة 95% تحت افتراض التوزيع الطبيعي.

نختار ثابت القطع على ما يناسب طريقة دالة الوزن بمقارنة قيمة ثابت القطع مع المسافة التربيعية للصف المعين بحيث تكون الطرائق الثلاث MCD وRMCD والطريقة المحورة متساويات تماماً في حالة البيانات غير الملوثة.

سيتم في هذا البحث الاعتماد على دالة وزن Huber التي تكون بالصيغة:

$$W(d_i) = \begin{cases} 1, & if\ d_i \le c \\ 0, & if\ d_i > c \end{cases}$$

حيث أن ثابت القطع (C) هو القيمة الجدولية لتوزيع مربع كاي بدرجة حرية p وبمستوى معنوية α بالنسبة لطريقة RMCD الحصينة، أي أن $C = \chi^2{}_{\alpha}(p)$، وستكون $\alpha = 0.7$ في كل البينات.

تتمثل طريقة (RMCD) في الخوارزمية الآتية:

1- في أول خطوة يتم تحديد مصفوفة Y التي تحتوي على n من الصفوف للمشاهدات و p من الأعمدة (المتغيرات).

2- يتم إيجاد متجه الأوساط الحسابية الموزون μ_{usual} ومصفوفة التباين – التباين المشترك الموزونة S_{usual} بالاعتماد على طريقة المربعات الصغرى الاعتيادية (الكلاسيكية) كقيم أولية، حيث أن:

$$\mu_{j_{usual}} = \frac{\sum_{i=1}^{n} W_i Y_{ij}}{\sum_{i=1}^{n} W_i}$$

$$S_{j_{usual}} = \frac{1}{\sum_{i=1}^{n} W_i - 1}\sum_{i=1}^{n} W_i\left(Y_{ij}-\mu_{j_{usual}}\right)'\left(Y_{ij}-\mu_{j_{usual}}\right)$$

حيث أن: $W_i = W(d_i)$.

3- يتم بعد ذلك حساب مسافات مهالونوبيس (Mahalanobis Distances) التربيعية حسب الصيغة الآتية:

$$d_i^2 = (Y_{ij}-\mu_{usual})S_{usual}^{-1}(Y_{ij}-\mu_{usual})'$$

4- بعدها يتم ترتيب المسافات التربيعية تصاعدياً وكالآتي:

$$(d_i^2)_{1:n} \le (d_i^2)_{2:n} \le \cdots \le (d_i^2)_{n:n}$$

5- يتم اختيار عدد ثابت h والذي يمثل عدد الصفوف المرتبة حسب المسافات التربيعية التي لها أقل محدد مصفوفة تباين مشترك ويحسب كالآتي:

$$h = int\left(\frac{n+p+1}{2}\right)$$

6- يتم تكرار الخطوات (1-5) r من المرات وذلك لضمان استقرار النموذج وفي كل مرة يتم حساب $\mu_{RMCD,r}$ و $S_{RMCD,r}$.

7- بعد ذلك يتم حساب تقدير مصفوفة المعلمات (B) ومتجه المعلمات (A) وكالأتي:

$$\hat{B}_{(q-1\times p)} = S_{XX}^{-1}{}_{(q-1\times q-1)} S_{XY\,(q-1\times p)},$$

$$\hat{A}_{(1\times p)} = \mu_{Y\,(1\times p)} - \mu_{X\,(1\times q-1)}\hat{B}_{(q-1\times p)}$$

حيث أن:

$$S_{XX_{j\,(q\times q)}} = \frac{1}{\sum_{i=1}^{n} W_i - 1}\sum_{i=1}^{n} W_i(X_{ij}-\mu_X)'(X_{ij}-\mu_X)$$

$$S_{XY_{j\,(q\times p)}} = \frac{1}{\sum_{i=1}^{n} W_i - 1}\sum_{i=1}^{n} W_i(X_{ij}-\mu_X)'(Y_{ij}-\mu_Y)$$

8- بعد ذلك يتم أخذ معدلات لـ r من المرات وبذلك يتم تقدير مصفوفتي المعلمات بطريقة RMCD.

(3-7) الطريقة المقترحة (RSMCD):

ذكرنا أن ثابت القطع (C) يمثل درجة الحصانة الذي له تأثير كبير جداً على حصانة التقديرات، فيتم اختيار القيمة الجيدة لثابت القطع بصورة تؤدي إلى زيادة حصانة المقدرات وتحليل التباين، لذا فقد تم اقتراح قيمة لثابت القطع (C^*) بتعديل دالة وزن Huber كالآتي:

$$W^*(d_i) = \begin{cases} 1, & if\ d_i \le C^* \\ \dfrac{C^*}{d_i}, & if\ d_i > C^* \end{cases}$$

حيث أن C^* هي قيمة ثابت القطع (cutoff value) المقترحة وقد تم تحديدها وفق تجارب قام بها الباحث بالاعتماد على نتائج التقدير وتستخدم في حالة التوزيع الطبيعي المتعدد الملوث من جهتين بالصيغة الآتية وبنفس خطوات طريقة RMCD:

$$C^* = 0.412 - 0.0028n + 0.009p + 0.033q$$

حيث أن p تمثل عدد المتغيرات التفسيرية، q تمثل عدد متغيرات الاستجابة.

وبعد ذلك تمت مقارنة النتائج بين الطرائق الثلاث بالاعتماد على معيار متوسط مربع الخطأ (MSE).

8. المحاكاة (Simulation) [12]:

يمكن تعريف المحاكاة بصورة عامة على أنها عبارة عن الحلول للمشكلات الرياضية من خلال بناء نموذج مشابه للمشكلة الأصلية ثم تطبيق المعاينة عليه، وقد استخدم هذا الأسلوب كثيراً في مجالات الإحصاء المختلفة لدراسة وتطوير الطرائق الإحصائية المختلفة فيما بينها. إن استخدام المحاكاة كأسلوب للتحليل يتم عند عدم التمكن من استخدام أساليب التحليل الأخرى وكذلك توفيراً للمال والجهد والوقت.

8. النتائج:

الحالة الأولى: عندما تكون المتغيرات غير ملوثة.

الحالة الثانية: عندما تكون البيانات ملوثة من كلا الجانبين.

وتتلخص النتائج التي تم الحصول عليها في الجداول الآتية:

جدول رقم (1) يوضح قيم MSE للتشكيلات المدروسة لمتغيرات طبيعية متجانسة وغير ملوثة

N			20			80		
q	p	Method/ρ	0.000002	0.5	0.9	0.000002	0.5	0.9
2	2	MCD	1.61E-06	1.76E-06	9E-07	2.15E-07	5.88E-07	2.63E-06
		RMCD	1.61E-06	1.76E-06	9E-07	2.15E-07	5.88E-07	2.63E-06
		RSMCD	1.61E-06	1.76E-06	9E-07	2.15E-07	5.88E-07	2.63E-06
	3	MCD	6.57E-06	7.35E-07	4.61E-06	1.3E-06	4.28E-06	6.53E-07
		RMCD	6.57E-06	7.35E-07	4.61E-06	1.3E-06	4.28E-06	6.53E-07
		RSMCD	6.57E-06	7.35E-07	4.61E-06	1.3E-06	4.28E-06	6.53E-07
3	2	MCD	1.25E-05	1.74E-05	2.26E-05	5.32E-06	1.46E-06	1.41E-06
		RMCD	1.25E-05	1.74E-05	2.26E-05	5.32E-06	1.46E-06	1.41E-06
		RSMCD	1.25E-05	1.74E-05	2.26E-05	5.32E-06	1.46E-06	1.41E-06
	3	MCD	7.23E-06	4.15E-06	2.85E-05	3.73E-07	3.18E-07	1.96E-06
		RMCD	7.23E-06	4.15E-06	2.85E-05	3.73E-07	3.18E-07	1.96E-06
		RSMCD	7.23E-06	4.15E-06	2.85E-05	3.73E-07	3.18E-07	1.96E-06

جدول رقم (2) يوضح قيم MSE للتشكيلات المدروسة لمتغيرات طبيعية غير متجانسة وغير ملوثة

N			20			80		
q	p	Method/ρ	0.000002	0.5	0.9	0.000002	0.5	0.9
2	2	MCD	2.16E-06	1.7E-06	2.32E-06	1.53E-07	1.65E-06	3.11E-06
		RMCD	2.16E-06	1.7E-06	2.32E-06	1.53E-07	1.65E-06	3.11E-06
		RSMCD	2.16E-06	1.7E-06	2.32E-06	1.53E-07	1.65E-06	3.11E-06
	3	MCD	4.41E-06	1.09E-05	4.16E-06	3.5E-07	3.47E-06	8.67E-08
		RMCD	4.41E-06	1.09E-05	4.16E-06	3.5E-07	3.47E-06	8.67E-08
		RSMCD	4.41E-06	1.09E-05	4.16E-06	3.5E-07	3.47E-06	8.67E-08
3	2	MCD	8.33E-06	1.27E-05	1.17E-05	6.48E-06	4.07E-06	9.25E-07
		RMCD	8.33E-06	1.27E-05	1.17E-05	6.48E-06	4.07E-06	9.25E-07
		RSMCD	8.33E-06	1.27E-05	1.17E-05	6.48E-06	4.07E-06	9.25E-07
	3	MCD	1.83E-05	1.12E-05	1.21E-05	3.64E-07	2.54E-06	7.19E-07
		RMCD	1.83E-05	1.12E-05	1.21E-05	3.64E-07	2.54E-06	7.19E-07
		RSMCD	1.83E-05	1.12E-05	1.21E-05	3.64E-07	2.54E-06	7.19E-07

جدول رقم (3) يوضح قيم MSE للتشكيلات المدروسة لمتغيرات طبيعية متجانسة وملوثة من جهتين

N			20			80		
q	p	Method/ρ	0.000002	0.5	0.9	0.000002	0.5	0.9
2	2	MCD	1.90E-03	1.69E-03	8.24E-04	1.68E-03	1.32E-03	5.69E-04
		RMCD	1.90E-03	1.69E-03	8.24E-04	2.03E-03	1.69E-03	5.69E-04
		RSMCD	1.78E-03	1.64E-03	7.89E-04	1.85E-03	1.50E-03	5.31E-04
	3	MCD	2.20E-03	1.92E-03	1.63E-03	1.24E-03	1.20E-03	1.18E-03
		RMCD	2.20E-03	1.92E-03	1.63E-03	1.22E-03	1.18E-03	1.12E-03
		RSMCD	1.67E-03	1.66E-03	1.33E-03	9.24E-04	8.92E-04	1.01E-03
3	2	MCD	4.68E-03	4.17E-03	3.46E-03	3.97E-03	3.93E-03	3.24E-03
		RMCD	4.68E-03	4.17E-03	3.46E-03	3.18E-03	3.93E-03	3.07E-03
		RSMCD	4.23E-03	4.00E-03	3.31E-03	3.51E-03	3.62E-03	2.97E-03
	3	MCD	4.19E-03	4.27E-03	3.83E-03	3.48E-03	3.76E-03	3.48E-03
		RMCD	4.19E-03	4.27E-03	3.83E-03	3.48E-03	3.65E-03	3.63E-03
		RSMCD	4.05E-03	3.90E-03	3.52E-03	2.91E-03	3.32E-03	3.10E-03

جدول رقم (4) يوضح قيم MSE للتشكيلات المدروسة لمتغيرات طبيعية غير متجانسة وملوثة من جهتين

N			20			80		
q	p	Method/ρ	0.000002	0.5	0.9	0.000002	0.5	0.9
2	2	MCD	1.50E-03	1.18E-03	4.80E-04	1.25E-03	8.57E-04	1.39E-04
		RMCD	1.50E-03	1.18E-03	4.80E-04	1.25E-03	8.57E-04	1.63E-04
		RSMCD	1.31E-03	1.03E-03	4.33E-04	1.27E-03	8.55E-04	1.02E-04
	3	MCD	2.16E-03	1.62E-03	1.56E-03	1.16E-03	1.08E-03	1.06E-03
		RMCD	2.16E-03	1.62E-03	1.56E-03	1.11E-03	1.04E-03	1.03E-03
		RSMCD	1.66E-03	1.29E-03	1.16E-03	7.99E-04	7.99E-04	9.79E-04
3	2	MCD	4.58E-03	3.79E-03	2.80E-03	3.52E-03	3.63E-03	3.04E-03
		RMCD	4.58E-03	3.79E-03	2.80E-03	3.52E-03	3.63E-03	3.04E-03
		RSMCD	3.99E-03	3.17E-03	2.63E-03	3.51E-03	3.62E-03	2.97E-03
	3	MCD	4.11E-03	3.97E-03	4.67E-03	3.44E-03	3.56E-03	3.53E-03
		RMCD	4.11E-03	3.97E-03	4.67E-03	3.52E-03	3.45E-03	3.51E-03
		RSMCD	3.23E-03	3.76E-03	4.03E-03	3.15E-03	3.06E-03	3.23E-03

9. تحليل النتائج:

الحالة الأولى:

نلاحظ من خلال الجدول (1) ما يأتي:

عندما تكون p=2 وq=2 فإن قيمة MSE تكون أفضل كلما زاد حجم العينة وقل معامل الارتباط في أغلب الحالات.

عندما تكون p=3 وq=2 فإن قيمة MSE تكون أفضل كلما زاد حجم العينة ومعامل الارتباط في أغلب التشكيلات.

عندما تكون p=2,3 وq=3 فإن قيمة MSE تكون أفضل كلما زاد حجم العينة ومعامل الارتباط في أغلب التشكيلات في هذه الحالة.

كما نلاحظ من خلال الجدول (2) ما يأتي:

عندما تكون p=2 وq=2 فإن قيمة MSE تكون أفضل كلما زاد حجم العينة ومعامل الارتباط في أغلب التشكيلات في هذه الحالة.

عندما تكون p=3 وq=2 فإن قيمة MSE تكون أفضل كلما زاد حجم العينة في أغلب الحالات المدروسة.

عندما تكون p=2 وq=3 فإن قيمة MSE تكون أفضل كلما زاد حجم العينة في أغلب التشكيلات في هذه الحالة.

عندما تكون p=3 وq=3 فإن قيمة MSE تكون أفضل عندما يقل معامل الارتباط ويزداد حجم العينة لأغلب التشكيلات في هذه الحالة.

كما نلاحظ من خلال الجداول (1) و(2) أن متوسط مربع الخطأ متساوٍ لجميع الطرائق الحصينة وبذلك تكون نتائجها متطابقة تماماً.

الحالة الثانية:

نلاحظ من خلال جدول رقم (3) ما يأتي:

عندما تكون p=2 وq=2 فإن قيمة MSE تكون أفضل كلما زاد حجم العينة ومعامل الارتباط، ونلاحظ تقدّم الطريقة المقترحة على الطرائق الأخرى ومن ثم طريقة MCD ومن ثم طريقة RMCD لأغلب التشكيلات في هذه الحالة.

عندما تكون p=3 وq=2 فإن قيمة MSE تتأثر بتغير كل من حجم العينة ومعامل الارتباط ولكن بصورة متذبذبة ونلاحظ تقدم الطريقة المقترحة على RMCD و MCD لأغلب التشكيلات في هذه الحالة.

عندما تكون p=2 وq=3 فإن قيمة MSE تكون أفضل كلما زاد حجم العينة ومعامل الارتباط لأغلب التشكيلات، ونلاحظ أن الطريقة المقترحة هي أفضل طريقة في أغلب التشكيلات في هذه الحالة وبعدها طريقة RMCD و MCD في هذه الحالة.

عندما تكون p=3 وq=3 فإن قيمة MSE تكون أفضل كلما زاد حجم العينة ولا تتأثر قيمته بتغير معامل الارتباط كثيراً، ونلاحظ أن الطريقة المقترحة وطريقة MCD هما الأفضل في كل التشكيلات في هذه الحالة.

نلاحظ من خلال الجدول (4) ما يأتي:

عندما تكون p=2 وq=2 فإن قيمة MSE تكون أفضل بالنسبة للطريقة المقترحة لأغلب التشكيلات ومن ثم طريقة MCD ومن ثم طريقة RMCD، وإن قيمته تكون أفضل كلما زاد حجم العينة ومعامل الارتباط لكل التشكيلات في هذه الحالة.

عندما تكون p=3 وq=2 فإن قيمة MSE تكون أفضل بالنسبة للطريقة المقترحة أيضاً ومن ثم طريقة MCD وRMCD، وكذلك فإن قيمة MSE تكون أفضل عند زيادة معامل الارتباط وحجم العينة لأغلب التشكيلات في هذه الحالة.

عندما تكون p=2 وq=3 فإن قيمة MSE تكون أفضل بالنسبة للطريقة المقترحة أيضاً ومن ثم طريقة MCD وطريقة RMCD، كما نلاحظ أن قيمة MSE تكون أفضل كلما زاد معامل الارتباط وحجم العينة لأغلب التشكيلات في هذه الحالة.

عندما تكون p=3 وq=3 فإن قيمة MSE تكون أفضل بالنسبة للطريقة المقترحة ومن ثم لطريقتي MCD وRMCD، نلاحظ أن

قيمة MSE تكون أفضل كلما ازداد حجم العينة ولا تتأثر كثيراً بتغير قيم معامل الارتباط لأغلب التشكيلات في هذه الحالة.

10. الاستنتاجات:

1- تمت ملاحظة أن المقدرات في حالة وجود الملوثات أن طريقة (RMCD) تلغي المشاهدات الشاذة من البيانات تماماً حيث يتم تعويض 0 بدلها بالاعتماد على قيمة القطع، بينما في الطريقة المقترحة فإنها تقرّب النقاط الشاذة إلى الخط الطبيعي ولا تلغيها تماماً.

2- كما لوحظ أن أفضل طريقة هي الطريقة المقترحة في أغلب الحالات بقيمة القطع الخاصة بها في كل أنواع المتغيرات.

11. التوصيات:

من خلال الاستنتاجات والملاحظات السابقة لابد أن نشير لبعض الأمور كتوصيات يمكن إن تؤخذ بنظر الاعتبار من قبل الباحثين في مجال التقديرات الحصينة مستقبلا، ومنها:

1- محاولة تطبيق هذا البحث على توزيعات غير طبيعية متعدد المتغيرات مثل التوزيع الآسي وغيرها من التوزيعات.

2- محاولة استبدال أو تعديل ثابت القطع المستخدم واستعماله على نفس دالة الوزن المقترحة في هذا البحث.

المصادر:

[1] السلامي، گورگيس شهيد، "الاستدلال الإحصائي الحصين لمعلمات الانحدار الخطي متعدد المتغيرات" أطروحة دكتوراه، كلية العلوم، الجامعة المستنصرية، 2006.

[2] العبادي، محمود محمد طاهر، "استخدام معلومات المجتمع في تقدير معالم أنموذج الانحدار المتعدد بالاعتماد على الانحدار التقسيمي مع التطبيق" المجلة العراقية للعلوم الإحصائية، العدد – 19، 2011.

[3] David A. Freedman, " Statistical Models Theory and Practice", Cambridge, New York, 2009.

[4] الدباغ، ظافر عاصم، "تحليل تباين حصين للنماذج الخطية" أطروحة دكتوراه، كلية الإدارة والاقتصاد، جامعة بغداد، 1999.

[5] Huber, P. J., "Robust Regression", Annals of Statistical, 1, pp. 799-821, 1973.

[6] Rousseeuw, P. J. & Leroy, A. M., "Robust Regression and Outlier Detection", John Wiley & Sons, New York, 1987.

[7] علوان، إقبال محمود، "مقارنة لبعض أساليب المقارنات المتعددة الحصينة" رسالة ماجستير، كلية الإدارة والاقتصاد، جامعة بغداد، 2002.

[8] Scheffe, H., "The Analysis of Variance", John Wiley & Sons, New York, 1959.

[9] Bradly, J. V., "Distribution - free Statistical Tests", Prentice-Hall International, Inc. London, 1986.

[10] Rousseeuw, P.J., & Katrin V. D., "A fast Algorithm for the Minimum Covariance Determinant Estimator", Technometrics, 41, pp. 212-223, 1999.

[11] ججو، نضال قرياقوش، "مخمنات حصينة لنموذج الانحدار الخطي" رسالة ماجستير، كلية التربية – ابن الهيثم، جامعة بغداد، 1989.

[12] رشيد، هدى عبد الله، "مقارنة الاختبارات الحصينة لعدم التجانس في البيانات المزدوجة" رسالة ماجستير في الإحصاء، كلية الإدارة والاقتصاد، الجامعة المستنصرية، 1997.

Estimation of Multivariate Linear Regression Parameters Using Robust Methods (Comparison Study)

Jubran A. Khuttar	Gorgeas Shaheed Mohammad	Mohammad Salim Ismail
Statistics and Informatics Dep.	Mathematics Dep.	Mathematics Dep.
Computer Science and Mathematics	College of Education	Computer Science and Mathematics
	University of Al-Qadisiya, Qadisiya, Iraq	
jubalaa98@yahoo.com	dunya9000@yahoo.com	mohammad.salim1991@yahoo.com

Abstract

In this paper In this Thesis has been use several robust methods – less sensitivity with outliers values – to estimate parameters matrix. These were compared by using criterion mean square random error by relying on simulation to get the data. Data mimic the reality and on which the comparison of methods is built. It was generated two kinds of variables: non-contaminated natural variables, contaminated natural variables from two sides. It was reached through the comparison between the results of the three methods. That data in the case of non-polluting match the results of the three robust methods. As in the case of polluting data, the suggested method - which has been assumed a constant cutoff value suggested for it - often give best results of two way methods: Estimate of minimum covariance determinate (MCD) & (MCD) re-Weight.

Keywords: Multivariate Linear Regression, Outliers, Mean Square Error, Robust Method, Data Contamination.