Iraqi Journal of Statistical Sciences, Vol. 21, No. 1, 2024, Pp (32-45)

Iraqi Journal Of Statistical Sciences

المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

طريقة المربعات الصغرى الموزونة الحصينة باستخدام صيغ مختلفة لمقدرات M (RWLSM) دراسة مقارنة

فاطمة محمد احمد و بشار عبدالعزيز الطالب

قسم الاحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

معلومات النشر

تاريخ المقالة:
تم استلامه في27 اذار 2023
تم القبول في 7 ايار 2023
متاح على الانترنيت 1 حزيران 2024

الكلمات الدالة:
المربعات الصغرى الموزونة، مقدرات M، اختيار المتغيرات، الكشف عن الشواذ، الانحدار الحصين.

المراسلة:
فاطمة محمد احمد
fatemamohammd738@gmail.com

الخلاصة

تم في هذا البحث تقليل أو إستبعاد تأثير عدم تحقق فرض التوزيع الطبيعي للبيانات، بسبب وجود أنواع من القيم الشاذة فيها عند الرغبة في إختيار أفضل معادلة إنحدار بالطرق الحصينة، وتمَّ تحقيق ذلك من خلال إدخال أوزان من طرق حصينة في التقدير واختبار حصانتها وملاءمتها للنموذج مسبقاً، ومن ثم إختيار الأوزان الناتجة من أعلى الطرق الحصينة كفاءة وإدخال هذه الأوزان في مراحل طرق أختيار أفضل معادلة إنحدار، فينتج عن ذلك نموذج يحقق صفتين في آن واحد وهما الحصانة وتقليل الأبعاد مقابل زيادة الكفاءة.

وقد تمَّ استخدام أسلوب المحاكاة على نماذج بأبعاد مختلفة وأحجام عينات مختلفة ونسب تلويث مختلفة في المتغير المعتمد مرة، وفي المتغيرات المستقلة مرة أخرى وفي الاثنين معاً مرة ثالثة، مع التركيز على دراسة احتمال تأثير وجود القيم الشاذة على المتغيرات التي ستبقى في الأنموذج والمتغيرات التي سيتمُّ حذفها.

ولتحقيق فكرة البحث تمّتْ مقارنة عدد من طرق التقدير الحصينة ومقارنة النتائج مع طريقة المربعات الصغرى الإعتيادية (OLS) وطريقة LASSO الحصينة المكيفة على بيانات تجريبية بأستخدام المحاكاة وكذلك على بيانات لعينة من مرضى الثلاسيميا في محافظة نينوى.

DOI 10.33899/IQJOSS.2024.0183230

1– المقدمة Introduction

إنَّ تقليل عدد المتغيرات التوضيحية التي تُستَخدم في نماذج الإنحدار يَعمَل على التقليل من الجهد والوقت والثمن المستغرق في تقدير وتحليل وتفسير الأنموذج، ويركز على المتغيرات ذات التأثير المعنوي الحقيقي على متغير الاستجابة، وكذلك يضمن لنا سهولة التحليل والفهم لطبيعة العلاقة بين المتغير المعتمد وأهم المتغيرات المستقلة تأثيراً على المتغير التابع إذ إنَّ أية زيادة في عدد المتغيرات التوضيحية غير المهمة في الأنموذج ستؤدي إلى ضياع الجهد دون جدوى، وعليه يجب أنْ يكون هناك توازن بين عملية تقليل عدد المتغيرات

التوضيحية وبين زيادة عددها للحصول على نتائج تنبؤية دقيقة، ويكون من الأفضل اختيار المعادلة المُفسِرة بأقل عدد من المتغيرات التوضيحية بحيث تكون هذه المتغيرات مُهِمة ولها تأثير معنوي فعلي على المتغير التابع (Al–Subaihi, 2004).

وبرزت فكرة البحث التي تأخذ بنظر الاعتبار حصانة وكفاءة المقدرات ولهذا ستنصب آليات التحليل على تحصين طرق أختيار أفضل معادلة إنحدار بالطرق التقليدية ضد القيم الشاذة وذلك من خلال تطبيق مجموعة من دوال الوزن وطرق التقدير ومقدرات التباين والمقدرات الأبتدائية البديلة ومن ثم أنتخاب أعلى المقدرات الموزونة كفاءة وتوظيف الاوزان المستحصلة منها في مراحل أختيار أفضل معادلة انحدار.

2– التقدير الحصين Robust Estimation

هي طرائق التقدير التي تعمل بشكل جيّد ليس فقط تحت ظروف مثالية، ولكن أيضاً في ظلّ ظروف تمثِّل خروجاً عن التوزيع أو الأُنموذج المفترض.

إنّ الهدف الأساسي من الإحصاء الحصين هو تطوير الإجراءات التي تبقى موثوقة وفعالة بشكل معقول في ظلّ الإنحرافات الصغيرة عن الأُنموذج. أي عندما يقع التوزيع الأساسي بالقرب من الأنموذج المفترض.

الطرائق الإحصائية الحصينة هي امتداد للطرائق المعلمية ، مع الأخذ بنظر الإعتبار أنَّ النماذج المعلمية هي أفضل تقريب للواقع ولكنّها تفقد كفاءتها عند عدم تحقق الفروض التي تقوم عليها(Hurn and Mirosevich, 2008).

3– بعض طرائق التقدير الحصين Robust Estimation Methods

إنَّ إتباع الطرائق التقليدية لتقدير معالم الأنموذج تكون غير دقيقة في تحليل البيانات عند وجود القيم الشاذة أو وجود خلل في إحدى فرضيات الإنحدار أو أنَّ توزيع الخطأ يكون غير طبيعي، حيث إنَّ وجود قيمة شاذه واحدة في البيانات سوف يؤدي إلى خلل في خصائص مقدرات المربعات الصغرى وإنَّ المقدر الحصين هو الذي يحافظ على الخصائص المرغوب بها للمقدرات عند خرق بعض فروض الإنحدار وسنلجأ إلى بعض طرائق التقدير الحصينة التي تمَّ تطبيقها في هذه الرسالة:

1- مقدرات M M-estimators

تعدُّ طريقة التقدير M واحدة من أهم الطرائق الحصينة شائعة الإستخدام، إذ أشارت أغلب الدراسات إلى إنَّ هذه الطريقة تعدُّ من أكثر الطرائق الحصينة سواء في كفاءتها المكافئة لطريقة المربعات الصغرى عندما تتوزَّع الأخطاء توزيعاً طبيعياً بوسط (صفر) وتباين σ^2 ،وتكون كفاءتها أعلى من كفاءة المربعات الصغرى عندما لاتتوزع الأخطاء توزيعاً طبيعياً أو عند وجود قيم شاذة في البيانات. وقد وسع (Huber،1973) نتائجه للتقدير الحصين من معلمة الموقع إلى حالة الإنحدار الخطي. وقد اكتسبت هذه التقديرات شهرة أكثر من بقية المقدرات الحصينة الأُخرى لأنَّها أكثر مرونة، وكذلك توفر إمكانية تعميمها مباشرة إلى الإنحدار المتعدّد.

وتبدأ عملية التقدير وفق هذه الطريقة (M–estimation) إبتداءاً بإستخدام طريقة المربعات الصغرى الموزونة تكرارياً (Iteratively Re– weighted Least Squares IRWLS) مع إستخدام إحدى دوال وزن مقدرات M، وهي كثيرة، ومن أشهرها هي دوال (Huber, Hampel, Tukey's Bisquare)، علماً أنَّ دالة Huber تضمن الحصول على مقدرات وحيدة (حل وحيد Unique Solution)، أما الطريقتان الأخريتان (Hampel و Tukey's Bisquare) فينتج عنهما عدة نهايات صغرى (تعدد الحلول المثلى)، وعندها يصبح من الضروري إستخدام قيم إبتدائية جيّدة بالشكل الذي نضمن فيه حصول تقارب سريع في المقدرات.

ومن الناحية الرياضية تقوم فكرة طريقة التقدير(M) على تحوير الأُسلوب المتبع في طريقة المربعات الصغرى اِلتي تهدف إلى تصغير المقدار الآتي :

$$Min \sum_{i=1}^{n} e_i^2 \qquad (1)$$

أي أنَّ

$$Min \sum_{i=1}^{n} \left(\underline{y}_i - \sum_{j=1}^{p} \underline{X_{ji}\hat{\beta}_j} \right)^2$$

أما طريقة مقدرات (M) فتهدف إلى تصغير المقدار

$$Min \sum_{i=1}^{n} \rho(e_i)$$

$$Min \sum_{i=1}^{n} \rho\left(\underline{y}_i - \sum_{j=1}^{p} \underline{x_{ji}\hat{\beta}_j}\right) \quad (2)$$

إذ تمثِّل ρ دالة بدلالة الأخطاء، ولتصغير المعادلة (1) نشتقها جزئياً بالنسبة لمتجه المقدرات $\hat{\beta}$ ومساواتها بالصفر، وكما يلي:

$$\sum_{i=1}^{n} x_i \ \Psi\left(\underline{y} - \underline{X\hat{\beta}}\right) \quad (3)$$

إذ تمثِّل Ψ المشتقة الجزئية للدالة ρ بالنسبة للمعلمات في المعادلة (2)، وتمثِّل منظومة مكونة من P من المعادلات، وتحلُّ باستخدام إحدى الطرق العددية المعروفة أو طريقة المربعات الصغرى الموزونة (Weighted Least Squares Method) ، ولإيجاد مقدرات M يتمُّ استخدام الصيغة الآتية:

$$\underline{\hat{\beta}_M} = \left(\underline{X}'\underline{W}\,\underline{X}\right)^{-1}\underline{X}'\underline{W}\,\underline{Y} \quad (4)$$

إذ تمثِّل W مصفوفة الأوزان، وهي مصفوفة قطرية (n×n) عناصر قطرها الرئيسية معطاة بالصيغة الآتية:

$$W_i = \frac{\Psi(e_i)}{(e_i)} \quad (5)$$

$$W_i = \frac{\left[\Psi\left(\underline{y}_i - \sum_{j=1}^{p} \underline{x_{ji}\hat{\beta}_j}\right)\right]}{\left(\underline{y}_i - \sum_{j=1}^{p} \underline{x_{ji}\hat{\beta}_j}\right)} \quad (6)$$

إذ تمثِّل β^0 القيم الإبتدائية لمتجه معلمات الأنموذج، ويتمُّ استخدامها لتحديد الأوزان، ويمكن استخدام مقدرات المربعات الصغرى كقيم ابتدائية وقد تم في هذه البحث أستخدام مقدرات المربعات الصغرى الاعتيادية OLS ومقدرات المربعات الصغرى المبتورة أو المشذبة LTS، ومن التكرار الأول نجد قيمة $\hat{\beta}^1$ ، أما في التكرار الثاني فنستخدم $\hat{\beta}^1$ في إيجاد الأوزان التي ستستخدم لإيجاد $\hat{\beta}^2$وهكذا تستمر عملية التكرار حتى نحصل على مقياس التقارب(Convergence) المعرف بالصيغة الآتية :

$$Max\left[\left|\hat{\beta}_j^{(r)} - \hat{\beta}_j^{(r-1)}\right|\right] < \delta \quad (7)$$

إذ تمثِّل δ قيمة صغيرة جدا،r تمثِّل عدد مرات التكرار، أي أنَّ الحلَّ يتوقف عندما يصبح الفرق المطلق بين المعلمات المقدرة في المرحلة الحالية والمعلمات المقدرة في المرحلة السابقة أصغر من القيمة المختارة δ أو يساويها، ولجعل مقدرات(M) تمتلك خاصية ثبات التباين Invariant Scale، فإنَّ الدالة المطلوب تصغيرها هي:

$$Min \sum_{i=1}^{n} \rho\left(\underline{y}_i - \sum_{j=1}^{p} \underline{X_{ji}\hat{\beta}_j}\right)/\hat{\sigma} \quad (8)$$

حيث نقوم بإشتقاق الدالة (8) بالنسبة للمتجه $\hat{\beta}_j$ ومساواتها بالصفـر فنحصل على دالة الوزن، وكما يلي:

$$\sum_{i=1}^{n} X_{ij}\Psi\left(y_i - x_i\beta_j\right)/\hat{\sigma} = 0 \quad (9)$$

ويمكن حلُّ المعادلة أعلاه باستخدام الصيغة (2) اذ إنَّ الأوزان يتمُّ إيجادها وفق الصيغة الآتية:-

$$W_i = \frac{[\Psi(e_i/\hat{\sigma})]}{(e_i/\hat{\sigma})} = \frac{\left[\Psi\left(\frac{y_i - \sum_{j=1}^{p} x_{ji}\hat{\beta}_j}{\hat{\sigma}}\right)\right]}{\left(\frac{y_i - \sum_{j=1}^{p} x_{ji}\hat{\beta}_j}{\hat{\sigma}}\right)} \tag{10}$$

لإيجاد $(\hat{\sigma})$ في المعادلة أعلاه، والتي تمثِّل قيمة الإنحراف المعياري، وأنَّ هذه القيمة تقّدر مرة واحدة فقط باستخدام القيم الأولية قبل البدء بالتكرار وهناك عدة صيغ لتقديرها منها:

$$\hat{\sigma} = 1.5\, Med|e_i| = Med|e_i|/0.6745 \tag{11}$$

$$\hat{\sigma} = 1.2\, Med|e_i| \tag{12}$$

$$\hat{\sigma} = 1.4825[Med|e_i - Mede_i|] = Med|e_i - Med(e_i)|/0.6745 \tag{13}$$

إذ تمثِّل e_i البواقي و Med يشير الى الوسيط، ولقد اقترح الباحثون عدداً من الدوال $\rho(.)$ أو مشتقاتها $\Psi(.)$ ، بحيث تجعل نتائج التقدير حصينة لا تتأثر بوجود الشواذ، وفيما يلي بعض الدوال المهمة لهذا النوع من المقدرات والمعرفة بدلالة الدالة $\Psi(e_i)$، وبافتراض أنَّ وسيط الأخطاء المطلقة(MAD: Median Absolute Deviation) الوارد في الصيغة (13) أعلاه كمقدر للانحراف المعياري، وكما يلي:

$$MAD = Med|e_i - Med(e_i)|/0.6745$$

اذ أن :0.6745 تمثل وسيط التوزيع الطبيعي القياسي.

وقد أوجد الباحثون (Montgomery et. al., 2001) الصيغة القياسية للبواقيStandardized Residuals باستخدام $eis = \frac{e_i}{MAD}$، أنَّ ثابت القطع(C: Tunning Constant) الذي يجعل التباين المقدرMAD غير متحيّز تقريباً ل σ عندما يكون حجم العينة كبير والخطأ يتوزع طبيعياً) .(Hasan and Ridha, 2011).

بعض دوال الأوزان لمقدرM

A. دالة Tukey Bisquare (Beatone and Tukey,1974):

$$\Psi_{Bisquare}(e_{is}, c) = \begin{cases} \left[1 - \left(\frac{e_{is}}{c}\right)^2\right]^2 & if|e_{is}| \le c \\ 0 & otherwise \end{cases} \tag{14}$$

اذ إنَّ c تأخذ القيمة الإفتراضية $c = 4.685$ ، دالة Huber (Huber,1964):

$$\Psi_{Huber}(e_{is}, c) = \begin{cases} 1 & if|e_{is}| \le c \\ \frac{c}{|e_i|} & otherwise \end{cases} \tag{15}$$

إذ أنَّ c تأخذ القيمة الإفتراضية $c = 1.345$ ، دالة Hampel (Grubbs,1969):

$$\Psi_{Hampel}(e_{is}, c) = \begin{cases} 1 & if|e_{is}| \le c \\ \frac{a}{|e_i|} & if\ a < |e_{is}| \le b \\ \frac{a(c-|e_{is}|)}{|e_{is}|(c-b)} & b < |e_{is}| \le c \end{cases} \tag{16}$$

إذ إنَّ القيم الإفتراضية لثوابت القطع (Tuning Constants) هي a = 2 b = 4 and c = 8

علماً أنَّ ثابت التوليف (Tunning Constant) لكلِّ دالة يستخدم لتعديل كفاءة المقدرات الناتجة لتوزيعات محدّدة ويحقق كفاءة تقريبية مقدارها (% 95) عندما تتبع الاخطاء التوزيع الطبيعي، وأنَّ الإختيار الجيّد لقيمة هذا الثابت يؤدي إلى زيادة حصانة المقدرات، لأنَّ لهذا الثابت تأثيراً كبيراً على حصانة المقدرات، وإنَّ قيمته تتراوح بين انحراف معياري واحد إلى انحرافين معياريين لقيم المشاهدات أو الاخطاء.

2- طريقة S الحصينة:

اقترح مقدر للإنحراف المعياري للأخطاء بالإعتماد على فكرة (Rousseeuw and Yohai,1984)و (Rousseeuw and Leroy,1987) الذين اقترحوا مقدرات S باعتبارها الحلّ الذي يوجد أقل قيمة تشتت ممكن للأخطاء، أي أنّه يقوم بإيجاد

$$Min_{\underline{\hat{\beta}}}\, S\left(e_1\left(\underline{\hat{\beta}}\right), e_2\left(\underline{\hat{\beta}}\right), \dots, e_n\left(\underline{\hat{\beta}}\right)\right) \quad (17)$$

بالتوازي مع مقدرات المربعات الصغرى التي تقوم بتصغير تباين الأخطاء، ويجب أنْ يكون واضحاً بأنَّ مقدرات المربعات الصغرى الإعتيادية OLS يمكن النظر إليها كحالة خاصة أقل حصانة من مقدرات S، فعندما نبدأ بتصغير تباين الأخطاء فإنَّ مقدرات S الحصينة ستقوم بتصغير مقدر التباين الحصين للأخطاء أي أنَّه يقوم بتصغير قيمة المعادلة التالية:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{\hat{\sigma}_e}\right) = b \quad (18)$$

اذ إنَّ b قيمة ثابت تعرف على أنّها $b = E_{\emptyset}[\rho(e)]$، حيث$\emptyset$ تمثّل التوزيع الطبيعي القياسي، وبعد أخذ الإشتقاق لهذه المعادلة وحلِّها نحصل على

$$\frac{1}{n}\sum_{i=1}^{n}\Psi\left(\frac{e_i}{\hat{\sigma}_e}\right) = b \quad (19)$$

حيث إنّ Ψ يتمُّ استبدالها بدالة وزن مناسبة. كما هو الحال بالنسبة لأكثر مقدرات M ، ومن أمثلة دوال الوزن دالتي Huber أو biweightأو Hampel وغيرهم. وعلى الرغم من أنَّ مقدرات S لها نقطة انهيار تبلغ 0.5 فإنَّ كلفة ذلك هو أنَّ هذه المقدرات تمتلك كفاءة منخفضة جداً حيث تبلغ كفاءتها مقارنة بطريقة المربعات الصغرى 30% فقط حسب مابين كلٍّ من (Croux et. al., 1994).

3- طريقة MM الحصينة (Rousseeuw and Leroy,1987) :

تقوم هذه الطريقة بتقدير معلمات الإنحدار باستعمال تقدير (S الحصين)، وذلك بتصغير minimize مقياس الإنحراف المعياري للبواقي من طريقة M. إنَّ طريقة (MM) تهدف للحصول على مقدرات ذات قيم عالية الدقة أو أكثر كفاءة. قبل استعمال المشاهدات في الأنموذج وهي:

$$\sum_{i=1}^{n}\acute{\rho}\,(e_{is})x_{ij} = 0 \qquad or$$

$$\sum_{i=1}^{n}\acute{\rho}\,\left(\left(\frac{y_i - \sum_{j=1}^{k}x_{ij}\hat{\beta}_j}{s_{MM}}\right)x_{ij} = 0 \quad (20)$$

إذ إنَّ s_{MM} هو الإنحراف المعياري الذي يتمُّ الحصول عليه من بواقي تقدير S و ρ هي دالة Tukeys الموزونة

$$\rho(e_{is}) = \begin{cases} \frac{e_{is}^2}{2} - \frac{e_{is}^4}{2c^2} + \frac{e_{is}^6}{6c^2} & -c \le e_{is} \le c \\ \frac{c^2}{6} & e_{is} < -c \quad or \quad e_{is} > c \end{cases} \quad (21)$$

4- طريقة المربعات الصغرى المشذبة (LTS) :Least Trimmed Squares ethod

إنَّ العديد من الباحثين لم يدرك أنَّ أداء المربعات الصغرى الإعتيادية OLS (Ordinary Least Squares) يمكن أنْ يكون ضعيفاً جداً عندما يكون شكل التوزيع الطبيعي للبيانات ذات ذيول (أطراف) ثقيلة (Heavy tails)، والتي تنشأ من القيم الشاذة حتى إذا كانت هناك قيمة شاذة واحدة فقط فانه سوف يكون لها تأثير كبير على تقديراتOLS. وللتغلب على هذه المشكلة سوف يتمُّ استخدام مقدر حصين له نقطة انهيار عالية كبديل عن طريقة OLS. ومن البدائل المتوفرة مقدر LTS الذي أُقترح من قبل (Rousseeuw and Yohai,1984) الذي يكون لديه نقطة انهيار(Breakdowm point) عالية تساوي $BP = [\{n - k/2 + 1\}/n]$.(Uraibi.et.,2009) $\hat{\beta}_{LTS}$ نحصل عليه بالشكل الآتي:

$$Min \sum_{i=1}^{h} e_i^2$$

إذ إنّ e_i^2 تمثِّل مربعات حدِّ الخطأ. وهذه الطريقة تقلل من نسبة تأثير القيم الشاذة (α) في البيانات. لتشذيب (لقطع) نسبة α اقترح (Rousseeuw and Leroy,1987) اختيار h حسب الصيغة الآَتية:-

$$h = n/2 + (p + 1)/2 \tag{22}$$

إذ إنَّ p تمثِّل عدد المعلمات. ومن مميّزات استخدام طريقة LTS السيطرة على مستوى التشذيب الذي يعتمد على تشابه النسبة المئوية المبتورة لنسبة القيم الشاذة. إذا كان هناك شك في إنَّ البيانات تحتوي على مايقرب 10% من القيم الشاذة، فإنَّ LTS سوف تقطع 10% من الطرفين أي أنَّ عملية التقدير تتمُّ بما نسبته 80% من المشاهدات الأصلية (Arif ,2012).

5- طريقة LASSO المكيفة (Wang et al. 2013) (Zou. 2006):

$$\arg \min_{\hat{\beta}} \sum_{i=1}^{n} (1 - \exp\{-(y_i - X_i^T\beta)^2/\gamma_n\}) + n\sum_{i=1}^{d} \lambda_{nj}\beta_j \tag{23}$$

إذ أن : γ_n : معلمة الضبط (التوليف) معلمة التنظيم λ_{nj} :

التطبيق العملي

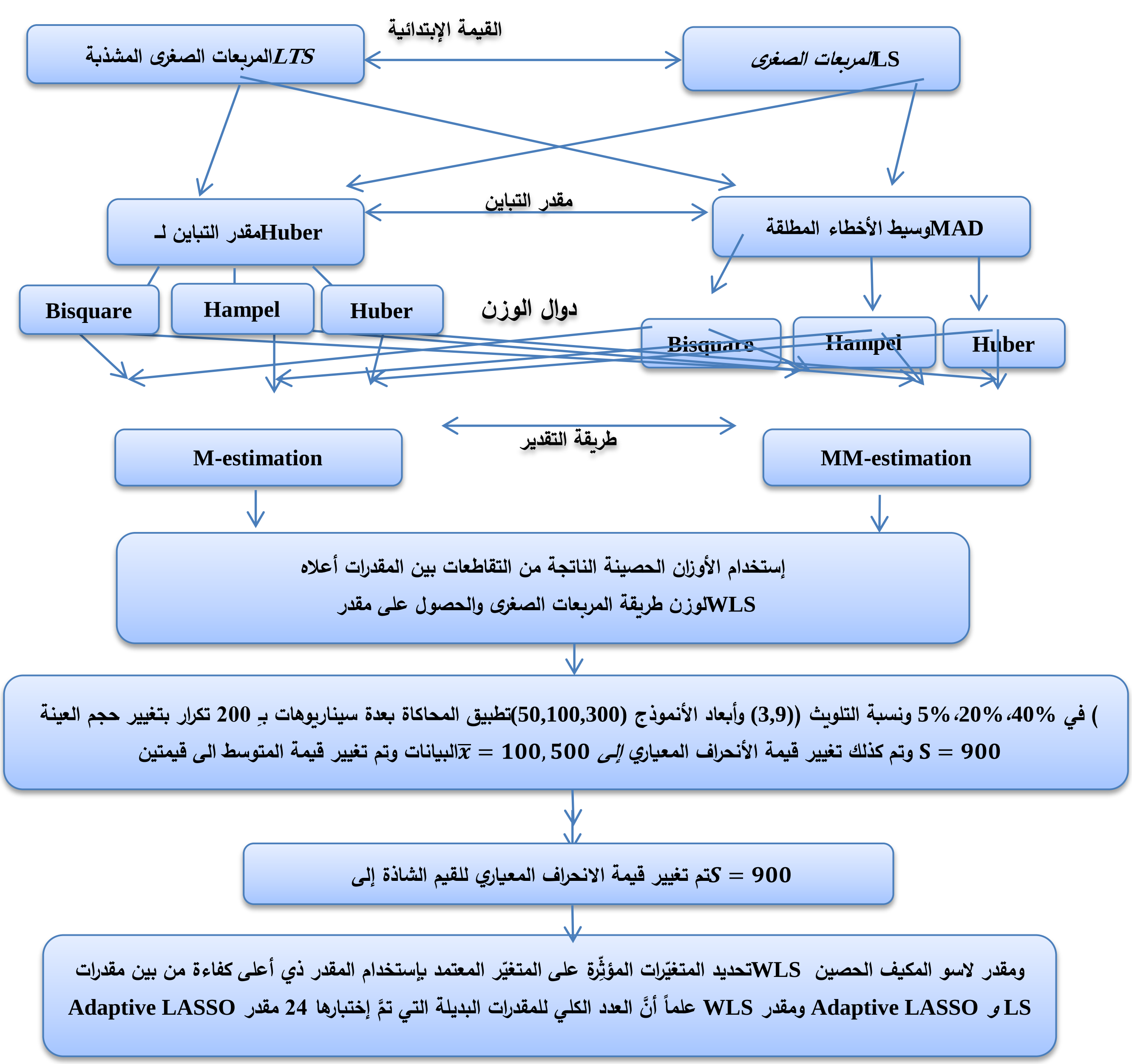
1. جانب المحاكاة (الخوارزمية المقترحة للجانب التجريبي) Simulation Side

تقوم فكرة الخوارزمية التي تسعى البحث إلى تقديمها إلى تقدير المعلمات بأسلوب M وبدوال أوزان مختلفة من خلال إدخال كلٍّ المتغيرات في الأنموذج وتطبيق عدة حالات تتمثّل بتغيير المقدرات الإبتدائية ومقدرات التباين والأوزان المستخدمة. وقد تمَّ اللجوء إلى إثنين من أهم المقدرات الحصينة، وهما مقدر M–estimation الحصين ضد القيم الشاذة في المتغيّر المعتمد (y–outliers)، وكذلك مقدر MM–estimation الذي يمتلك حصانة ضد القيم الشاذة في كلٍّ من المتغيرين المعتمد، وكذلك المتغير المستقل (تدعى القيم الشاذة في فضاء المتغيّرات المستقلة بالقيم الجاذبة أو الفعالة أو المخلة (X–leverage points)، حيث أنَّ مقدر MM–estimation هو عبارة عن مقدر M–estimation بدالة وزن Tukey's biweight معتمدة على أحد مقدرات S (S–estimator) وفي أدناه مخطط إنسيابي للخوارزمية المقترحة.

وتم استخدام ثلاثة احجام عينات $n = 50, 100, 300$، وتم أختبار بُعدين لأنموذج الانحدار كما هو واضح في الخوارزمية أعلاه وهما $m = 3, 9$ وتم استخدام ثلاثة نسب للتلويث $\alpha = 5\%, 20\%, 40\%$ وتمت عملية التلويث باستخدام طريقة تلويث صف كامل وهو الاسلوب الذي إتبعه (Toka et.al.,2021) والذي يطلق عليه casewise وتم تغيير الوسط الحسابي للقيم الملوثة مرتين، حيث أن بيانات المحاكاة قد تم توليدها بوسط حسابي = صفر وعند تلويث البيانات تم تبديل الوسط الحسابي بقيمتين 100, 500 وتم تغيير قيمة الأنحراف المعياري من $s = 100$ إلى $s = 900$ وبذلك أصبحت القيم الملوثة مختلفة من حيث قيمة المتوسط والتباين وكل الملوثات قد تم توليدها من التوزيع الطبيعي وتم تكرار توليد $B = 200$ عينة لأختبار كل طريقة من طرق التقدير المقترحة وكانت سيناريوهات التلويث تتضمن تلويث y مره والمتغيرات المستقلة X'S مرة،

وكليهما معاً مرة ثالثة وبلغ إجمالي نماذج المحاكاة التي تم توليدها في دراسة المحاكاة هو 561,600، حيث تم توليدها وفقاً للتالي : $561,600 = 26 \times 3 \times 200 \times 2 \times 2 \times 3 \times 2 \times 3$ أنموذج رياضي مقدر (علماً أن كل مقدر حصين يحتاج إلى ما أقصاه 20 تكراراً ليصل الى التقارب، أي أن عدد النماذج المقدرة فعلياً قد يصل إلى $11,232,000 = 20 \times 561,600$ أنموذجاً مقدراً).
المخطط الانسيابي الوارد في البحث عبر عن الخوارزمية المقترحة حيث اعتمدت الباحثة على مفهوم الخوارزمية الذي يمثل مجموعة من الخطوات الرياضية والمنطقية والمتسلسلة اللازمة لحل مشكلة البحث لتحصين عملية اختيار المتغيرات في معادلة الانحدار وتم رسم المخطط الانسابي Flow Chart للتعبير عن خطوات الخوارزمية.

المخطط الانسيابي للطريقة المقترحة لإختيار أفضل مقدر لاختيار أفضل معادلة إنحدار



لغرض أختيار الأنموذج الأفضل تم تنفيذ على بيانات مولدة بأستخدام برنامج R حيث تم أختيار الطريقة التي أثبتت أعلى كفاءة (أقل RMSE) في تجارب المحاكاة وتمت مقارنتها مع طريقة Adaptive LASSO الحصينة مقارنة بالطريقة الأكثر شيوعاً، وقد ثبت بأنها قد تغلبت على طريقة LASSO بعد أختيار أعلى دوال المربعات الصغرى الموزونة كفاءةً.

جدول (1) : الطرق التي حققت أفضل أداء (أقل RMSE) في حالة وجود قيم شاذة في المتغير المعتمد

أكفأ طريقة	قيمة المتوسط	حجم العينة	محور الشواذ	نسبة التلويث
لنموذج بثلاثة متغيرات				
WLS.LS.Bisquare.Mad.M	عند قيمة وسط حسابي = 100	50	Y-outlier	6%
WLS.LTS.HAMPEL.Huber.M				20%
				40%
WLS.LS.Bisquare.Mad.M		100		5%
				20%
				40%
WLS.LS.Hampel.Huber.M		300		5%
WLS.LS.Hampel.Mad.M				20%
WLS.LS.Bisquare.Mad.M				40%
لم تتفوق أي من الطرق على نظيراتها	عند قيمة وسط حسابي = 500	50		6%
WLS.LS.Hampel.Mad.M				20%
WLS.LTS.HAMPEL.Huber.M				40%
WLS.LS.Hampel.Huber.M WLS.LS.Bisquar.Mad.M		100		5%
WLS.LS.Bisquar.Mad.M WLS.LS.Hampel.Huber.M				20%
WLS.LS.Bisquar.Mad.M				40%
WLS.LS.Hampel.Huber.M		300		5%
				20%
WLS.LS.Hampel.Mad.M WLS.LTS.HAMPEL.Huber.M				40%
لنموذج بتسعة متغيرات				
WLS.LS.HAMPEL.Huber.M	عند قيمة وسط حسابي = 100	50	Y-outlier	6%
WLS.LS.Bisquar.Mad.M				20%
				40%
WLS.LS.HAMPEL.Huber.M		100		5%
WLS.LTS.HAMPEL.Huber.M				20%
WLS.LS.Bisquar.Mad.M				40%
WLS.LS.HAMPEL.Huber.M		300		5%
WLS.LS.Bisquar.Mad.M				20%
				40%
WLS.LS.HAMPEL.Huber.M	عند قيمة وسط حسابي = 500	50		6%
WLS.LS.Bisquar.Mad.M				20%
				40%
WLS.LS.HAMPEL.Huber.M		100		5%
WLS.LTS.HAMPEL.Huber.M				20%
WLS.LS.Bisquar.Mad.M				40%
WLS.LS.HAMPEL.Huber.M		300		5%
WLS.LS.Bisquar.Mad.M				20%
				40%

جدول (2) :الطرق التي حققت أفضل أداء (أقل RMSE) في حالة وجود قيم إنعطاف (شواذ) في المتغيرات المستقلة

أكفأ طريقة	قيمة المتوسط	حجم العينة	محور الشواذ	نسبة التلويث
لنموذج بثلاثة متغيرات				
WLS.LTS.HAMPEL.Huber.M	عند قيمة وسط حسابي = 100	50	X-leverage	6%
				20%
				40%
		100		5%
				20%
				40%
		300		5%
				20%
WLS.LS.Hampel.Mad.M LS				40%
WLS.LTS.HAMPEL.Huber.M	عند قيمة وسط حسابي = 500	50		6%
				20%
				40%
		100		5%
				20%
				40%
		300		5%
				20%
LS				40%
لنموذج بتسعة متغيرات				
WLS.LTS.HAMPEL.Huber.M	عند قيمة وسط حسابي = 100	50	X-leverage	6%
				20%
				40%
		100		5%
				20%
				40%
		300		5%
				20%
				40%
WLS.LTS.HAMPEL.Huber.M	عند قيمة وسط حسابي = 500	50		6%
				20%
				40%
		100		5%
				20%
				40%
		300		5%
				20%
				40%

جدول (3) : الطرق التي حققت أفضل أداء (أقل RMSE) في حالة وجود قيم شاذة في المتغيرين X وY

لنموذج بثلاثة متغيرات				
WLS.LTS.HAMPEL.Huber.M	عند قيمة وسط	50	X-leverage	6%
				20%

WLS.LS.Bisquar.Mad.M				40%
WLS.LTS.HAMPEL.Huber.M		100		5%
				20%
WLS.LS.Bisquar.Mad.M				40%
WLS.LTS.HAMPEL.Huber.M		300		5%
WLS.LTS.HAMPEL.Huber.M WLS.LS.Bisquar.Mad.M				20%
WLS.LS.Hampel.Huber.M				40%
WLS.LTS.HAMPEL.Huber.M	عند قيمة وسط حسابي = 500	50		6%
				20%
WLS.LS.Bisquar.Mad.M				40%
WLS.LTS.HAMPEL.Huber.M		100		5%
				20%
				40%
		300		5%
				20%
LS				40%
لنموذج بتسعة متغيرات				
WLS.LTS.HAMPEL.Huber.M	عند قيمة وسط حسابي = 100	50	X-leverage Y-outlier	6%
				20%
				40%
		100		5%
				20%
WLS.LS.Hampel.Mad.M LS				40%
WLS.LTS.HAMPEL.Huber.M		300		5%
				20%
LS				40%
WLS.LTS.HAMPEL.Huber.M	عند قيمة وسط حسابي = 500	50		6%
				20%
WLS.LS.Bisquar.Mad.M				40%
WLS.LTS.HAMPEL.Huber.M		100		5%
				20%
WLS.LS.Bisquar.Mad.M				40%
WLS.LTS.HAMPEL.Huber.M		300		5%
				20%
LS				40%

وعند النظر إلى الجداول الثلاثة (1) و (2) و (3) أعلاه الذين يمثلون ملخص نهائي لكافة تجارب المحاكاة لكل احجام النماذج والعينات ونسب التلويث ودوال الوزن وقيمتي المتوسط، حيث تظهر الطرق الأعلى كفاءة لكل حالة وتعطي الباحثين خارطة طريق حول طريقة التقدير المناسبة لكل حالة. وإذا وددنا إستخلاص طريقة متفوقة إجمالاً لكل حالة من حالات تلويث البيانات بالقيم الشاذة وكما هو مبين فيما يلي:

A. **كانت طريقة LS.Bisquare.Mad.M** أكثر الطرق التي حققت أعلى كفاءة (اقلهم بقيمة RMSE) من بين كل الطرق التي تمت مقارنتها في دراسة المحاكاة عند وجود شواذ في المتغير المعتمد y–outliers في اغلب الحالات التي تمت مقارنتها

B. **كانت طريقة LTS.HAMPEL.Huber.M** أكثر الطرق التي حققت أعلى كفاءة (اقلهم بقيمة RMSE) من بين كل الطرق التي تمت مقارنتها في دراسة المحاكاة عند وجود شواذ في المتغيرات المستقلة X–leverage points في اغلب الحالات التي تمت مقارنتها

C. **كانت طريقة LTS.HAMPEL.Huber.M** أكثر الطرق التي حققت أعلى كفاءة (اقلهم بقيمة RMSE) من بين كل الطرق التي تمت مقارنتها في دراسة المحاكاة عند وجود شواذ في المتغير المعتمد y–outliers وفي المتغيرات المستقلة X–leverage points في اغلب الحالات التي تمت مقارنتها

والنقاط الواردة في أعلاه تعطي خارطة طريق موجزة يمكن للباحثين إستخدامها لاختيار طريقة التقدير الملائمة في حالة عدم توفر معلومات واضحة حول نسب الشواذ في البيانات ونوعها.

2. الجانب التطبيقي Applied Side

جمع البيانات (Al-Nuaimi, 2005) :

طبقت هذه الدراسة على بيانات جمعت من مستشفى ابن الأثير التعليمي للولادة والأطفال في الموصل ، من مرضى مصابين بفقر دم البحر الأبيض المتوسط من نوع بيتا او ما يسمى الثلاسيميا ، وكان عدد المشاهدات (150) حيث أن المتغير المعتمد أو المستجيب في الدراسة مثل عمر العظم مقاساً بالشهر Age of the Bone (Month) كما واختير عدد من المتغيرات التي يعتقد بأنها تؤثر فيه بعد مراجعة أطباء اختصاصيين وممارسين في مرض الثلاسيميا.

Variable Name	أسم المتغير	
Real Age (Month)	العمر الحقيقي (مقاساً بالشهر)	X1
Onset of Disease (Month)	عمر المريض عند ظهور المرض (مقاساً بالشهر)	2X
c.m)(Enlargement of liver	تضخم الكبد (مقاساً بالسنتمتر)	3X
Hemoglobin	هيموكلوبين الدم	4X
Packed cell volume	مكداس الدم (خلايا الدم المضغوط)	5X
Reticulacyte	الخلايا الشبكية	6X
Normooblast	الأرومة الحمراء	7X
Fetal Hemoglobin	الهيموكلوبين الجيني	8X
Number of Blood units	عدد وحدات الدم	9X
Onset of Blood Trans fusion To According	بداية نقل الدم حسب العمر (مقاساً بالشهر)	10X

جدول (4): كفاءة النماذج النهائية لكل الطرق المقارنة

الطريقة	X.Intercept.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	R^2	R^2_{adj}	AIC	p – value
LS(كل الانحدارات الممكنة بدون أوزان)	1	1	0	0	1	0	0	0	0	1	0	0.817	0.814	1274.478	2.2e-16
LS(الحذف العكسي بدون أوزان) LS (الاختيار الامامي بدون أوزان) LS (الانحدار التدريجي بدون أوزان)	1	1	1	0	1	1	0	1	0	1	0	0.836	0.8291	836.72	2.2e-16
WLS.LS.Bisquare.Mad.M	1	1	0	0	1	0	0	0	0	1	0	0.882	0.88	1205.105	2.2e-16

(كل الانحدارات الممكنة)														
WLS.LS.Bisquare.Mad.M (طريقة الحذف العكسي) WLS.LS.Bisquare.Mad.M (طريقة الإختيار الأمامي) WLS.LS.Bisquare.Mad.M (الإنحدار التدريجي)	1	1	1	1	1	1	0	1	0	1	0	0.9036	0.8989	729.29	2.2e-16
Robust Variable Selection Exponential Squared Loss	1	1	0	0	1	0	0	0	0	0	0	0.7415	0.7229	-181.9097	Not Available

وبالنظر إلى الجدول (4) أعلاه نلاحظ أن الطرق المختلفة قد أنتجت نماذجاً مختلفة، فمنها ما قام بإستبعاد أغلب المتغيرات مثل طريقتي كل الانحدارات الممكنة وطريقة LASSO اللتان أبقيتا 3 و 2 متغيرات في الأنموذج على التوالي وهذا خلاف ماتمت مناقشته مع أطباء الإختصاص الذين أكدوا بأن المتغيرات العشرة الموجودة في الأنموذج مهمة وتؤثر على الإصابة بمرض الثلاسيميا وعليه فإن طرق الاختيار الامامي والحذف العكسي والانحدار التدريجي التي تم وزنها بالأوزان الحصينة والتي حققت أعلى كفاءة في دراسات المحاكاة في غالبية الحالات وهي طريقة WLS.LS.Bisquare.Mad.M التي جمعت بين الحصانة والكفاءة فضلاً عن إتفاقها مع الرؤية الطبية لهذا المرض.

ومن ناحية أخرى فإن المقارنة بين مقاييس الكفاءة بين نماذج تختلف في عدد متغيراتها غير مجدية لأن تلك المقاييس ترتبط بعدد المتغيرات الموجودة في الأنموذج، فعلى سبيل المثال فإن قيمة معامل التحديد تزداد بزيادة عدد المتغيرات وذلك لأن كل متغير مضاف إلى الأنموذج يمتلك مجموع مربعات موجب وبالتالي فإن إضافته ستؤدي إلى إضافة قيمة موجبة إلى البسط، وبما أن مقام حساب معامل التحديد يحتوي على مجموع المربعات الكلي فبالتالي فهو قيمة ثابتة وزيادة البسط ستؤدي إلى زيادة القيمة الكلية للكسر وعليه تميل الباحثة إلى إعتماد نتائج المحاكاة التي تم فيها تنفيذ عملية التقدير مئات الالاف من المرات وأعتماد الأنموذج النهائي الذي نتج من طريقة WLS.LS.Bisquare.Mad.M كنموذج نهائي يتم إعتماده لتشخيص أهم المتغيرات المؤثرة على الاصابة بالثلاسيميا. ونفس الامر ينطبق على معيار AIC والذي ترتبط عملية حسابه أيضاً بعدد المتغيرات في الأنموذج كون أن زيادة عدد المتغيرات سيطرح رقماً موجباً من البسط $(sse \times (n-p-1))$ وبذلك فإن طرح رقم صغير أو كبير من n سيكون له تأثير كبير على قيمة لوغارتم المقدار فتنتج لدينا أحياناً قيمٌ سالبة تجعل قيمة AIC سالبة. أما لو كانت النماذج التي تتم مقارنتها من نفس البعد (نفس عدد المتغيرات) لأمكنت مقارنتها باستخدام تلك المعايير ولكن هذا لايتحقق في طرق إختيار أفضل معادلة إنحدار التي تعمل على الموازنة بين تقليل عدد المتغيرات في الأنموذج وبين زيادة كفاءة الأنموذج المقدر.

الاستنتاجات

بناءاً على ماتمّ التوصل إليه في الجانبين التجريبي والتطبيقي وصل البحث إلى مجموعة من الاستنتاجات، وكما يلي:

1. حافظت طريقة المربعات الصغرى الموزونة على خصائصها وإمكانية تطبيقها في كلِّ الحالات وللنماذج التي تعاني من كل أنواع الشواذ سواءاً كانت في المتغير المعتمد أو المتغيرات المستقلة أو كليهما معاً.
2. لم تكن طريقة LASSO الحصينة موفقة في كل الحالات، فضلاً عن أنَّها تستبعد الكثير من المتغيرات التي قد يكون بعضها مهماً، بالإضافة إلى أنها تعاني من بعض المشاكل الحسابية المتعلقة بالتقارب حيث توقفت الدالة عن التنفيذ في بعض الحالات.
3. تمكن البحث ومن بين 561,600 تجربة محاكاة من تثبيت بعض المقدرات الأعلى كفاءة والتي ينصح الباحثين باستخدامهم حسب موقع الشواذ وفي حالتي النماذج بالابعاد القليلة أو العالية لحصول مقدراتها على أقل قيمة لجذر متوسط مربعات الخطأ، وكما يلي:

A. في حالة وجود الشواذ في المتغيرات المستقلة (القيم الفعالة Leverage Points) يفضل عند إيجاد مقدر M الحصين إستخدام مقدر المربعات الصغرى المشذبة (LTS) كقيمة ابتدائية وإستخدام دالة Hampel كدالة وزن، ومقدر Huber كمقدر للتشتت، وطريقة M في التقدير كما تم تطبيقه في طريقة LTS.HAMPEL.Huber.M.

B. في حالة وجود الشواذ في المتغيرالمعتمد (Y-outliers) يفضل عند إيجاد مقدر M الحصين إستخدام مقدر المربعات الصغرى (LS) كقيمة ابتدائية وإستخدام دالة Bisqure كدالة وزن، ومقدر وسيط الأخطاء المطلقة كمقدر للتشتت، وطريقة M في التقدير كما هو الحال في مقدر LS.Bisquare.Mad.M الذي تم إستخدامه في البحث.

C. في حالة وجود الشواذ في المتغير المعتمد والمتغيرات المستقلة (القيم الفعالة Leverage Points والقيم الشاذة y-outliers) يفضل عند إيجاد مقدر M الحصين إستخدام مقدر المربعات الصغرى المشذبة (LTS) كقيمة ابتدائية وإستخدام دالة Hampel كدالة وزن، ومقدر Huber كمقدر للتشتت، وطريقة M في عملية التقدير كما تم بيانه في المقدر LTS.HAMPEL.Huber.M.

بناءاً على ماورد في أعلاه نستنتج أن الكفاءة في حساب الاوزان تمكن طريقة المربعات الصغرى (الموزونة WLS) من إنتاج مقدرات كفوءة كما تم إثباته وبالتالي يجعل إستخدام الطرق التقليدية في إختيار أفضل معادلة إنحدار ممكناً في تحديد أهم المتغيرات المؤثرة على المتغير المعتمد وكما تم إثباته في الجانبين التجريبي والتطبيقي حيث تم إختزال النماذج بشكل جمع بين الكفاءو والحصانة.

Reference

1. Al-Subaihi, Ali, A. (2004). " Variable Selection in Multivariable's Regression Using SAS/IML ", Institute of Public Admin station, Saudi Arabia.
2. Hurn , D. and Mirosevich , V. , (2008) . "Modern Robust Statistical Methods" , The American Psychologist Association , Vol. 63, No. 7, PP. 591–601.
3. Huber,P.J.(1973). "Robust estimation of location parameter" , Annals of Mathematical Statistics, 35: 73-101.
4. Montgomery D. C., Peck E. A., and Vining G. C. (2001). " Introduction to Linear Regression Analyses", 3rd ed., New york: John Wiley&Sons Inc, Library of congress, U.S.A.
5. Hasan, T. A. and Ridha, M. S. (2011). "Using Robust Regression to Find the Most Appropriate Model to Represent Meteorological Data in the City of Erbil during (1998-2010)", Journal of Administration and Economics, thirty four years, 39 year, Issue: 39.
6. Beaton, A.E., and Tukey, J.W., (1974)."The Fitting of Power series Meaning Polynomials, Illustrated on band – Spectroscopic Data", Techno metrics, 16, PP 147 – 185.
7. Huber,P.J. (1964)."Robust Estimation of location parameter".Ann .math .stat.35, PP 73-101.
8. Grubbs,F.E.(1969)."Procedures for Detecting Outlying observation", Technimetrics.
9. Croux, C. Rousseeuw, P. J. and Hossjer, O. (1994)." Generalized S-Estimators" Journal of the American Statistical Association , 89(428), pp.1271-1281.
10. Rousseeuw, P. J .and Leroy, A. M. (1987). Robust Regression and Outlier Detection, Wiley-Interscience , New York (series in Applied probability and statistics) , 329 Pages . ISBNO-471-852333.
11. Rousseeuw P.J. and Yohai V.J. (1984). "Robust regression by means of S-estimators. in Robust and Nonlinear Time Series Analysis", eds. J. Franke. W. Härdle and R.D. Martin. Lecture Notes in Statistics 26, New york: Springer Verlag. <ftp://ftp.win.ua.ac.be/pub/preprints/84/ Robreg84.pdf>
12. Uraibi , H. S. , Talib , B. A. , and Yousif , J. H. , (2009)."Linear Regression Model Selection Based on Robust Bootstrapping Technique" , American Journal of Applied Sciences , 6 (6), PP.1191-1198.

13. Arif, A. L. (2012). "Using Types of Robust Regression to Deal with Outying Values in Simple Linear Regression", Dhi Qar Journal of Agricultural Research, College of Science - Dhi Qar University, Iraq, 1 (2) , PP 93-103.

14. Wang, X., Yunlu J., Mian H., and Heping Z. (2013). "Robust Variable Selection with Exponential Squared Loss." *Journal of the American Statistical Association* 108(502):632–3. https://doi.org/10.1080/01621459.2013.766613.

15. Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101 (476): 1418–29. https://doi.org/10.1198/016214506000000735.

16. Toka, O., Çetin, M., and Arslan, O. (2021). "Robust regression estimation and variable selection when cellwise and casewise outliers are present" ,Hacettepe Journal ,Volume 50 (1), 289 – 303.

17. Al-Nuaimi, A. M. (2005). "Variable Selection in Ridge Regression", Un-Published MSc. Thesis, Faculty of Computer Science and Mathematics, University of Mosul – Iraq.

Robust Weighted Least Squares Method using different schemes of M-estimators (RWLSM), A comparative Study

Fatima M. Selevany and Bashar A. Al-Talib

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Abstract: In this research, it was reducing or excluding the effect of not satisfying the assumption of normal distribution of the data, due to the presence of types of outlying values in it when we wish to choose the best regression equation by robust methods, and this was achieved by introducing weights from the robust methods in the estimate and testing their robustness and suitability for the model in advance, And then selecting the weights resulting from the highest efficient robust methods and introducing these weights in the stages of selecting best regression equation, which results in a model that achieves two characteristics at the same time, which are robustness and reducing dimensions in return for increasing efficiency.

The simulation approach was used on models with different dimensions, different sample sizes, and different contamination rates in the dependent variable once, in the independent variables again, and in both together, with a focus on studying the possible impact of the presence of outliers on the variables that will remain in the model and the variables that will be deleted. .

To achieve the idea of the paper, a number of robust estimation methods were studied, and the results were compared with the ordinary least squares method (OLS) and the robust adaptive LASSO method on experimental data using simulation, as well as on data for a sample of thalassemia patients in Nineveh province.

Keyword: Variables Selection ،(M) M-estimators ،Weighted Least Squares ،Robust Regression ،Outliers Detection.