تخمين حصين لمعلمتي أنموذج العلاقة الدالية الخطية البسيطة عند تكرار المشاهدات

ROBUST ESTIMATION OF THE PARAMETERS OF SIMPLE LINEAR FUNCTIONAL RELATION MODEL WITH REPLICATION OF OBSERVATIONS

خلود صباح عميد، شوقي شاكر حسين

قسم الرياضيات – كلية العلوم – جامعة بغداد. بغداد – العراق

الخلاصة

يتطرق هذا البحث للسيرة التاريخية لمسالة تخمين معلمتي أنموذج العلاقة الدالية الخطية البسيطة التي تعود إلى عهد ليس بالقريب فهي تمتد إلى قرابة قرن وربع القرن من الزمن. إذ يتناول أهم من عمل على هذه المسالة والطرائق التي استحوذت على حلها والمخمنات التي تولدت منها وذلك وفق ضوابط متنوعة حول توزعات المتغيرات العشوائية التي يضمها الأنموذج. كما ويتضمن العديد من المخمنات الحصينة المقترحة عند توفر مشاهدات مكررة في وجود الشوارد. كذلك سعينا إلى المقارنة بين مخمنات مختلف الأساليب باستخدام أسلوب المحاكاة وذلك نظراً لغياب الجانب النظري من اجل الوقوف على كفاءة أدائها والتي تظهر النتائج بان الأساليب المقترحة مشجعة ويمكن الركون إليها.

1. مقدمة

هب توفر العلاقة الخطية البسيطة بين المتغيرين الحقيقيين X^* و Y^* والمعطاة كما في الصيغة الرياضية الآتية: $Y^* = \alpha + \beta X^*$, حيث α و β قيمتان حقيقيتان مجهولتان يمثل كل منهما على التوالي تقاطع وميل الخط المستقيم لهذه العلاقة. كذلك افرض أن أي قيمة من القيم الحقيقية لكل من X^* و Y^* لا يمكن مشاهدتها وان ما يمكن مشاهدته بالنسبة إلى X^* هو قيمة X مضافاً إليها قيمة المتغير العشوائي δ وبالنسبة إلى Y^* هو قيمة Y مضافاً إليها قيمة المتغير العشوائي $\in$؛ أي أن وبتعبير رياضياتي يمكن مشاهدة المتغيرين العشوائيين X و Y المعطيين في العلاقتين الآتيتين: $X = X^* + \delta$ و $Y = Y^* + \in$. تدعى هذه العلاقة بين X^* و Y^* في حالة وجود خطأ عشوائي في مشاهدة كل منهما بأنموذج العلاقة الدالية, كما تدعى كل من α و β بمعلمة الأنموذج.

أن جوهر مسالة هذا الأنموذج يكمن في إيجاد مخمن مناسب لكل من α و β وذلك اعتماداً على عينة بحجم n من مشاهدات المتجه العشوائي (X, Y) ولتكن (X_i, Y_i)؛ $i = 1,2,...,n$, حيث كل منها يحقق صيغ العلاقات الآتية:

$$Y_i^* = \alpha + \beta X_i^*;\ X_i = X_i^* + \delta_i;\ Y_i = Y_i^* + \in_i;\ i = 1,2,...,n$$

وان $(\delta_i, \in_i)$؛ $i = 1,2,...,n$ عينة تتبع المتجه العشوائي $(\delta, \in)$.

على وفق ضوابط معينة حول توقع وتباين δ و $\in$ والترابط بينهما وعلى مسيرة تاريخية امتدت إلى حوالي قرن وربع القرن هيمنت مسالة تخمين α و β على اهتمام العديد من الباحثين. إذ كانت ولا تزال تعد محورا مهماً في حقل الاستدلال الإحصائي لما لهذا الأنموذج من تطبيقات عملية في نواحي حياتية مختلفة.

تاريخياً وكما تحدث ولد Wald [1] عام 1940 أن هذه المسالة قد أثيرت من قبل Adcock عام 1877 حيث اقترح فكرة التحليل الهندسي لحل المسالة والتي تعتمد على تصغير مجموع مربعات المسافة العامودية بين المشاهدات والخط المستقيم, وان هذا الأسلوب كان قد اعتمده أيضاً بيرسون Pearson عام 1901. ولكن روز Rose عام 1928 كان قد انتقد هذا الحل وبالمقابل أعطى خلاصة لعدة طرائق (فسرت بعد ذلك هندسياً من قبل Jones عام 1937) موضحاً فيها حقيقة هامة: أن طرائق حل المسالة محكومة بمعرفة احد اثنين على الأقل وهما إما تباين δ وليكن σ_δ^2 وتباين $\in$ وليكن $\sigma_\in^2$ أو نسبة $\sigma_\in^2$ إلى σ_δ^2 ولتكن λ; عام 1934 طور Frisch الجانب النظري لطرائق لا تعتمد مبدأ الاحتمال, في حين Koopmans عام 1937 ربط عمل Frisch بالجانب النظري للاحتمال سعياً منه في الحصول على بناء اشمل مفترضاً توفر شرط التوزع الطبيعي ومشدداً على عدم جدوى هذه الطرائق دون معرفة قيمة λ. إما Allen عام 1939 فقد تبنى أسلوبا آخر يعتمد توفر قيم اثنين من ثلاثة: $\sigma_\in^2$, σ_δ^2, ومعامل الارتباط بين δ و $\in$.

فيما يخص بحث Wald فانه كان قد أعطى حلاً وفق الضوابط الآتية: $\delta_1,\delta_2,...,\delta_n$ عينة غير مترابطة كل منها يتبع δ وان $E\delta=0$ و σ_δ^2 منته، $\in_1,\in_2,...,\in_n$ عينة غير مترابطة كل منها يتبع $\in$ وان $E\in=0$ و $\sigma_\in^2$ منته, وكذلك العينتان غير مترابطتين. فان مخمني α و β يعطيان على التوالي كما في الصيغتين التاليتين:

$$\hat{\alpha}=\bar{Y}-\beta\bar{X}\ ;$$

$$\hat{\beta}=\left(\sum_{i=1}^{m}Y_i-\sum_{i=m+1}^{n}Y_i\right)\Big/\left(\sum_{i=1}^{m}X_i-\sum_{i=m+1}^{n}X_i\right),$$

حيث $m=\frac{n}{2}$ عندما n عدد زوجي وفي حالة n فردي تهمل المشاهدة $\frac{n+1}{n}$, وان لا يكون المقام في الصيغة الثانية صفراً, ثم برهن Wald أنهما مخمنين متسقين, إن نظرة متفحصة للمخمن $\hat{\beta}$ تكشف لنا أن Wald يقوم بتقسيم المشاهدات إلى مجموعتين متساويتين (عشوائياً أو بعد ترتيب تبعاً لقيم X) ثم حساب ميل الخط المستقيم الواصل بين معدل المجموعتين. أشار Bartlett [2] عام 1949 أن كفاءة مخمني أسلوب Wald تزداد إذا تم تقسيم المشاهدات إلى ثلاث مجموعات بدلاً من مجموعتين مرتبة تبعاً لقيم X ومن ثم إهمال المجموعة الوسطى في حساب $\hat{\beta}$.

على غرار ما قام به Wald اقترح Theil [3] عام 1950 أسلوباً جديداً يدعى أسلوب ثايل التام لتخمين المعلمتين α و β وهما:

$$\hat{\alpha}=med\left\{Y_i-\hat{\beta}X_i\,;i=1,2,...,n\right\};$$

$$\hat{\beta}=med\left\{\frac{Y_j-Y_i}{X_j-X_i};X_j\neq X_i;1\le i<j\le n\right\},$$

أي أن المخمن $\hat{\beta}$ هو وسيط كل الميول المتاحة. يعاب أسلوب ثايل العدد الكبير للميول المراد حسابها وهو اقل أو يساوي $\binom{n}{2}$ ولتجاوز هذه المعضلة اقترح ثايل الأسلوب غير التام وذلك باستخدام جزء من الميول وكما يلي:

$$\hat{\beta}=med\left\{\frac{Y_{m+i}-Y_i}{X_{m+i}-X_i};i=1,2,...,m\right\}$$

حيث $m=\frac{n}{2}$ عندما n عدد زوجي وفي حالة n عدد فردي تهمل المشاهدة $\frac{n+1}{n}$ حيث المشاهدات مرتبة تبعاً لقيم X.

عام 1959 أوضح مادنسكي Madansky [4] إمكانية استخدام أسلوب الارجحية العظمى لتخمين α و β (كذلك علم 1961 خصص كريبل Graybill [5] فصلاً في كتابه – الفصل التاسع حول نفس الموضوع) وفق الضوابط التالية: $\delta_1,\delta_2,...,\delta_n$ عينة عشوائية تتبع $\delta\sim N(0,\ \sigma_\delta^2)$ مستقلة عن $\in_1,\in_2,...,\in_n$ عينة عشوائية تتبع $\in\sim N(0,\ \sigma_\delta^2)$ ومعرفة قيمة λ. فان مخمني α و β يعطيان على التوالي في الصيغتين التاليتين:

$$\hat{\alpha}=\bar{Y}-\hat{\beta}\bar{X}\,;\hat{\beta}=\frac{\left(S_Y^2-\lambda S_X^2\right)^2+\left[\left(S_Y^2-\lambda S_X^2\right)^2+4\lambda S_{XY}^2\right]^{\frac{1}{2}}}{2S_{XY}}$$

حيث أن:

$$S_X^2=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2;S_Y^2=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2;S_{XY}=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right).$$

إما في عام 1966 فقد استخدم سبرينت Sprent [6] أسلوب اصغر المربعات الموزونة لتخمين معالم الأنموذج مفترضاً أن δ و $\in$ مترابطان, هذا وتناول في كتابه المنشور عام 1969 [7] بعض جوانب حل هذه المسالة. في حين عام 1970 أعطى موران Moran [8] خلاصة وافية حول الموضوع, كذلك كيندال وستيوارت Kendall and Stuart [9] في كتابهما –الفصل 29–المحدث عام 1979. وفي عام 1987 ظهر كتاب فولر Fuller [10] الموسوم "نماذج قياس

الخطأ" تناول فيه شتى الضوابط وسبل تخمين α و β والذي يعد اليوم واحدا من المصادر المهمة للموضوع.

وحيث أن العقود الأخيرة من القرن الماضي قد شهدت تطوراً كبيراً في مجال المخمنات الحصينة فان أنموذج العلاقة الدالية لابد أن يشمله هذا النوع من التخمين. الملاحظ أن مخمنات ثايل تعد حصينة وذلك كون الوسيط اقل تأثراً بالشوارد Outliers من المعدل, كما وان مخمنات Wald عند إحلال الوسيط بدل المعدل تعد هي الأخرى حصينة. درس مختار Mokhtar [11] عام 1989 إمكانية استخدام أسلوب M- المحور إلى أسلوب W- مفترضاً معرفة قيمة λ مع دراسة مقارنة باستخدام المحاكاة في وجود الشوارد وذلك كما في الصيغ التالية:

$$\hat{\alpha}^{(k+1)} = \bar{Y}^{(k)} - \hat{\beta}^{(k+1)}\bar{X}^{(k)};\ \hat{\beta}^{(k+1)} = \frac{\left(S_Y^{2(k)} - \lambda S_X^{2(k)}\right) + \left[\left(S_Y^{2(k)} - \lambda S_X^{2(k)}\right)^2 + 4\lambda S_X^{2(k)}\right]^{1/2}}{2S_{XY}^{(k)}}$$

حيث:

$$\bar{X}^{(k)} = \frac{\sum_{i=1}^{n} W_i^{(k)} X_i}{\sum_{i=1}^{n} W_i^{(k)}};\ \bar{Y}^{(k)} = \frac{\sum_{i=1}^{n} W_i^{(k)} Y_i}{\sum_{i=1}^{n} W_i^{(k)}};\ S_X^{2(k)} = \sum_{i=1}^{n} W_i^{(k)}\left(X_i - \bar{X}^{(k)}\right)^2$$

,

$$S_Y^{2(k)} = \sum_{i=1}^{n} W_i^{(k)}\left(Y_i - \bar{Y}^{(k)}\right)^2;\ S_{XY}^{(k)} = \sum_{i=1}^{n} W_i\left(X_i - \bar{X}^{(k)}\right)\left(Y_i - \bar{Y}^{(k)}\right)^2$$

,

$$W_i^{(k)} = \frac{\Psi\left(r_i^{(k)}/S^{(k)}\right)}{r_i^{(k)}/S^{(k)}};\ r_i^{(k)} = \frac{Y_i - \hat{\alpha} - \hat{\beta}^{(k)}X_i}{\lambda + \hat{\beta}^{2(k)}};\ S^{(k)} = \frac{med\left\{\left|r_i^{(k)}\right|\right\}}{0.6745}$$

وان Ψ هي إحدى الدوال المتناظرة التي تستخدم في أسلوب M- مثل الدوال: هوبر Huber, اندروز Andrews, توكي Tukey وغيرها (انظر المصدر [12]), هذا مع اختيار قيمة ابتدائية مناسبة لكل من α و β. وكأسلوب حصين آخر في عام 1994 أقدم كل من لندر وبابو Linder and Babo [13] على تخمين معالم الأنموذج باستخدام أسلوب Booststrapping حيث ينبغي أن تكون قيمة λ معلومة.

وقبل أن نختم هذا البند لابد من الإشارة إلى أن هنالك العديد من البحوث التي تناولت هذه المسالة ولم نشر إليها (انظر المصدر [14]), كذلك يرافق معظم البحوث المنشورة لهذه المسالة أنموذجاً آخر يدعى أنموذج العلاقة البنيوية Structural Relation حيث فيه كل من X^* و Y^* متغير عشوائي بدلاً من متغير حقيقي وهو ليس ضمن دراستنا هذه.

2. مخمنات مقترحة لتخمين معلمتي أنموذج العلاقة الدالية عند تكرار المشاهدات

يعرف أنموذج العلاقة الدالية عند تكرار المشاهدات كما في الصيغ التالية:

$$Y_i^* = \alpha + \beta X_i^*;\ X_{ij} = X_i^* + \delta_{ij};\ Y_{ij} = Y_i^* + \epsilon_{ij};\ i = 1,2,...,n;\ j = 1,2...,m_i$$

حيث $X_j^* \neq X_i^*$ لكل $i \neq j$, أي لكل قيمة من قيم $\left(X^*, Y^*\right)$ ولتكن $\left(X_i^*, Y_i^*\right)$, $i = 1,2,...,n$ فان هذا الزوج المرتب يقابله m_i من مشاهدات المتجه العشوائي $\left(X, Y\right)$ وهي $\left(X_{ij}, Y_{ij}\right)$, $j = 1,2,...,m_i$.

يتيح لنا هذا الأنموذج فرز مشاهدات $\left(X, Y\right)$ المتمثلة بالأزواج المرتبة $\left(X_{ij}, Y_{ij}\right)$, $i = 1,2,...,n$, $j = 1,2,...,m_i$ إلى أربع فصائل يمكن تمثيلها بالمجموعات الآتية:

$$A_i = \left\{X_{ij};\ j = 1,2,...,m_i\right\} \quad i = 1,2,...,n$$

$$B_i = \left\{Y_{ij};\ j = 1,2,...,m_i\right\} \quad i = 1,2,...,n$$

$$C_i = \left\{\left(X_{ij}, Y_{ij}\right);\ j = 1,2,...,m_i\right\} \quad i = 1,2,...,n$$

$$D = \left\{\left(X_{ij}, Y_{ij}\right);\ i = 1,2,...,n;\ j = 1,2,...,m_i\right\} = \bigcup_{i=1}^{n} C_i$$

مما تقدم يلاحظ انه في حالة أنموذج العلاقة الدالية عند عدم تكرار المشاهدات لم يتوفر لدينا سوى مجموعة واحدة هي $\left\{\left(X_i, Y_i\right);\ i = 1,2,...,n\right\}$. عليه يتبع بالضرورة إثارة تساؤل مشروع ألا وهو: هل أن تكرار المشاهدات على هذا الشكل يسهم في تعريف مخمن لكل من α و β بحيث يكون أكثر ركانة من سواه؟ إننا نعتقد أن توفر مثل هذه البيانات يتيح لنا طرق أبواب لم تكتسب الاهتمام سابقاً وبالذات في معالجة تأثير الشوارد سواء كانت في مشاهدات المتغير العشوائي X أو المتغير العشوائي Y أو كليهما.

من الواضح أن أنموذج العلاقة الدالية عند تكرار المشاهدات ما هو إلا أنموذج علاقة دالية مما يتبع إمكانية سلوك أي مسلك تقليدي لتعريف مخمن متسق لكل من α و β الأمر الذي لا يتطلب سوى إعادة صياغة مخمن كل أسلوب تضمنه البند الأول بدلالة مجموعة المشاهدات D . وحيث أن بعض الأساليب ينبغي معرفة قيمة λ فإننا نقترح مخمن ذو صفات جيدة لكل من σ_δ^2 و σ_ϵ^2 وبالتالي مخمن متسق للمعلمة λ على التوالي كما في الصيغ الآتية:

$$S_\delta^2 = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(X_{ij} - \bar{X}_{i\cdot}\right)^2 \Big/ (M-n);\ S_\epsilon^2 = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 \Big/ (M-n)$$

$$\hat{\lambda}=S_{\epsilon}^{2}/S_{\delta}^{2};\ \ \bar{X}_{i\cdot}=\sum_{j=1}^{m_i}X_{ij}/m_i;\ \ \bar{Y}_{i\cdot}=\sum_{j=1}^{m_i}\bar{X}_{ij}/m_i;\ \ M=\sum_{i=1}^{n}m_i$$

ولغرض الاستفادة من المجموعات A_i, B_i و C_i وتوظيفهما في حل معضلة أنموذج العلاقة الدالية نقترح أسلوباً مغايراً لما هو معتاد ألا وهو أسلوب المرحلتين. أن هذا المسلك الجديد على الرغم من تشعباته التي يمكن أن ترد فيه إلا انه وبصورة عامة يعتمد الإجراءات الآتية عند كل مرحلة من مرحلتيه وفق ضوابط بحث وولد التي وردت في البند الأول فيما يخص المتغيرات العشوائية δ_{ij} و $\in_{ij}$.

المرحلة الأولى:

لكل $i=1,2,...,n$ يمكن التوصل إلى ما يأتي:

1. عناصر المجموعة A_i تشكل عينة غير مترابطة تتبع المتغير العشوائي X عند الدليل i وليكن X_i حيث $EX_i=X_i^*$ وان $\operatorname{var}X_i=\sigma_\delta^2$. وحيث أن X_i^* قيمة حقيقية مجهولة تشكل والحالة هذه معلمة تموضع Location المتغير العشوائي X_i, عليه يمكن توظيف عناصر المجموعة A_i في الحصول على مخمن مناسب ذو صفات جيدة أن أمكن ذلك (وذلك من خلال أساليب تخمين معلمة التموضع) لتخمين المعلمة X_i^* وليكن $\hat{X}_i^*$.
2. على غرار ما جاء في (1) فان عناصر المجموعة B_i تشكل أيضاً عينة غير مترابطة تتبع المتغير العشوائي Y عند الدليل i وليكن Y_i حيث $EY_i=Y_i^*$ و $\operatorname{var}Y_i=\sigma_\epsilon^2$, وان Y_i^* قيمة حقيقية مجهولة يمكن اعتبارها معلمة مجهولة تشكل أيضاً معلمة تموضع المتغير العشوائي Y_i. وباستخدام عناصر المجموعة B_i يمكن الحصول على مخمن مناسب للمعلمة Y_i^* وليكن $\hat{Y}_i^*$.
3. عناصر المجموعة C_i تشكل عينة غير مترابطة تتبع المتجه العشوائي (X,Y) عند الدليل i وليكن (X_i,Y_i) حيث $E(X_i,Y_i)=(X_i^*,Y_i^*)$ وان X_i و Y_i غير مترابطين. وحيث الزوج المرتب (X_i^*,Y_i^*) قيمة حقيقية مجهولة يمكن عدها معلمة مجهولة تشكل معلمة تموضع المتجه العشوائي (X_i,Y_i). وباستخدام عناصر المجموعة C_i يمكن الحصول على مخمن مناسب لهذه المعلمة وليكن $(\hat{X}_i^*,\hat{Y}_i^*)$.

مما تقدم سواء من خلال (1) و (2) أو من (3) فانه يتم الحصول على مخمن مناسب لكل زوج مرتب حقيقي مجهول (X_i^*,Y_i^*) $i=1,2,...,n$ متمثلة بالمخمنات $(\hat{X}_i^*,\hat{Y}_i^*)$ $i=1,2,...,n$, وبهذا تتم المرحلة الأولى على وفق هذا الأسلوب. إننا بهذه المرحلة نقلص عدد المشاهدات من M إلى n من المشاهدات حيث كل من m_i من المشاهدات تنتهي بمشاهدة واحدة تتمثل بمشاهدة قيمة المخمن $(\hat{X}_i^*,\hat{Y}_i^*)$. إننا نعتقد بهذا العمل وبمساعدة أسلوب تخمين مناسب سوف ننتهي بمشاهدات تكون قريبة (قدر الإمكان) من القيم الحقيقية المجهولة للأزواج المرتبة (X_i^*,Y_i^*) $i=1,2,...,n$ الأمر الذي يتيح لنا استخدام هذه المشاهدات بدلا من المشاهدات الرئيسية في تخمين كل من α و β وذلك كما سيرد في المرحلة الثانية.

من أساليب التخمين الممكنة لمعلمة التموضع (X_i^*,Y_i^*) $i=1,2,...,n$ وذلك من خلال مجموعة المشاهدات A_i, B_i أو C_i وعلى افتراض توفر شرط التوزع الطبيعي لكل من δ و $\in$ مخمن أسلوب الأرجحية العظمى وهو:

$$\left(\hat{X}_i^*,\hat{Y}_i^*\right)=\left(\bar{X}_{i\cdot},\bar{Y}_{i\cdot}\right);\ \ i=1,2,...,n$$

والذي يتسم بالصفاة غير منحاز وذو اقل تباين إضافة إلى الاتساق. إما عند غياب شرط التوزع الطبيعي وفي وجود الشوارد يمكن تطبيق أساليب حصينة عديدة والتي جميعها تتمتع بميزة مشتركة وهي استخدام أسلوب الموازنة بين المشاهدات للتخفيف من تأثير الشوارد (خصوصاً إذا كانت ضمن المشاهدات العزومية Leverage حيث تأثيرها يكون اشد) الأمر الذي يقلل من تأثر المخمنات بها. وذلك من خلال أقران أوزان اقل مع المشاهدات التي يعتقد بشروديتها من تلك التي تقرن مع بقية المشاهدات.

هنالك العديد من المخمنات التي تنطوي تحت هذا الأسلوب, فعند استخدام المجموعة A_i وكذلك المجموعة B_i فإننا نسقط عناصر المجموعة C_i التي تنتمي إلى المستوى الديكارتي R^2 على محور السينات بالنسبة لعناصر A_i كذلك محور الصادات بالنسبة لعناصر B_i, أي تحويل مشاهدات الأنموذج عند تكرار المشاهدات إلى مشاهدات تنتمي إلى R بدلاً من R^2 مما يمهد الطريق أمامنا إلى توظيف الأساليب الحصينة التقليدية المصنفة إلى عدة طرائق منها أسلوب M,L التي يتاح لنا بها إيجاد مخمن جيد لكل من X_i^* و Y_i^* $i=1,2,...,n$. فعند تطبيق أسلوب -L على المجموعة A_i والمجموعة B_i فان مخمنات -L تعرف بأنهما التركيبتان الخطيتان الآتيتان على التوالي:

$$\hat{X}_i^*=\sum_{j=1}^{m_i}a_jX_{i(j)};\hat{Y}_i^*=\sum_{j=1}^{m_i}a_jY_{i(j)};i=1,2,...,n$$

حيث يمثل كل من $X_{i(j)}$ و $Y_{i(j)}$ إحصاءات مرتبة لعناصر A_i و B_i, وان a_j أوزان بحيث $\sum_{j=1}^{m_i} a_j = 1$ وان $a_j \geq 0$. وبتغير هذه الأوزان يمكن تكوين صيغ مخمنات مختلفة, ومن المفضل اختيارها بحيث تتناسب عكسياً مع مقدار بعد المشاهدة عن مركز المشاهدات – أي ينبغي لها أن تتناقص باقتراب j من 1 أو من m_i – وهذا يجعل الشوارد أن وجدت ضمن المشاهدات المتطرفة اقل وزنا من سواها. هنالك العديد من هذه المخمنات منها الوسيط الذي هو عبارة عن الإحصاءات الوسيطة لكل من الإحصاءات المرتبة $X_{i(j)}$ و $Y_{i(j)}$ في حالة m_i عدد فردي أو معدل الإحصائتين الوسيطتين عندما m_i عدد زوجي. وهنالك مخمن آخر يدعى الوسط الحسابي المشذب وهو عبارة عن الوسط الحسابي بعد تشذيب عدد من الإحصاءات المرتبة من كلا الطرفين, في حين مخمن ونسر يستبدل قيم المشاهدات المشذبة بقيمة اقرب مشاهدة.

أما عند تطبيق أسلوب M- على المجموعة A_i فان قيمة مخمن M- للمعلمة $X_i^*; i=1,2,...,n$ هي قيمة المخمن $\hat{X}_i^*$ التي تحقق المعادلة:

$$\sum_{j=1}^{m_i}\Psi\left[\left(X_{ij}-\hat{X}_i^*\right)/S_{Xi}\right]=0;\quad S_{Xi}=\underset{j}{\mathrm{med}}\left\{\left|X_{ij}-\underset{j}{\mathrm{med}}\{X_{ij}\}\right|\right\}/0.6745$$

في حين عند تطبيق أسلوب M- على المجموعة B_i فان قيمة مخمن M- للمعلمة $Y_i^*; i=1,2,...,n$ هي قيمة المخمن $\hat{Y}_i^*$ التي تحقق المعادلة:

$$\sum_{j=1}^{m_i}\Psi\left[\left(Y_{ij}-\hat{Y}_i^*\right)/S_{Yi}\right]=0;\quad S_{Yi}=\underset{j}{\mathrm{med}}\left\{\left|Y_{ij}-\underset{j}{\mathrm{med}}\{Y_{ij}\}\right|\right\}/0.6745$$

بانتقاء دالة Ψ المناسبة وإحدى الطرائق العددية يمكن إيجاد القيمة التخمينية لكل من المعلمة $X_i^*, Y_i^*; i=1,2,...,n$ وإننا هنا نقترح دالة Ψ جديدة المعرفة كما في الصيغة التالية:

$$\Psi(u)=\begin{cases}-a\exp-\left[(u+a)/b\right]^c & ;u<-a\\ u & ;|u|\le a\\ a\exp-\left[(u+a)/b\right]^c & ;u>a\end{cases}$$

حيث أن a,b,c ثوابت التعديل تأخذ القيم $a=1.5$, 1.8 أو 2.1 , $b=0.75$ أو 1, $c=2$. يلاحظ انه عندما $|u|\le a$ فان هذا الجزء من الدالة هو دالة الأرجحية العظمى حيث الأوزان متساوية لكافة المشاهدات, ثم تأخذ الأوزان بالتناقص تدريجياً عندما $u<-a$ أو $u>a$. كما أن الدالة المقترحة تتمتع بمرونة التحكم بتناقص هذه الأوزان من خلال قيم b و c.

الملاحظ أن الأسلوبين L و M يمران بخطوتين لتخمين X_i^* و Y_i^* حيث يتم تطبيق كل منهما على المجموعة A_i ومن ثم على المجموعة B_i. يقودنا هذا إلى التفكير في أسلوب آخر يقوم بتخمين (X_i^*, Y_i^*) بخطوة واحدة; أي بمعنى آخر استخدام المجموعة C_i, وهي فكرة أسلوبنا المقترح الآخر والذي سوف يدعى بأسلوب القشرة المحدبة. وهو محاولة لتعميم أسلوب L- المبني على ترتيب المشاهدات بصورة تصاعدية. أن المشكلة في الأنموذج الحالي هي انتشار المشاهدات في المستوي حيث لا معنى هنا لوحدانية الترتيب مما يجعل الأمور أصعب بكثير من انتشارها على الخط المستقيم, لذا سوف نعتمد مجموعات القشرة المحدبة كأسلوب لترتيب المشاهدات. فإذا كان عدد عناصر C_i منته فان مجموعة القشرة المحدبة هي ببساطة اصغر مضلع مغلق رؤوسه عبارة عن نقاط تنتمي إلى C_i والتي تمثل النقاط المتطرفة في المشاهدات.

أن أسلوبنا الجديد هذا يعتمد على تقشير (حذف) نقاط القشرة المحدبة (لان الشوارد العزومية أن وجدت تكون ضمن هذه النقاط) وتخمين (X_i^*, Y_i^*) من المجموعة المتبقية من المشاهدات وذلك بأخذ المعدل لهذه المشاهدات, أي أن مخمن هذا الأسلوب يشابه مخمن الوسط الحسابي المشذب في أسلوب L-. أن أسلوب التقشير هذا يمكن أن يأخذ الصورة التكرارية وذلك عند التقشير لأكثر من مرة والذي حينها يدعى بأسلوب التقشير المتعدد.

أن ما يؤخذ على أسلوب القشرة المحدبة هو حذفه للعديد من المشاهدات التي لا تعد شوارد, لذا ولتجنب عملية الحذف نقترح أسلوب القشرة المحدبة المطوية. تتلخص فكرة هذا الأسلوب بإنشاء مثلثات متتالية عددها مساو لعدد نقاط القشرة المحدبة ومن ثم اخذ مركز الثقل لهذه المثلثات واستبدال نقاط القشرة المحدبة بمراكز الثقل هذه. أن هذا الأسلوب يقوم بعملية تقريب المشاهدات من بعضها بدلاً من حذفها ومن ثم تخمين (X_i^*, Y_i^*) بأخذ المعدل لهذه المشاهدات. رياضياتياً يمكن صياغة هذا الأسلوب كالآتي هب أن:

$$Q_i=\left\{(X_{ik},Y_{ik});k=1,2,...,p;3\le p<m_i\right\}$$

تمثل مجموعة نقاط القشرة المحدبة للمجموعة C_i مرتبة حول أو عكس القشرة. فان مراكز ثقل المثلثات مبتدئين عند أي نقطة من نقاط Q_i وعندما $p\ge 4$ هي:

$$\left(\tilde{X}_{ij},\tilde{Y}_{ij}\right)=\left(\frac{X_{ij}+X_{ij+1}+X_{ij+2}}{3},\frac{Y_{ij}+Y_{ij+1}+Y_{ij+2}}{3}\right);\ j=1,2,...,p-2$$

$$\left(\tilde{X}_{ip-1},\tilde{Y}_{ip-1}\right)=\left(\frac{X_{ip-1}+X_{ip}+X_{i1}}{3},\frac{Y_{ip-1}+Y_{ip}+Y_{i1}}{3}\right)$$

$$\left(\tilde{X}_{ip},\tilde{Y}_{ip}\right)=\left(\frac{X_{ip}+X_{i1}+X_{i2}}{3},\frac{Y_{ip}+Y_{i1}+Y_{i2}}{3}\right)$$

وعندما $p=3$ تضاف نقطة وهمية رابعة ولتكن $\left(X_{iu},Y_{iu}\right)$ وهي عبارة عن مركز ثقل المثلث الذي رؤوسه النقاط الثلاث أي أن:

$$\left(X_{iu},Y_{iu}\right)=\left(\frac{X_{i1}+X_{i2}+X_{i3}}{3},\frac{Y_{i1}+Y_{i2}+Y_{i3}}{3}\right)$$

ومن ثم تكوين ثلاثة مثلثات كل منها تشترك هذه النقطة الوهمية مع نقطتين في تكوينه وبعد ذلك تحسب مراكز الثقل لهذه المثلثات. وعلى افتراض أن Q_i' تمثل مراكز ثقل المثلثات فان مخمن هذا الأسلوب هو:

$$\left(\hat{X}_i^*,\hat{Y}_i^*\right)=\left(\left[\sum_{C_i-Q_i}X_{ij}+\sum_{Q_i'}\tilde{X}_{ij}\right]\Big/m_i,\left[\sum_{C_i-Q_i}Y_{ij}+\sum_{Q_i'}\tilde{Y}_{ij}\right]\Big/m_i\right)$$

كذلك فان هذا الأسلوب يمكن تكراره للحصول على تقارب أكثر للمشاهدات.

أما المرحلة الثانية أصبح لدينا مجموعة مشاهدات المتجهات العشوائية $\left(\hat{X}_i^*,\hat{Y}_i^*\right)$, $i=1,2,...,n$. دع:

$$\zeta_i=\hat{X}_i^*-X_i^*,\ \eta_i=\hat{Y}_i^*-Y_i^*;\ i=1,2,...,n$$

فانه يصبح لدينا الأنموذج:

$$Y_i^*=\alpha+\beta X_i^*;\ \hat{X}_i^*=X_i^*+\zeta_i;\ \hat{Y}_i^*=Y_i^*+\eta_i;\ i=1,2,...,n$$

وهو أنموذج علاقة دالية يحمل معلمتي الأنموذج القديم نفسهما إلا أن δ_{ij} و $\in_{ij}$ حل محلهما المتغيران العشوائيان ζ_i و η_i على التوالي, كذلك X_{ij} و Y_{ij} حل محلهما $\hat{X}_i^*$ و $\hat{Y}_i^*$ على التوالي. إن هذا الأنموذج الجديد ليس مستقلاً عن الأنموذج الرئيسي حتماً حيث يعتمد تباين ζ_i وكذلك تباين η_i على تباين $\hat{X}_i^*$ وتباين $\hat{Y}_i^*$ وبدورهما يعتمدان حتماً على تباين δ_{ij} وتباين $\in_{ij}$. أي إننا بصدد أنموذج علاقة دالية غير محدد القيود ما لم تحدد صيغة المخمن $\hat{X}_i^*$ والمخمن $\hat{Y}_i^*$.

عند استخدام أسلوب L- أو أسلوب M- في المرحلة الأولى فانه لا مسلك لنا سوى استخدام أسلوب وولد أو أسلوب ثايل التام أو غير التام في هذه المرحلة لتخمين α و β , وذلك لصعوبة تحديد صيغة تباين كل من ζ_i و η_i وبالتالي إيجاد مخمن مناسب للمعلمة λ وبالذات عند استخدام أسلوب M-.

أما عند استخدام أسلوب الأرجحية العظمى في المرحلة الأولى فينبغي التحيط هنا وذلك لكون الأنموذج بات في هذه المرحلة:

$$Y_i^*=\alpha+\beta X_i^*;\ \bar{X}_{i\cdot}=X_i^*+\bar{\delta}_{i\cdot};\ \bar{Y}_{i\cdot}=Y_i^*+\bar{\in}_{i\cdot};\ i=1,2,...,n$$

وان:

$$\operatorname{var}\bar{\delta}_{i\cdot}=\frac{\sigma_\delta^2}{m_i};\ \ \operatorname{var}\bar{\in}_{i\cdot}=\frac{\sigma_\in^2}{m_i};\ \ i=1,2,...,n$$

أي تظهر هنا الحالة اللاتجانسية. وللتغلب على هذه المعضلة ليس أمامنا إلا حالة واحدة فقط وهي افتراض تساوي قيم m_i ; $i=1,2,...,n$ لتطبيق أسلوب الأرجحية العظمى كمرحلة ثانية لتخمين α و β حيث يمكن استخدام المخمن $\hat{\lambda}$ المعرف سابقاً عند عدم معرفة قيمة المعلمة λ مسبقاً وذلك لكون:

$$\lambda=\frac{\operatorname{var}\bar{\in}_{i\cdot}}{\operatorname{var}\bar{\delta}_{i\cdot}}=\frac{\sigma_\in^2}{\sigma_\delta^2}$$

ولكن ذلك يغدو دون جدوى إذ يمكن البرهنة وببساطة أن مخمني أسلوب الأرجحية العظمى للمعلمتين α و β هذا هما نفس مخمني أسلوب الأرجحية العظمى عند استخدامه كمرحلة واحدة.

كذلك يمكن البرهنة عند استخدام الأرجحية العظمى في المرحلة الأولى وعند تساوي قيم m_i يتبعه أسلوب وولد في المرحلة الثانية لتخمين المعلمة β هو مخمن أسلوب وولد ذو المرحلة الواحدة لهذه المعلمة نفسها.

وبالرغم من المعوقات أعلاه فان الطريق ما زال مفتوحاً أمامنا عند استخدام أسلوب الأرجحية العظمة في المرحلة الأولى وتساوي قيم m_i وذلك باستخدام أسلوب W- في المرحلة الثانية حينما تكون قيمة λ معلومة مسبقاً أو تخمينها من خلال المخمن $\hat{\lambda}$. وفي كل الأحوال عند تساوي أو عدم تساوي قيم m_i يمكن استخدام أسلوب ثايل التام أو غير التام في المرحلة الثانية لتخمين كل من α و β عندما يستخدم أسلوب الأرجحية العظمى في المرحلة الأولى.

وقبل أن ننهي هذا البند لا بد أن نشير إلى أن تعميم أسلوب التخمين ذي المرحلتين ممكن وببساطة لتخمين معالم أنموذج العلاقة الدالية المتعددة, وذلك من خلال جعل الأسلوب متعدد المراحل.

دراسة مقارنة باستخدام أسلوب المحاكاة

هب أن $Y_i^*=1+2X_i^*$ (أي القيمة الحقيقية لكل من α و β هي 1, 2 على التوالي) وان:

$$X_{ij}=X_i^*+\delta_{ij};\ Y_{ij}=Y_i^*+\in_{ij};\ i=1,2,...,n;\ j=1,2,...,m$$

و $\delta_{nm},...,\delta_{12},\delta_{11}$ عينة عشوائية تتبع المتغير العشوائي δ الذي يتوزع أما $N(0,1)$ باحتمال p_1 وأما $N(0,k_1^2)$ باحتمال $(1-p_1)$ مستقلة عن العينة العشوائية $\in_{nm},...,\in_{12},\in_{11}$ التي تتبع المتغير العشوائي $\in$ الذي يتوزع أما $N(0,1)$ باحتمال p_2 وأما $N(0,k_2^2)$ باحتمال $(1-p_2)$ حيث $0\le p_1,p_2\le 1$ و $1\le k_2,k_1$.

يتضح من تعريف أنموذج العلاقة الدالية هذا أن δ تتوزع توزعاً خليطاً من توزعين طبيعيين يشتركان بالمعدل نفسه إلا أنهما يختلفان في التباين إذ يسهم الأول بنسبة p_1 بينما يسهم الثاني بنسبة $(1-p_1)$, وحيث أن تباين الثاني اكبر من الأول بمقدار k_1^2 من المرات فانه من المتوقع الحصول على مشاهدة باحتمال $(1-p_1)$ من التوزع الثاني التي قد تكون شاردة بالمقارنة مع مشاهدات التوزع الأول. كما ويلاحظ أن الحالة $p_1=1$ أو $k_1=1$ تعني خلو $\delta_{nm},...,\delta_{12},\delta_{11}$ من الشوارد. والكلام ذاته يمكن أن يقال عن المتغير العشوائي $\in$.

تم توليد 200 عينة كل منها بحجم nm من المشاهدات:

$$\left(X_{ij},Y_{ij}\right)=\left(X_i^*+\delta_{ij},1+2X_i^*+\in_{ij}^*\right);\ i=1,2,...,n;\ j=1,2,...,m$$

حيث $X_i^*=i;\ 1=1,2,...,n$ (أي أن قيم X_i^* متساوية الأبعاد) وان قيم δ_{ij} و $\in_{ij}$ قد تم توليدها في الحاسوب بإتباع أسلوب بوكس – مولر.

لكل قيمة منتخبة من قيم p_2,p_1,k_2,k_1,m,n تم تخمين المعلمة β وذلك باستخدام أساليب التخمين: (ML) الأرجحية العظمى ذو المرحلة الواحدة و λ مخمنة, (ML-W) الأرجحية العظمى في المرحلة الأولى يتبعه أسلوب -W في المرحلة الثانية, (L-Th) أسلوب -L (الوسيط) في المرحلة الأولى يتبعه أسلوب ثايل التام و (M-Th) أسلوب -M في المرحلة الأولى حيث الدالة Ψ هي الدالة المقترحة يتبعه أسلوب ثايل التام في المرحلة الثانية. تكرر هذه العملية مائتي مرة للحصول على مائتي قيمة للمعلمة β.

وللتعرف على مدى كفاءة المخمنات سوف نعتمد مؤشر مجموع مربعات الخطأ لهذه المخمنات أي:

$$SSE\left(\hat{\beta}\right)=\sum_{i=1}^{200}\left(\hat{\beta}_i-2\right)^2$$

إن قيم هذا المؤشر تم تبويبها في الجدول التالي والذي قسم أفقياً إلى أربعة أقسام وذلك حسب عدم وجود شوارد, وجود شوارد في قيم X_{ij} فقط, وجود شوارد في قيم Y_{ij} فقط ووجود شوارد في قيم كل من X_{ij} و Y_{ij}. وفيما يأتي النتائج التي يمكن قراءتها عند كل قسم من هذا الجدول.

قيم $SSE\left(\hat{\beta}\right)$ لمخمنات المعلمة β

n	m	k_1	k_2	p_1	p_2		
						ML	ML-W
10	5	1	1	1	1	7.82	4.43
10	15	1	1	1	1	7.20	3.85
10	5	4	1	0.9	1	8.56	5.31
10	5	4	1	0.8	1	9.79	6.66
10	5	9	1	0.9	1	12.70	9.08
10	5	9	1	0.8	1	19.33	17.49
10	15	4	1	0.9	1	7.50	4.11
10	15	4	1	0.8	1	7.99	4.53
10	15	9	1	0.9	1	8.97	5.17
10	15	9	1	0.8	1	11.22	7.24
10	5	1	4	1	0.9	8.83	5.84
10	5	1	4	1	0.8	9.28	6.23
10	5	1	9	1	0.9	9.47	6.10
10	5	1	9	1	0.8	10.24	7.58
10	15	1	4	1	0.9	7.14	4.62
10	15	1	4	1	0.8	7.42	5.12

عند عدم وجود شوارد (القسم الأول):

1. جميع المخمنات ذات أداء حسن ولكن يلاحظ أن مخمنات الأساليب ذات المرحلتين هي أفضل من أداء أسلوب ML, وان أسلوب (M-Th) هو الأرجح أداءاً.
2. كفاءة جميع المخمنات تزداد بازدياد حجم العينة.

عند وجود شوارد في قيم X_{ij} فقط (القسم الثاني):

1. تأثر جميع المخمنات وان أداء أساليب المرحلتين لازال أكفأ من أسلوب ML ذو المرحلة الواحدة.
2. تأثر أسلوب (ML-W) أكثر من أسلوبي (L-Th) و (M-Th) وذلك لكون هذا الأسلوب يستخدم المعدل في المرحلة

الأولى منه الذي يكون شديد التأثر بالشوارد, والغلبة دائماً لأسلوب (M-Th).

3. يزداد تأثير المخمنات عند ازدياد عدد الشوارد $p_1 = 0.8$ أو ازدياد قيمة k_1, في حين يتحسن أداء جميع المخمنات عند ازدياد حجم العينة.

عند وجود شوارد في قيم Y_{ij} (القسم الثالث):

1. لا يزال أساليب المرحلتين أكفأ من أسلوب ML.
2. تأثر المخمنات اقل وذلك مقارنة مع الحالة السابقة, وان هذا التأثر يزداد عند زيادة عدد الشوارد $p_2 = 0.8$ أو ازدياد قيمة k_2.
3. أساليب المرحلتين دائماً تغلب أسلوب ML والأفضلية المطلقة لأسلوب (M-Th). كما ويتحسن أداء جميع المخمنات بازدياد حجم العينة.

عند وجود شوارد في قيم كل من X_{ij} و Y_{ij} (القسم الأخير):

1. تأثر جميع المخمنات وبخاصة أسلوب ML والذي يعزى إلى أن مخمنات أساليب المرحلتين حصينة. وان هذا التأثر يزداد بازدياد عدد الشوارد $p_1 = p_2 = 0.8$ أو عند ازدياد قيم K_1 أو k_2.
2. تفوق أداء أسلوب (L-Th) و (M-Th) وتتحسن كثيراً خصوصاً عند ازدياد حجم العينة.

قبل أن نختم هذا البند لابد أن نشير هنا إلى إننا لم نتمكن من برمجة الجزء الخاص بالمرحلة الأولى وفيما يخص أساليب القشرة المحدبة, لذا سوف نترك هذا الشأن كمسألة مفتوحة للدارسين مع الاعتقاد بأنه أسلوب واحد.

المصادر

1. Wald, A. **1940**: *The Fitting Of Straight Lines If Both Variables Are Subject To Errors,* Ann. Math. Statistics Vol. 11, 284-300
2. Bartlett, M. S. **1949**: *Fitting A Straight Line When Both Variables Are Subject To Error,* Biometrics, Vol. 5, 207-212.
3. Thiel, H. **1950**: *A Rank Invariant Method Of Linear And Polynomial Regression,* Proc. Kon. Nederi. A. Kad Wetensch, 53, 386-392.
4. Madansky, A. **1959**: *The Fitting Of Straight Line When Both Variables Are Subject To Error,* JASA, 173-205.
5. Graybill, F. A. (1961): *An Introduction To Linear Statistical Model*, McGraw-Hill book company.
6. Sprent, P. A. **1966**: *A Generalized Least Squares Approach To Linear Functional Relationships,* J. Roy. Statist. Soc, Ser B, 28, 278-297.
7. Sprent, P. A. **1969**: *Models In Regression And Related Topics,* London, Methuen.
8. Moran, P. A. **1970**: *Estimating Structural And Functional Relationship,* J. Multivariate Anal. 1, 232-255.
9. Kendall, M. G. and Stuart, A. **1979**: *The Advanced Theory Of Statistics,* Vol. II, London Charles Griffith Co.
10. Fuller, W. A. **1987**: *Measurement Error Models,* Wiley, New York.
11. Mokhtar, B. A.**1989**: *On W-Estimators Of A Linear Functional Relationship,* Commun. Statist-theory Math., Vol. 18, 287-314.
12. Andrew, et al ,**1972**: *Robust Estimates Of Location,* princetion university press, princention.
13. Linder, E. and Babo, G. J. **1994**: *Bootstrapping The Linear Functional Relationship With Known Error Variance Ratio,* Scandinavian J. of Stat. Vol. 21, 21-39.
14. Ameed, K. S. **2003**: *Estimation Of The Parameters Of Simple Linear Functional Model With Replication Of Observations,* Ph. D. thesis, Dept of Math., college of science, University Of Baghdad.