مقارنة بين طريقتي (GM) الحصينة المعممة (Generalized M) وطريقة (S) الحصينة لتقدير انموذج الانحدار للبيانات ذات التوزيع الطبيعي الملتوي

خديجة عبد الكريم هادي أ.م.د رواء صالح محمد

الجامعة المستنصرية/ كلية الادارة والاقتصاد

Khadigaabdalkareemhade@gmail.com

rshnss٦٩@yahoo.com

المستخلص:-

تم في هذا البحث دراسة طريقتين حصينتين لتقدير معالم انموذج الانحدار الخطي وعليه تطرقنا الى الطرق الحصينة لبيان اثر وجود القيم الشاذة في قيم المشاهدات وحددت المشكلة وجود القيم الشاذة في البيانات تؤثر في توزيع البواقي في انموذج الانحدار لذلك سيصبح توزيعا ملتويا حيث تم سحب عينة مكونة من (٢٤٣) مريضا مصابا بأمراض الدم متمثلا بالمتغير المعتمد الهيموجلوبين(HGB) والمتغيرات المستقلة (RBC,HCT,MCV,MCHC,MCH) بأستعمال معيار متوسط مربعات الخطأ (Mse) وذلك من خلال البرنامج الاحصائي Matlab ٢٠١٥ وتوصلنا الى افضل طريقة للتقدير كانت طريقة (GM) الحصينة المعممة.

المصطلحات الرئيسية للبحث : مفهوم الانحدار الحصين ، القيم الشاذة ، الكشف عن القيم الشاذة ، التوزيعات الملتوية ، طرق التقدير الحصينة ، معيار المقارنة

ABSTRACT

In this research, two strong methods were studied for estimating the parameters of the linear regression model, and accordingly we touched on the strong methods to show the effect of the presence of anomalies in the values of observations. ٢٤٣) patients with blood diseases represented by the dependent variable hemoglobin (HGB) and the independent variables (RBC, HCT, MCV, MCHC, MCH) using the mean squared error (Mse) criterion through the statistical program Matlab ٢٠١٥ and we reached the best method of estimation was the GM method the generalized fortress

١ – المقدمة :–

ان اسلوب بناء نماذج الانحدار الخطي من المواضيع المهمة في التحليل الاحصائي ، عندما يرتبط اثنان او اكثر من المتغيرات التوضيحية في علاقة الانحدار تحدث مجموعة من المشاكل والتي يكون فيها خرق لأحد الفروض الأساسية مما يؤدي الى الحصول على تقديرات متحيزة او اقل دقة وهنالك عده طرق اقترحت لمعالجة هذه المشاكل نذكر منها الطرائق الحصينة في تحليل الانحدار.

*بحث مستل من رسالة ماجستير

في عام (١٩٦٤) قدم Huber أول نظرية عامة في الحصانة لذلك قام بتقديم فصل من المقدرات المرنة وقدم Humple

(The Infinitesimal Approach) ووضع نظرية اخرى في عام (١٩٧١) وتعتمد هذه النضرية على ثلاث مفاهيم

مركزية هي (دالة التأثير ونقطة الانهيار والحصانة النوعية) اذا تشير دالة التأثير الى الحصانة النوعية عند نقطة معينة .

٢- مفهوم الانحدار الحصين[٤](The concept of Robust Regression) :-

هو طريقة بديلة لتقدير معالم انموذج الانحدار تستخدم في حالة و جود القيم الشاذة في البيانات لان تقدير معالم انموذج الانحدار الاعتيادي في ظل وجود هذه الشواذ تكون غير كفؤه لأنه تحدث حالة عدم تطابق بين بيانات موضوع الدراسة و الفروض الاساسية الواجب توافرها في الانموذج وبذلك تفقد الطرائق التقليدية مثل طريقة المربعات الصغرى والامكان الاعظم خصائصها الجيدة لتقدير معالم انموذج الانحدار المدروس حيث أن هذه المشكلة تستوجب إيجاد طرائق بديلة للتقدير لذا تم إيجاد طرائق إحصائية بديلة تكون مقاومة لمعالجة هذه المشكلة وهي ما تسمى بالطرائق الحصينة اما المقدرات الناتجة عن هذه الطريقة البديلة فتسمى بالمقدرات الحصينة وتكون غير حساسة تجاه الشواذ أضافة الى إنها تعطي مقدرات كفؤة تعادل كفاءة طرائق التقدير الاعتيادية مثل طريقه المربعات الصغرى(OLS) والامكان الأعظم (ML) في حالة عدم وجود الشواذ و تعطي مقدرات اكثر كفاءة في حالة وجودها في البيانات.

٣-القيم الشاذ[٩] [٧](Outlier):-

هي المشاهدات التي تكون غير منطقية عندما تتم مقارنتها ببقية بيانات العينة أو هي مشاهدات العينة التي تكون كبيرة جدا أو صغيرة جدا بالمقارنة مع بقية مشاهدات العينة و أن وجودها يؤثر بصورة سلبية على كفاءة الطرائق الإحصائية المستخدمة ويجب الكشف عنها للحد من تأثيرها، وقد وضح (١٩٨٧) (Rousseeuw) بصورة بيانية عن كيفية تأثير القيمة الشاذة في مقدرات طريقة المربعات الصغرى وكيف أن هذه القيمة تغير من اتجاه خط الانحدار وقد عرف العديد من العلماء والباحثين القيم الشاذة منهم:-

(*Bross*) المشاهدات الشاذة كالاتي " تلك المشاهدة التي تظهر منحرفة بشكل كبير عن سائر مكونات العينة التي وجدت فيها تلك المشاهدة".

ومما سبق يمكن القول بأن القيمة الشاذة هي المشاهدة التي تقع بعيداً عن معظم مشاهدات العينة والتي لا تتطابق أو لا تنسجم مع بقية المشاهدات الموجودة في العينة تحت الدراسة.

وللقيم الشاذة عدة أنواع ولكن ما يُثير اهتمامنا هو نوعين من الشواذ ولا توجد تسميات أو مصطلحات محددة لتسميها ولكن هناك تسميات شائعة وهي

١.
واذ، أخطاء شاذة (*Outliers error*)، شواذ في المحور *Y*، شواذ الانحدار.

٢.
واذ في المحور *X*، نقاط جذب، نقاط التأثير، نقاط الرفع، نقاط الانعطاف.

٤-الكشف عن القيم الشاذة:-

يتم الكشف عن القيم الشاذة بواسطة طريقة الرسم الصندوقي .

٤-١الرسم الصندوقي[١١] [١] (Box plot):-

الرسم الصندوقي (Box plot) او في بعض الأحيان تسمى طريقة الصندوق والقطع المخططة مع الملخصات الخمسة (The Box Plots with ٥ Number summaries Whisker and) المقترحة من قبل Tukey عام ١٩٧٧ وهي من الطرائق (اللامعلمية) التي لا تفترض ان يكون هنالك توزيع معين على البيانات خطية بذلك احد الافتراضات الاساسية التي تعتمد عليها الطرائق الاخرى (المعلمية) وتعد هذه الطريقة من افضل الطرائق للتحري والتعامل مع القيم الشاذة في حالة المتغير الواحد وبافتراض عدم معرفة التوزيع الاحتمالي للمتغير ،ان هذه الطريقة تعتمد على المدى الربيعي (Quartile Rang) في إيجاد القيم الشاذة حيث ان الربيعيات يمكن ان تستخدم لتكوين مقايس اخر للتباينات، والمعتمدة على مدى الربيعي الثاني او نصف المدى الربيعي أي (Q٢) والمعرفة بالشكل الاتي

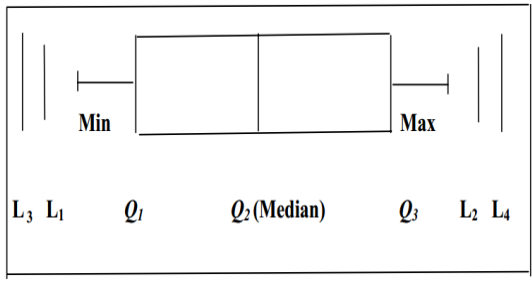
$$IQR = H = Q^{٣} - Q^{١} \qquad \dots(١)$$

والذي يقيس ٥٠٪من متوسط انحرافات للمشاهدات. حيث تشير القيم الكبيرة من هذا المقياس على أن الربيعي الاول والثالث لهما تشتت كبير والذي بدوره يشير الى ان المشاهدات تمتلك تباين عالي،

وهذه الربيعيات تستخدم في بوكس–ويسكر–بلوت للكشف عن المشاهدات الشاذة والمعتمدة على خمس احصائيات وهي:

١.
لحد الادنى للمشاهدات.
٢.
الحد الاعلى للمشاهدات.
٣.
لربيعي الأول.
٤.
لربيع الثاني
٥.
لربيع الثالث.

ويكون توزيعه على الشكل الخاص بهذهٍ الاحصائية كما يلي- : الخطوط الثلاثة العمودية منها تشير الى الربيعيات الثلاثة والخطوط المددة من جهته اليمنى واليسرى تسميان بالشوارب (Whiskers) بحيث اي قيمة تقع خارج هذا الشوارب تعتبر قيمة شاذة. كما هو موضح في الشكل رقم (١) التالي :



الشكل رقم (١) يوضح الشوارب واي قيمة تقع خارجه تعتبر شاذة

تمتد الشوارب الى الخارج حتى تصل الى اقل من ١.٥مرات من H او الى اقصى حد من المشاهدات المتطرفة. بمعنى اخر، إذا وقع اي مشاهده خارج مدى $(Q_1 - ١.٥\ IQR)$ و $(Q_3 + ١.٥\ IQR)$ في هذه الحالة تعتبر القيم من نوع الشواذ الضمنية اما إذا وقع خارج حدود $(Q_1 - ٣ * \mathrm{IQR})$ و $(Q_3 + ٣ * \mathrm{IQR})$ في هذه الحالة يعتبر وجود هذه المشاهدة الشاذة مشكلة في

البيانات. وايضا يمكن ان تستخدم كمقياس لمعرفة خاصية التوزيع الطبيعي للأخطاء العشوائية وذلك عن طريق ملاحظة نوع واتجاه الالتواء للشوارب (Whisker).

٥-التوزيعات الملتوية[٨][٦] :-

قام الباحثين Azzalini عام (١٩٨٥) باقتراح صنف جديد من التوزيعات، والتي تشترك بخصائص مماثلة مع التوزيع الطبيعي وهو حالة خاصة من التوزيع الطبيعي وأطلق عليه التوزيع الطبيعي الملتوي (SND) وعندما تكون الأخطاء العشوائية تمثل حد الخطأ الطبيعي الملتوي (Skew Normal Errors). وفي حالة معلمة الالتواء تساوي صفر يرجع التوزيع الطبيعي الملتوي (SND) الى أساسه متمثل بالتوزيع الطبيعي Normal Distribution .

٥-١التوزيع الطبيعي الملتوي[٢][٨][١٠] (Skew Normal Distribution):-

ان دالة التوزيع الطبيعي الملتوي تكون على نوعين الأول هو في حالة الالتواء الموجب او من الجهة اليمنى(RSND) Right Skew Normal Distribution والثاني هو في حالة الالتواء السالب او من الجهة اليسرى(LSND) Left Skew Normal Distribution وبين (Mudholkar و Huston) التوزيع(SN) بصيغة مختلفة عما جاء به Azzalini وأطلقوا عليها اسم التوزيع الطبيعي الملتوي بمعلمة الالتواء ايبسلون Epsilon Skew Normal (ESN) والذي تكون دالة الكثافة الاحتمالية له بالشكل التالي

$$f(x)=\begin{cases}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2(1+\varepsilon)^2}} & if\ x<\mu\\ \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2(1-\varepsilon)^2}} & if\ x\geq\mu\end{cases} \quad ,-1<\varepsilon<1 \quad ...(2)$$

حيث ان كل من σ,μ , ε تمثل معلمة الموقع (*location*)، القياس(*scale*)، الالتواء (*skewness*) على التوالي وان متوسط وتباين هذا التوزيع هما على النحو التالي

$$E(x) = \mu - \frac{٤\varepsilon\sigma}{\sqrt{٢\pi}} \qquad \dots (٣)$$

$$v(x) = \frac{\sigma^{٢}[(٣\pi - ٨)\varepsilon^{٢} + \pi]}{\pi} \qquad \dots (٤)$$

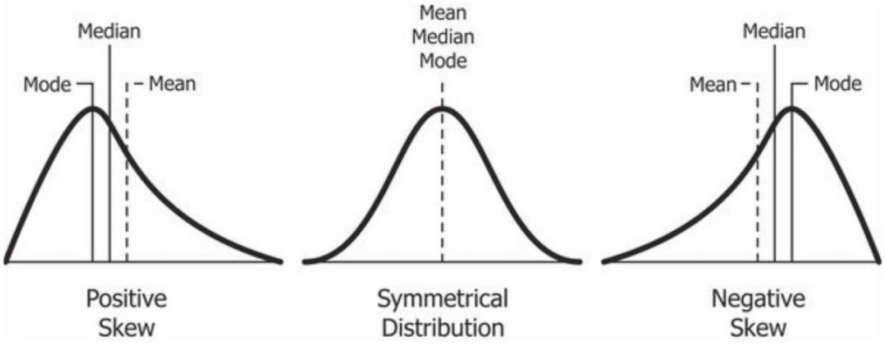
وان كل من معامل الالتواء (*skewness*) والتفلطح (*kurtosis*) وفق هذا التوزيع يكون على النحو الاتي

$$\gamma = \left[\frac{٢\sqrt{٢}\,\varepsilon\left[(٥\pi - ١٦)\varepsilon^{٢} - \pi\right]}{[(٣\pi - ٨)\varepsilon^{٢} + \pi]^{\frac{٣}{٢}}}\right]^{٢} \qquad \dots (٥)$$

$$\delta = \frac{\left(١٥\pi^{٢} + ١٦\pi - ١٩٢\right)\varepsilon^{٤} + ١٠\pi(٣\pi - ٨)\varepsilon^{٢} + ٣\pi^{٢}}{[(٣\pi - ٨)\varepsilon^{٢} + \pi]^{٢}} \qquad \dots (٦)$$

حيث ان (δ، γ)هما معامل الالتواء ومعامل التفلطح على التوالي

ان مشكلة تقدير المعالم مثل ($\mu, \sigma^{٢}$) لبيانات عشوائية كبيرة والتي تأخذ توزيعاً معيناً، كانت أحد الموضوعات التي تدرس بشكل معمق والشكل رقم (٢) التالي يوضح المنحنى عندما يكون هناك التواء



الشكل رقم (٢) يوضح شكل المنحنى عندما يكون هناك التواء نحو اليمين أو نجو اليسار

ونلاحظ من الشكل أعلاه ان في حال كون التوزيع متماثل فان الوسط الحسابي=الوسيط=المنوال ($\mu = median = mode$) اما إذا كان غير متماثل فيكون على النحو التالي إذا كان الالتواء

موجب فان الوسط الحسابي>الوسيط>المنول $(mode < median < \mu)$ اما إذا كان الالتواء سالب فان المنوال>الوسيط>الوسط الحسابي $(\mu < median < moed)$.

٦-طرق التقدير الحصينة[٥] :-

ونظراً لِحساسية الطرائق الاعتيادية تجاه هذه المشاكل تم إيجاد الطرائق الحصينة أو (المقاومة) كأسلوب بديل عن الطرائق الاعتيادية في التقدير وتتصف هذه الطرائق بأنها أقل حساسية تجاه القيم الشاذة، وأن هذه الطرائق الحصينة تكون مكافئة للطرائق الاعتيادية في حال عدم وجود قيم شاذة وأنها تنتج مقدرات أفضل من مقدرات الطرائق الاعتيادية في حالة وجود قيم شاذة .

٦-١طريقة (GM) الحصينة المعممة[١٣] :-

ان طريقة (M) هي من أشهر طرق التقدير الحصينة تجاه وجود القيم الشاذة في قيم المتغير المعتمد (ماثل) لأنها غير حصينة لوجود نقاط الرفع اقتُرحت مقدرات طريقة (M) المعممة (Generalized M) حيث ان الهدف كان هو إيجاد اوزان تأخذ بنظر الاعتبار القيم الشاذة في كل من المتغيرات المستقلة والمتغير المعمد ان مقدرات هذه الطريقة تكون بالصورة الاتية

$$\sum_{i=1}^{n} w_i(x_i)\Psi\left\{\frac{e_i}{\hat{\sigma}_e}\right\}x_i = 0 \qquad \dots(٧)$$

حيث ان Ψ هي الدالة الأساسية (كما في مقدرات طريقة M) وعادةً تُسمى دالة Huber، والاوزان w_i تعمد في المقام الأول على تقديرات المربعات الصغرى ويتم تحديثها بصورة مستمرة ويمكن الحصول على المعالم وفقا لهذه الطريقة من خلال

$$\hat{\beta} = (X'WX)^{-1}X'WY$$

٦-٢طريقة (S) الحصينة[١٢]:-

ان استعمال الوسيط كقيمة وزنية في طريقة (M) يؤدي الى فقدان توزيع البيانات وكذلك الدالة الاحتمالية للبيانات وهذه نقطة الضعف في طريقة (M) لذا لابد من إيجاد طريقة للتغلب على نقطة الضعف هذه، حيث تم وصف مقدرات S من قبل (yohai،Rousseeuw) عام ١٩٨٤

للتغلب على نقطة الضعف هذه. أن هذه الطريقة استعملت الانحراف المعياري للبواقي في طريقة (M) للتخلص من نقاط الضعف للوسيط

ان مقدر (S) يكون بالصيغة التالية

$$\hat{\beta}_S = \min_{\beta} \hat{\sigma}_s(e_1, e_2, \dots, e_n)$$

وبتحديد قيمة حصينة لـــ($\hat{\sigma}_s$) ويتحقق

$$\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{i=0}^{k} x_{ij}\beta_j}{\hat{\sigma}_s}\right) \quad \dots(٨)$$

$$\hat{\sigma}_s = \sqrt{\frac{1}{nk}\sum_{i=1}^{n} w_i e_i^2} \quad \dots(٩)$$

حيث ان $K = 0.199, w_i = w_\sigma = \frac{p(u_i)}{u_i^2}$

والتقدير الابتدائي لـــ($\hat{\sigma}_s$) هو

$$\hat{\sigma}_s = \frac{median\,|e_i - median(e_i)|}{0.6745} \quad \dots(١٠)$$

وبالاشتقاق الجزئي لـــ(β)

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}_s}\right) = 0, j = 0,1,\dots,k \quad \dots(١١)$$

حيث ان $\psi = \rho'$

$$\psi(u_i) = \rho(u_i) = \begin{cases} u_i\left[1 - \left(\frac{u_i}{c}\right)^2\right]^2, & |u_i| \le c \\ 0, & |u_i| > c \end{cases} \quad \dots(12)$$

إذ أن ψ_i دالة IRLS المرجحة

$$w_i(u_i) = \begin{cases} \left[1 - \left(\frac{u_i}{C}\right)^2\right]^2 & |u_i| \le C \\ 0 & |u_i| > C \end{cases} \quad \dots(13)$$

حيث ان $u_i = \frac{e_i}{\sigma_s}$, $c = 1.547$

ان مقدرات (S) هي أكثر حصانة من مقدرات (M) لأنها تمتلك تحيز قليل وتباين قليل أيضا في حال البيانات الملوثة (*Ehab M. Almetwally*)

ان نقطة الانهيار لمقدرات (S) تكون مساوية الى 50% ويمكن حساب على النحو التالي

$$\varepsilon_n^* = \frac{\frac{n}{2} - p + 2}{n} \quad \dots(14)$$

٧-معيار المقارنة[٣](MSE):-

لغرض المقارنة بين الطرائق المستخدمة في التقدير ومعرفة أي الطرق أفضل فأننا سوف نعتمد على المعيار الآتي وهو

متوسط مربعات الخطأ Mean Square Error (MSE) يعتبر هذا المعيار من المعايير المهمة في نماذج الانحدار
الخطية ويعرف بالصيغة الاتية :

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} \quad \dots(15)$$

٨-الجانب التطبيقي:-

تم اخذ عينه من المرضى المصابين بأمراض الدم في مستشفى الكرامة التعليمي في محافظه واسط بحجم (٢٤٣)مريض وتم تحديد المتغير المعتمد y بهيموجلوبين الدم HGB اما المتغيرات التابعة الاخرى فتمثل (HCT ، MCV ، MCHC ، MCH ، RBC) وهي كالاتي :-

(Y) يمثل المتغير المعتمد هيموجلوبين الدم HGB .

(X١) يمثل المتغير المستقل الاول HCT .

(X٢) يمثل المتغير المستقل الثاني MCV .

(X٣) يمثل المتغير المستقل الثالث MCHC .

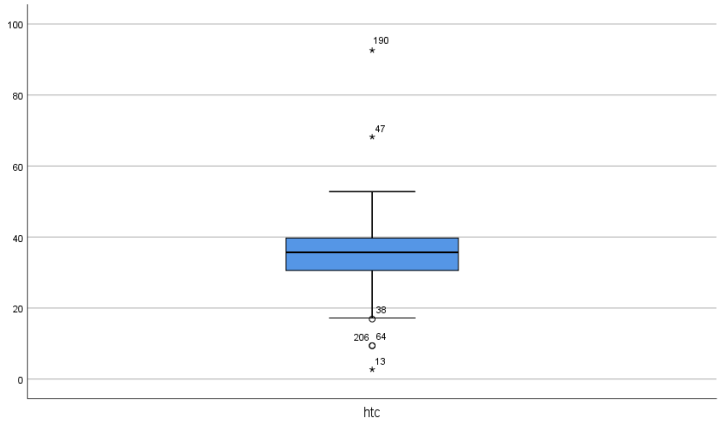
(X٤) يمثل المتغير المستقل الرابع MCH .

(X٥) يمثل المتغير المستقل الخامس RBC .

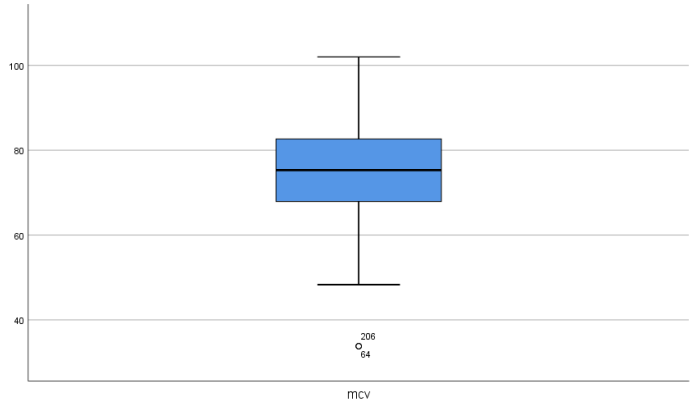
٨-١الكشف عن المشاهدة الشاذة بواسطة طريقة (Box plot):-

من خلال اجراء تحليل البيانات للدراسة للكشف عن وجود القيم الشاذة بالاعتماد على طريقة (Box plot) ومن الاشكال

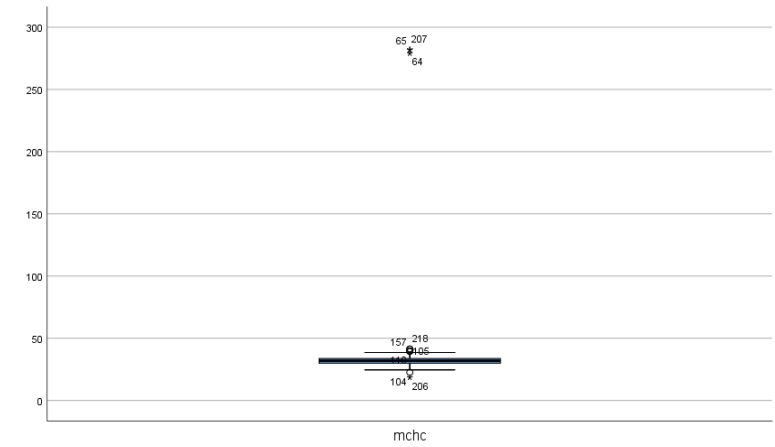
البيانية التالية (٣-٧) تبين وجود قيم شاذة وحسب كل متغير



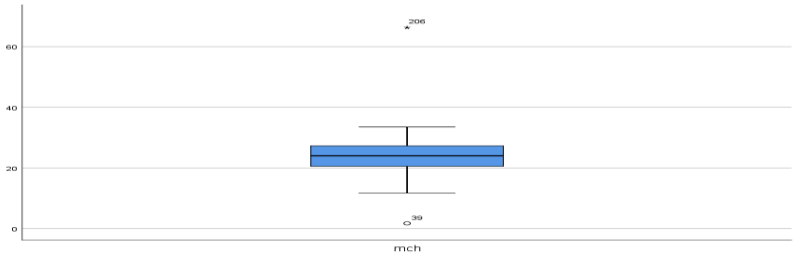
الشكل رقم (٣) يمثل القيم الشاذة للمتغير المستقل الاول (HTC)



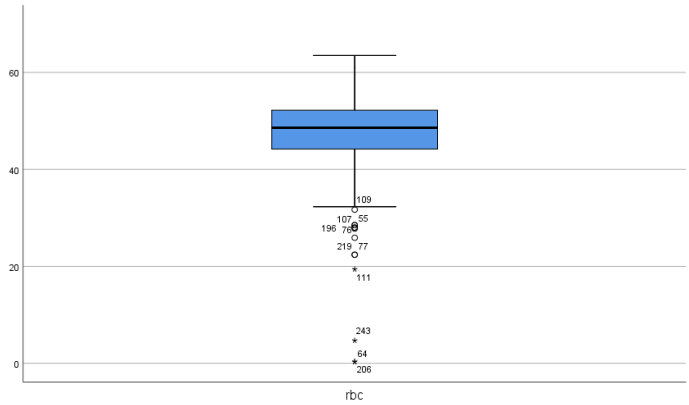
الشكل رقم (٤) يمثل القيم الشاذة للمتغير المستقل الثاني (MCV)



الشكل رقم (٥) يمثل القيم الشاذة للمتغير المستقل الثالث (MCHC)



الشكل رقم (٦) يمثل القيم الشاذة للمتغير المستقل الرابع (MCH)



الشكل رقم (٧) يمثل القيم الشاذة للمتغير المستقل الخامس (RBC)

٨-٢معالم انموذج الانحدار وفق طريقتي (GM) وطريقة(S):-

وبأجراء التحليل لبيانات الدراسة وبالاعتماد على البرنامج الاحصائي (Matlab-٢٠١٥) تم تقدير معالمانموذج الانحدار بوجود قيم شاذة بالاعتماد على الطريقتين (GM) المعممة الحصينة و(S) الحصينة وتوصلنا الى نتائج التقدير المبينة في جدول رقم (١) التالي :

جدول رقم (١) يوضح معالم انموذج الانحدار لطريقتين (GM) و(S)

b_o	b_1	b_2	b_3	b_4	b_5	
	٣		٠	١	٢	٠
	٤		٠	٢	٣	٣

ومن خلال الجدول اعلاه يمكن الحصول على النماذج المقدرة بالطريقتين وكالاتي :

انموذج الانحدار وفقاً لطريقة (GM) يمكن كتابة على الشكل الاتي

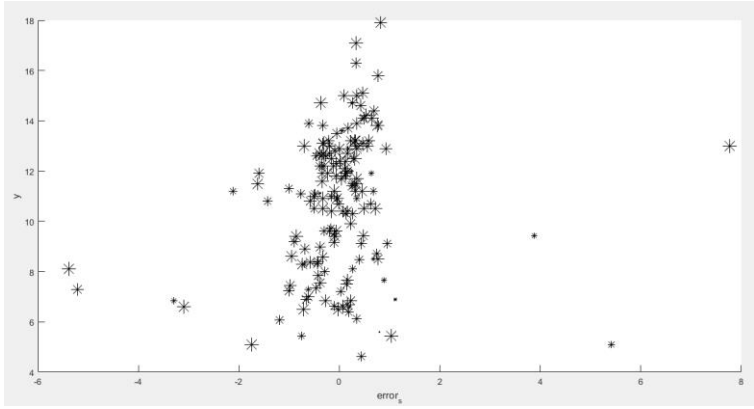
$$y_i = -٠.٩٦٨٣ + ٠.٠٠٣١x_1 - ٠.٠٠٦٠x_2 - ٠.٠٢٢١x_3 + ٠.٠٢١٢x_4 + ٠.٠٠٩٦x_5$$

انموذج الانحدار وفقاً لطريقة (S) يمكن كتابة على الشكل الاتي

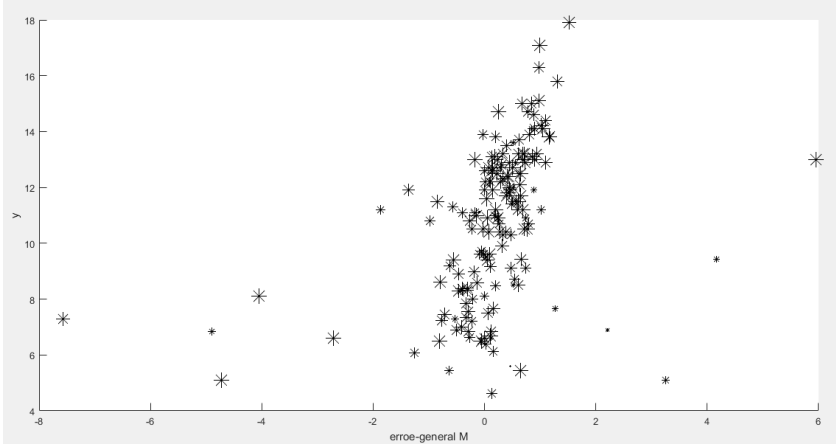
$$y_i = -١.٠٠٣٤ + ٠.٠٠٤١x_1 + ٠.٠٠٣٦x_2 + ٠.٠١٠٣x_3 - ٠.٠٠٠٣x_4 - ٠.٠٠٠٣x_5$$

٨-٣الانتشار للمتغير المعتمد مع البواقي وفق طريقتي (GM) وطريقة(S):-

يمثل رسم الانتشار القيم للمتغير المعتمد مع البواقي وفقا للطريقتين الحصينتين وحسب الاشكال (٨) (٩) كالتالي :



الشكل (٨) يمثل شكل الانتشار للمتغير المعتمد مع البواقي وفق طريقة (S)



الشكل (٩) يمثل شكل الانتشار للمتغير المعتمد مع البواقي وفق طريقة (GM)

٨-٤معيار المقارنة وفق طريقتي (GM) وطريقة(S):-

وللوصول الى افضل طريقة لتقدير معلمات الانحدار للطريقتين تم استخدام معيار المقارنة MSE وتم الحصول

على النتائج التالية حسب جدول رقم (٢) التالي :

جدول رقم (٢) يوضح معيار المقارنة (MSE) بين طريقتي (GM) وطريقة (S)

mse_{GM}	mse_S
٠.٠٠٢٣٤	٠.٠٠٢٦٧

يوضح جدول رقم (٢) المقارنة بين طريقتي (GM) وطريقة (S) باستعمال معيار متوسط مربعات الخطأ وتبين ان طريقة (GM) الحصينة المعممة هي الافضل لامتلاكها اقل قيمة لمعيار المقارنة MSE .

الاستنتاجات:-

تعد طريقة (Mse GM) كأفضل طريقة لتحقيق النتائج وحسب جدول (٢) الذي تبين قيمة المعيار ٠.٠٠٢٣٤

التوصيات:-

استعمال طرائق حصينة لم يتم التطرق لها مسبقا من قبل الباحثين بصورة كبيرة والتعرف على خصائصها بالأضافة الى حساسيتها تجاه وجود القيم الشاذة وكيفية التخلص منها .