تأثير القيم الشاذة ونقطة الأصل على نتائج تحليل الانحدار

ألاء عبد الستار داوود حمودات

قسم الإحصاء والمعلوماتية ، كلية علوم الحاسبات والرياضيات ، جامعة الموصل ،موصل ، العراق

(تاريخ الاستلام: ١٥ / ٢ / ٢٠٠٩ ، تاريخ القبول: ٢٥ / ١٠ / ٢٠٠٩)

الملخص

من المؤثرات على تحليل الانحدار هي القيم الشاذة ومرور خط أو مستوى الانحدار من نقطة الأصل ، وتم الكشف عن وجود القيم الشاذة بطريقة الرسم الصندوقي وعولجت بطريقة متوسط البتر ، واستخدام المقياسين معامل التحديد (R^2) ومتوسط مربعات الخطأ (MSe) للمقارنة بين نتائج تحليل الانحدار عند احتواء البيانات على القيم الشاذة واحتواء النموذج على الثابت β_0 وعدم احتوائهما تارة أخرى .

المقدمة

يعرف تحليل الانحدار (Regression analysis) بان مقياس رياضي لمتوسط العلاقة بين متغيرين أو أكثر بدلالة وحدات قياس المتغيرات ذات العلاقة وتمثل العلاقة بنماذج الانحدار ، ويعد تحليل الانحدار من الأدوات الإحصائية الأكثر استعمالاً لأنه يعطينا طريقة سهلة لتحديد العلاقة بين المتغيرات التي يمكن التعبير عنها بشكل معادلة بين متغير الاستجابة (y) مع واحد أو أكثر من المتغيرات التوضيحية المستقلة ($X_1, X_2, X_3, \ldots, X_n$) ، ويعد العالم الانكليزي(١٨٢٢-١٩١١) Francis Calton أول من استخدم مفهوم الانحدار في التطبيقات البايولوجية بهدف اكتشاف بعض العلاقات بين المتغيرات البايولوجية .

إن إجراء التحليل الاحصائي في جميع النواحي العلمية يعتمد بالدرجة الأساس على اختيار مجموعة من البيانات وتنقية هذه البيانات من القيم الشاذة التي تشكل انحرافاً واضحاً عن بقية المشاهدات.

إن طرائق الكشف عن القيم الشاذة ومعالجتها قد احتلت مساحة واسعة من البحوث الإحصائية، في حين تعتمد الجذور التاريخية لاكتشاف المشاهدات الشاذة الى عام 1755 عندما حاول (Bosocovich) تحديد اهليجية الأرض من خلال عشرة قياسات ولكنه تعتمد باستبعاد قياسين منهم لتطرفهما الشديد [1]. وفي عام 1972 توصل الى العالم[2] استخدام البواقي القياسية لاكتشاف مشاهدة شاذة واحدة في نماذج الانحدار الخطي البسيط. وفي عام 1977 اقترح Tukey [3] الطريقة الـ β_0 المعلمية المسماة طريقة الصندوق والقطع المخططة حيث تعتمد على الرسم في تحديد المشاهدات الشاذة . واقترح [4] طريقة لتحويل حالة المتغيرين الى حالة المتغير الواحد مع اخذ الانحرافات المطلقة ومن ثم اعتماد طريقة Tukey اللامعلمية لاكتشاف المشاهدات الشاذ . واقترحت [5] طريقة لتقدير المشاهدات الشاذة لنماذج الانحدار الخطي البسيط من خلال عمود البواقي $e_{(i)}$ بعد حذف المشاهدة (i) وباستخدام بعض مقاييس التشتت والمتوسط. واقترحا الباحثين[6] أسلوب لتقدير المشاهدات الشاذة العائدة لنماذج الانحدار الخطي البسيط والذي يعتمد على الشكل الدائري لانتشار النقاط حول مركزه المتمثل بمتوسطي المتغيرين التوضيحي والاستجابة . كما قدمت [1] دراسة عن الأخطاء وتأثيرها على نتائج تحليل الانحدار لمتغيرات المواليد الخدج .

الهدف

للتعرف على تأثير وجود بعض القيم الشاذة على نتائج تحليل الانحدار مقارنة بمعالجتها . قد تحتوي معادلة الانحدار التقديرية على الثابت β_0 وقد لا تحتوي عليه ويعتمد ذلك على طبيعة العلاقة بين المتغيرات واختبار الفرضية الخاصة بالثابت ، من هذا المنطلق وضع الهدف الثاني لدراسة التداخل بين احتواء المعادلة على الثابت من عدمه مع تلازم وجود القيم الشاذة وعند معالجتها ، ويطبق ذلك على بيانات تشير العلاقة بين متغيراتها وعلى حتمية استبعاد الثابت من المعادلة في حين تؤيد اختبار الفرضية الخاصة به وجوده في المعادلة .

تحليل الانحدار

يقتصر استخدام النموذج الخطي البسيط على تحليل العلاقة بين متغير الاستجابة وعلاقته بمتغير توضيحي واحد ، ويوصف المتغير التوضيحي X بالمتغير ألسببي (pridector variable) والمتغير الآخر هو متغير الاستجابة Y (response variable) أي إن هناك علاقة دالية بين Y وX أي Y=f(X) ولكن في الواقع هناك دراسات تتطلب وضع متغير الاستجابة كدالة لأكثر من متغير توضيحي واحد ، مثل هذه الدراسات تغطى بواسطة النموذج الخطي العام الخاص بالانحدار الخطي المتعدد والأسلوب الأخير هذا ما هو إلا عبارة عن امتداد طبيعي للنموذج الخطي البسيط ، حيث إن العلاقة الدالية function relationship بين متغير الاستجابة y والمتغيرات التوضيحية المستقلة ($X_1, X_2, \ldots, X_n$) في تحليل الانحدار المتعدد تأخذ الصيغة التالية :

$$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i3}+\ldots..+\beta_mX_{im}+U_i \quad i=1,2,3,\ldots..,m$$

حيث أن :-

Y_i: المتغير التابع أو قيمة متغير الاستجابة .

β_0 : نقطة تقاطع مستوى الانحدار بالمحور Y.

$\beta_0, \beta_1, \beta_2, \beta_3, \ldots.., \beta_m$: تمثل معالم النموذج المجهولة .

U_i : الأخطاء العشوائية .

ويمكن الحصول على معالم معادلة الانحدار التقديرية المستحصلة من بيانات عينة كالأتي وباستخدام اسلوب المصفوفات[7] :

$$\hat{\beta} = (x'x)^{-1}x'y$$

حيث إن :

$x'x$: مصفوفة المعلومات Information Matrix

$x'y$: متجه لحاصل ضرب المتغيرات التوضيحية في متغير الاستجابة.

وبذلك تكون معادلة الانحدار التقديرية هي :

$$\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_{i1} + \hat{B}_2 x_{i2} + ... + \hat{B}_m x_{im} + U_i$$

نقطة التقاطع خط الانحدار مع الاحداثي y

وهي عبارة عن نقطة تقاطع خط الانحدار بالمحور y ، كذلك β_0 تعطي متوسط الاستجابة عندما تكون قيم xi1 تساوي صفرا . وان توزيع $\hat{\beta}_0$ هو التوزيع الطبيعي بوسط حسابي قدره :

$$E(\hat{\beta}_0) = \beta_0$$

وتباين قدره :

$$S^2(\hat{\beta}_0) = MSe\left[\frac{1}{n} + \frac{(\overline{X})^2}{S_{xx}}\right]$$

MSe : متوسط مربعات الخطأ

أما اختبار الفرضية الخاصة بـ β_0 هي :

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

والمختبر الإحصائي t :

$$|t| = \frac{\beta_0}{S_{\hat{\beta}_0}}$$

اذا كانت القيمة المطلقة للمختبر الأحصائي أصغر من القيمة الجدولية $t_{\frac{1}{2}\alpha,n-2}$ تقبل فرضية العدم H_0 وترفض الفرضية البديلة H_1 بمعنى أن خط الانحدار يمر من نقطة الأصل β_0 وأن معادلة الانحدار التقديرية تحتوي على الثابت β_0 .

كذلك يمكن تنظيم جدول تحليل التباين وحسب الفرضية الآتية :

جدول تحليل التباين

S.O.V	d.f	S.S	M.S	F
$R(X_1,X_2,\ldots,X_m)$	m	SSR	SSR/m	MSR/MSe
Error	n-m-1	SSe	SSe/n-m-1	
Total	n-1			

حيث ان :

SST : مجموع المربعات الكلي

$$SST = \mathbf{Y'Y} - \frac{(\sum Y_i)^2}{n}$$

SSR: مجموع المربعات العائدة الى الانحدار

$$SSR(X_1 X_2 ... X_m) = \beta' X' Y - \frac{(\sum Y_i)^2}{n} \quad SSe :$$

مجموع المربعات العائدة الى الخطأ

$$SSR(X_1 X_2 ... X_m) - SSe = SST$$

فإذا كانت قيمة F المحسوبة مساوية أو اكبر من قيمة F الجدولية عند مستوى معنوية α يدل على أهمية النموذج في تفسير التغيرات في المتغير المعتمد وعلى الأقل احتواء النموذج على متغير واحد.

القيم الشاذة Outliers

يواجه الباحثون أحياناً مجموعة من المشاكل الاحصائية قد يكون بعضها واضحاً أمامه وبعضها غير واضح ، فيجد نفسه بحاجة الى طرائق جديدة تمكنه من تنظيم سير التجربة عن طريق جعل الخطأ الناتج اصغر ما يمكن وفي الوقت نفسه يحصل على تقدير غير متحيز للمقدار الذي يبحث عنه[7] . وابتدأت فكرة دراسة القيم الشاذة بأفكار بسيطة معتمدة على الحدس والتخمين ، فهي لم تأخذ بنظر الاعتبار فكرة وأغراض التباعد discordancy للقيم الشاذة والانعكاسات المختلفة على النموذج جراء وجود تلك القيم.

والقيم الشاذة هي تلك المشاهدات التي تبدو غير منطقية وتظهر انحرافاً كبيراً عن سائر مكونات العينة التي وجدت فيها تلك المشاهدة ، لقد ورد عن Barnett [8] المشاهدة الشاذة في مجموعة من البيانات بأنها تلك المشاهدة التي تبدو غير منطقية اذا ما قورنت بسائر مجموعة البيانات .

إذا تأكد أن النتيجة تعود الى واحد أو أكثر من الأسباب المذكورة سهل الأمر وسهلت المعالجة ، فقد يصار الى إزالة المسبب . من المهم فحص البيانات لتخليصها من القيم الشاذة قبل الدخول في التحليل الإحصائي وذلك للحصول على نتائج دقيقة وبالتالي اتخاذ قرارات صائبة .

الكشف عن القيم الشاذة :

تناول العديد من الباحثين في بحوثهم موضوع القيم الشاذة ودراسة تأثيراتها في دقة النتائج والوصول الى أفضل القرارات ، وتم تطبيق الكشف عن القيم الشاذة على نماذج الانحدار والتجارب المصممة واختبارها في بيانات متعدد المتغيرات وبأساليب مختلفة من حيث الكشف الكلي للمشاهدات أو الكشف الجزئي للأعمدة الشاذة وتنوعت الطرائق المكتشفة الحديثة في مجال القيم الشاذة وتحديدها ومن أهم هذه الطرائق طريقة الرسم الصندوقي وتستخدم العرض بالرسم لتعيين القيم الشاذة وتتلخص هذه الطريقة كالأتي :

١- **قيمة الوسيط (Median)** وتعتمد قيم الوسيط على عدد القيم اذا كانت فردية أم زوجية إذ بعد ترتيب قيم المتغير تصاعديا" يمكن تحديد رتبة الوسيط والمساوية لـ $\frac{n+1}{2}$ وتحديد قيمة الوسيط والتي هي القيمة المقابلة للترتيب الذي يتوسط البيانات في حالة البيانات الفردية ، وتكون رتبة الوسيط في حالة البيانات الزوجية والمساوية لـ $\frac{n}{2}$ وتحديد قيمة الوسيط وتمثل متوسط القيمتين الواقعتين وسط البيانات في حالة البيانات الزوجية .

٢- **قيمة الربيع الأدنى (Lower Quartile)** وتمثل قيمة الربيع الأدنى للصندوق ويمكن حسابه كالأتي :

$$Q_L = \text{INTERGER}\left[\frac{\overline{Me}+1}{2}\right]$$

$\overline{Me}$: قيمة الوسيط

Integer : إهمال الكسر الناتج من القسمة (أي العدد الصحيح)

٣-**قيمة الربيع الأعلى Upper Quartile** تمثل قيمة الربيع الأعلى الحد الأعلى للصندوق ويمكن حسابه كالأتي :

$$Q_u = Q_L + \overline{Me}$$

ويمثل الفرق بين الربعين (الربيع الأدنى والربيع الأعلى للصندوق) طول الصندوق أو متوسط الانتشار أو المدى الربيعي (H - spread)

$$H - spread = Q_u - Q_L$$

٤- أكبر قيمة غير شاذة (Upper Extreme) حيث أن أي قيمة في لبيانات أكبر من هذه القيمة تعد تلك القيمة شاذة بمعنى أنها تعد قيمة شاذة إذا وقعت خارج المدى الأتي :

$$U.E = Q_u + 1.5(H - spread)$$

٥- أصغر قيمة غير شاذة (Lower Extreme) أي قيمة في البيانات اصغر من هذه القيمة وتعد قيمة شاذة اذا وقعت خارج المدى الأتي :

$$L.E = Q_L - 1.5(H - spread)$$

أما بالنسبة الى عرض الصندوق فقد استخدم Tukey قاعدة تتناسب مع الجذر التربيعي لحجم العينة ($\sqrt{n}$) والسياج الداخلي (Inner Fence) يمثل بعد المسافة (R) عن الربيعين أي أن :

$$R = 1.5(H - spread)$$

$$\text{Inner Fence} = Q_L - R \text{ (السياج الداخلي)}$$

$$\text{Inner Fence} = Q_u - R \text{ (السياج الداخلي)}$$

وان السياج الخارجي يمثل بعد المسافة (2R) عن الربيعين أو بعد المسافة R عن السياج الداخلي أي أن :

$$\text{Outer Fence} = Q_L - 2R \text{ (السياج الخارجي)}$$

$$\text{Outer Fence} = Q_u - 2R \text{ (السياج الخارجي)}$$

ويمتاز الرسم الصندوقي بان شكله بسيط ولا يتأثر بالقيم الشاذة ويمكن أن يحددها بكل وضوح ويحتوي الصندوق المستطيل تقريبا" ٥٠% من القيم الوسطية من البيانات وعلى ٥٠% من القيم الأخرى تقع خارج الصندوق ويكون قسم منها بالخطين الجانبين والقسم الأخر إن وجد يكون قيم شاذة [9]

معالجة القيم الشاذة

بعد أن يتم الكشف عن القيمة الشاذة أصبح من الضروري معالجة هذه القيمة واستخدمت طريقة متوسط البتر Trimmed mean [10] وتمتاز هذه الطريقة بالدقة وسهولتها في التطبيق . وتتلخص الطريقة بترتيب مشاهدات العمود المشخص بأنه شاذاً تصاعدياً أو تنازلياً وبعدها يتم تحديد قيمة الوسيط للمشاهدات المرتبة ومن ثم تقدير القيمة الشاذة حسب صغرها أو كبرها مع بقية المشاهدات وكالاتي : إذا كانت القيمة الشاذة المراد تقديرها اصغر من الوسيط يتم حذف اكبر المشاهدات في العمود الشاذ فضلاً عن القيمة الشاذة المراد تقديرها وإيجاد الوسط الحسابي للمشاهدات المتبقية في العمود والذي يعد تقديراً لهذه القيمة ، أما اذا كانت المشاهدة الشاذة المراد تقديرها أكبر من الوسيط تحذف أصغر المشاهدات في العمود فضلا عن القيمة الشاذة المراد تقديرها وإيجاد الوسط الحسابي للمشاهدات المتبقية في العمود وهذا يكون هو التقدير للمشاهدة الشاذة ،

بعض مقاييس كفاءة النموذج المقدر :

أولا" – معامل التحديد (R^2)

Coefficient of Determination

بعد توفيق أو مطابقة النموذج الخطي للبيانات المعطاة يجب عمل تقييم لدقة المطابقة ، إن المقياس الأكثر اتساعاً في الاستخدام هو معامل التحديد R^2 الذي يعرف بأنه نسبة الانحرافات الموضحة بواسطة معادلة الانحدار التقديرية الى مجموع المربعات الكلي للمتغير المستجيب وهو نسبة مساهمة معادلة الانحدار التقديرية في تفسير أو شرح الانحرافات الكلية في قيم (y) حول الوسط الحسابي $\bar{y}$ ويمكن أن يعبر عن معامل التحديد بالصيغة [11]

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{y})^2}{\sum(Y_i - \bar{y})^2} = \frac{SSR\ (X_1\ X_2 \ldots X_m)}{S_{yy}}$$

وبما إن مجموع المربعات الكلي SST قد جزء الى مركبتين رئيسيتين هما مجموع مربعات الانحدار SSR ومجموع مربعات الأخطاء أو البواقي SSe .

$$SST = SSR + SSe$$

حيث ان هذه المجاميع هي مركبات موجبة لذلك نلاحظ أن R^2 له مدى بين 0 و 1 فعندما يكون النموذج ملائماً بشكل جيد للبيانات فمن الواضح أن قيمة تقترب من الواحد وعندما يكون معامل التحديد قريب من الصفر فان ذلك يعني أن النموذج لا يفسر التغيرات في متغير الاستجابة . وذكر [12] انه عند زيادة عدد المتغيرات الداخلة في النموذج يؤدي الى زيادة في قيمة R^2 وتكون تلك المتغيرات ذات أهمية عند إدخالها في النموذج . نلاحظ من البرهان الأتي تأثير إدخال المتغيرات في قيمة R^2 عند اختيار عدد مختلف من المتغيرات التوضيحية لإدخالها في النموذج نلاحظ أن المقام يبقى ثابتا حيث انه مجموع مربعات انحرافات قيم المتغير المعتمد عن وسطه الحسابي ، أي إن (S_{yy}) هي قيمة ثابتة لكافة النماذج المختارة .

عليه يكون التغيير في قيمة R^2 ناتجا عن البسط $SSR(x_1\ x_2 \ldots\ldots x_m)$ ونلاحظ ما يأتي :

$$SSR(x_1 \ldots\ldots x_k \mid x_{k+1} \ldots\ldots x_m) + SSR(X_{K+1} \ldots\ldots X_m) = SSR(x_1\ x_2 \ldots\ldots x_m)$$

وبما ان كافة الحدود في المساواة $SSR(x_1\ x_2 \ldots\ldots x_m)$ هي مجموع مربعات فهي إذن موجبة ، عليه عند حذف الموجب $SSR(x_1 \ldots\ldots x_k \mid x_{k+1} \ldots\ldots x_m)$ من الحد الأيمن عندئذ يكون :

$$SSR(x_1\ x_2 \ldots\ldots x_m) > SSR(x_{K+1}\ x_2 \ldots\ldots x_m)$$

عليه فأن :

$$\frac{SSR(x_1\ x_2 \ldots\ldots x_m)}{SST} > \frac{SSR(x_1\ x_2 \ldots\ldots x_k)}{SST}$$

$$R^2(x_1 \ldots\ldots x_m) > R^2(x_1 \ldots\ldots x_k)$$

عليه فان عند إضافة متغير له تأثير معنوي الى النموذج يعمل على زيادة في قيمة R^2 .

عليه يجب استبعاد تلك المتغيرات التوضيحية التي ليس لها تأثيرا" على المتغير المعتمد وبذلك يمكن استخدام R^2 كمقياس للحكم على معادلة الانحدار التقديرية .

ثانيا" – متوسط مربعات البواقي Mean Square Error (MSe)

احد مقاييس الدقة التي تستخدم للحكم على ملائمة معادلة الانحدار التقديرية هو متوسط مربعات الخطأ MSe :

$$MSe = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-m-1} = \frac{\text{SSe}}{\text{n - m -1}}$$

ومن خصائص معامل التحديد ان مجموع المربعات الكلي $SST=S_{yy}$ ثابت لكافة النماذج الجزئية التي تتفاعل مع نفس المتغير المعتمد وبذلك فان إضافة متغير يزيد من مجموع المربعات العائدة للانحدار وهذا بدوره يؤدي الى تقليل قيمة MSe ، وهذا يوضح العلاقة بين MSe و R^2 عند اختيار متغيرات النموذج ، فالنموذج المفضل هو النموذج الذي يحتوي على اقل قيمة لـ MSe وأعلى قيمة لـ R^2 وبأقل عدد متغيرات .

طريقة الانحدار المتدرج

وهي طريقة تربط بين طريقتي الاختيار الأمامي والاختيار الخلفي وهي عبارة عن طريقة اختيار أمامي لكن عند كل خطوة من العملية فان المتغيرات التي يتم اختيارها يعاد اختبارها وهذا مشابه لطريقة الحذف العكسي ، وهذه الطريقة تحتاج إلى قيمتين من قيم F الجدولية هما F(in) للاختيار الأمامي و F(out) للحذف العكسي وخطواته :

١- نحسب قيمة F الجزئية لكل متغير توضيحي ، ثم نختار أعلى F جزئية ذات معنوية إحصائية ، وعند عدم وجود هكذا متغير نوقف العمليات الإحصائية ونتخذ قرار بعدم وجود أي متغير مؤثر على الظاهرة .

٢- ولاختيار المتغير التوضيحي الثاني نعيد النقطة أعلاه وقبل إن نختار متغيراً توضيحياً ثالثاً نجري الطريقة العكسية لمعرفة أهمية بقاء المتغير الذي اختير في المرة الأولى وذلك لتحديد حذفه من المعادلة أو إبقاؤها فيها ، ويتم التوقف بإتباع خطوات هذه الطريقة عند عدم معنوية أي متغير في الاختيار الأمامي وبهذا الأسلوب يمكن الاستمرار الى أن نصل الى النموذج النهائي .

الجانب التطبيقي

استخدمت البيانات التي جمعت من سجلات مستشفى البتول التعليمي للولادة في محافظة نينوى من قبل [1] وقد تم الاعتماد على السجلات الخاصة بالأطفال حديثي الولادة والاستمارات الخاصة بالمعلومات حول أطفال الخدج من شعبة الخدج في المستشفى . حيث أخذت عينة عشوائية تتكون من (١٠٠) طفل ، وتمت عملية جمع للبيانات بالاعتماد على (٣) متغيرات توضيحية وهذه المتغيرات هي :

y : فترة بقاء الطفل الخديج بالحاضنة مقاساً بالأيام .

X_1 : عمر الطفل الخديج مقاساً بالأسابيع .

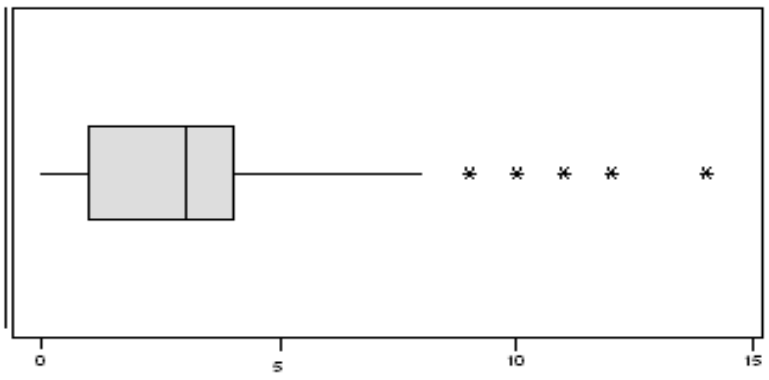
X_2 : وزن المولود الخديج مقاساً بالغرامات .

X_3 : عمر الأم مقاساً بالسنين.

حيث اجري تحليل الانحدار متغير الاستجابة (y) على الثلاث متغيرات التوضيحية وذلك للتعرف على وجود القيم الشاذة وكيفية معالجتها وتأثير الثابت β_0 على نتائج تحليل الانحدار عند وجوده في النموذج وعدم وجوده باستخدام البرامجيات الحاسوبية Minitab v.13.2 كما استخدم تحليل الانحدار المتدرج Stepwise Regression لتحديد المتغيرات المؤثرة على حياة الخديج من بقاءه حياً أو وفاته .

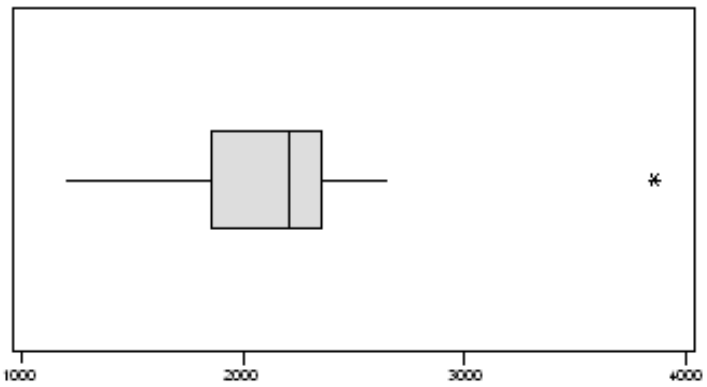
الكشف عن القيم الشاذة .

اظهر الرسم الصندوقي (Box plot) خمسة قيم شاذة في المتغير المعتمد y (period) كما في الشكل (١) .



الشكل (١) يوضح القيم الشاذة في المتغير (y)

وقيمة شاذة واحدة في المتغير التوضيحي (X_2) (الوزن weight) كما في الشكل (٢) .



الشكل (٢) يوضح القيم الشاذة في المتغير (X2)

تحليل الانحدار قبل معالجة القيم الشاذة

ستتم ملاحظة تأثير القيم الشاذة على نتائج تحليل الانحدار عند احتواء النموذج على الثابت β_0 وعند عدم احتواءه عليه .

١- احتواء النموذج على الثابت β_0

يشير الجدول (١) على النتائج التحليلية الآتية :

الجدول (١) : نتائج تحليل الانحدار للمتغيرات الثلاث في المتغير المعتمد

Regression Analysis: PERIOD(Y) versus AGE(X1); WEIGHT(X2); M.AGE(X3)

```
The regression equation is
PERIOD(Y) = 22.8 - 0.437 AGE(X1) - 0.00206 WEIGHT(X2) - 0.0377 M.AGE(X3)

Predictor        Coef     SE Coef          T        P
Constant       22.768       1.996      11.41    0.000
AGE(X1)       -0.43659     0.08179      -5.34    0.000
WEIGHT(X2   -0.0020590   0.0007815      -2.63    0.010
M.AGE(X3      -0.03774     0.02618      -1.44    0.153

S = 1.823       R-Sq = 55.5%     R-Sq(adj) = 54.1%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	397.56	132.52	39.86	0.000
Residual Error	96	319.19	3.32		
Total	99	716.75			

R^2=0.5546

من قيم P في الجدول (١) يتبين وجود دلالة إحصائية للمتغيرين X1 و X2 في تأثيرهما على الاختلافات في المتغير المعتمد y عند مستوى معنوية ٠,٠١ في حين لا يوجد أهمية للمتغير X3 ، كما يلاحظ قيمة **تحديد أفضل المتغيرات** ولتحديد أفضل المتغيرات المؤثرة في المتغير المستجيب y استخدمت طريقة الانحدار المتدرج الذي يشير الى نتائجها الجدول (٢) .

معامل التحديد R^2=0.5546 أي ان حوالي ٥٥% من الاختلافات الموجودة بين قيم (y) تعود أسبابها إلى تأثير المتغيرات التوضيحية المدروسة .

الجدول (٢) : نتائج التحليل التدريجي

Stepwise Regression: PERIOD(Y) versus AGE(X1); WEIGHT(X2); M.AGE(X3)

```
Backward elimination.  Alpha-to-Remove: 0.05
Response is PERIOD(Y on  3 predictors, with N =  100
Step          1        2
Constant      22.77    21.85
AGE(X1)      -0.437   -0.423
T-Value       -5.34    -5.18
P-Value       0.000    0.000

WEIGHT(X2   -0.00206 -0.00235
T-Value       -2.63    -3.09
P-Value       0.010    0.003

M.AGE(X3     -0.038
T-Value       -1.44
P-Value       0.153

S              1.82     1.83
R-Sq          55.47    54.50
R-Sq(adj)     54.08    53.56
C-p             4.0      4.1
```

أشارت طريقة الانحدار المتدرج الى أهمية المتغيرين X_1 (عمر الخديج) و X_2 (وزن الخديج) ليكونوا في نموذج الانحدار وهذه النتيجة تطابق على ما أشارت اليه النتائج في الجدول (١) .

٢ – عدم احتواء النموذج على الثابت β_0

وضعت نتائج تحليل الانحدار في الجدول (٣) .

الجدول (٣) : نتائج تحليل الانحدار للمتغيرات الثلاث في المتغير المعتمد

Regression Analysis: PERIOD(Y) versus AGE(X1); WEIGHT(X2); M.AGE(X3)

```
The regression equation is
PERIOD(Y) = 0.290 AGE(X1) - 0.00378 WEIGHT(X2) + 0.0571 M.AGE(X3)

Predictor        Coef     SE Coef          T        P
Noconstant
AGE(X1)       0.29028     0.07828       3.71    0.000
WEIGHT(X2    -0.003776    0.001171      -3.23    0.002
M.AGE(X3      0.05708     0.03789       1.51    0.135

S = 2.784
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	957.27	319.09	41.17	0.000
Residual Error	97	751.73	7.75		
Total	100	1709.00			

R^2=0.560

من قيم P في الجدول (٣) يتبين وجود دلالة إحصائية للمتغيرين X1 و X2 في تأثيره على الاختلافات في المتغير المعتمد y عند مستوى معنوية ٠,٠١ في حين لا يوجد أهمية للمتغير x3، اما قيمة معامل التحديد R^2=0.560 أي ان حوالي 56% من الاختلافات الموجودة بين قيم (y) تعود أسبابها إلى تأثير المتغيرات التوضيحية المدروسة ، ونلاحظ ايضا" ارتفاع قيمة F=41.17 المحسوبة .

اختيار افضل المتغيرات

ولاختبار افضل المتغيرات استخدمنا طريقة الانحدار كما في جدول (٤) .

الجدول (٤) : نتائج التحليل التدريجي

Stepwise Regression: PERIOD(Y) versus AGE(X1); WEIGHT(X2); M.AGE(X3)

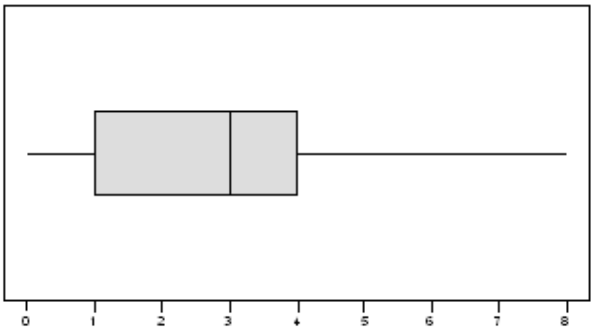
```
Backward elimination.  Alpha-to-Remove: 0.05

Response is PERIOD(Y on  3 predictors, with N =  100

Step          1        2
No constant
AGE(X1)       0.290    0.317
T-Value        3.71     4.12
P-Value       0.000    0.000

WEIGHT(X2    -0.0038  -0.0034
T-Value       -3.23    -2.96
P-Value       0.002    0.004

M.AGE(X3      0.057
T-Value        1.51
P-Value       0.135

S              2.78     2.80
C-p             3.0      3.3
```

أشارت طريقة الانحدار المتدرج الى أهمية المتغيرين X_1 (عمر الخديج) و X_2 (وزن الخديج) ليكونوا في نموذج الانحدار وهذه النتيجة تطابق مع نتائج النموذج المحتوي على β_0 .

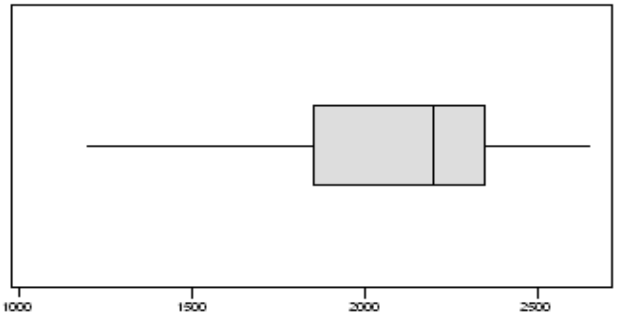
معالجة القيم الشاذة

وبعد ان تم الكشف عن القيم الشاذة أصبح من الضروري معالجة هذه القيم وتم استخدام طريقة متوسط البتر (Trimmed mean) لمعالجتها من خلال تشخيصها حذف اكبر الشاهدات في العمود فضلاً عن القيم الشاذة وإيجاد الوسط الحسابي الذي يمثل تقديراً للقيمة الشاذة. ظهر من تطبيق طريقة الرسم الصندوقي وباستخدام البرنامج الجاهز Minitab المشاهدات (١٠و١٢ و٤ و٤ و١١) والعائدة للمتغير y (period) هي قيم شاذة والتي تقع خارج الحدين الأدنى والأعلى للسياج الداخلي ، واستخدمت طريقة متوسط البتر لمعالجتها فكانت القيم (3.1589 و ٣,٣٢٣٠ و ٣,٤٢٠٥ و ٣,٠٩٢٥ و ٣,٢٣٥٧) على التوالي تمثل تقديراً لهم ،وقيمة المشاهدة (3850) والعائدة للمتغير (weight) هي القيمة الوحيدة الشاذة فكانت القيمة التقديرية لها هي (2131.9) .

والشكلين (٣) و (٤) يوضحان خلو المتغيرين y و X من القيم الشاذة وعلى التوالي .



الشكل (3) يمثل خلو متغير y من القيم الشاذة



الشكل (4) يمثل خلو المتغير x_2 من القيم الشاذة

تحليل الانحدار عند عدم وجود القيم الشاذة (بعد معالجة القيم الثابتة)

١- احتواء النموذج على الثابت β_0

تم إدراج نتائج تحليل الانحدار في الجدول (٥) .

الجدول (٥) : نتائج تحليل الانحدار للمتغيرات الثلاث في المتغير المعتمد

Regression Analysis: PERIOD(Y)_1 versus AGE(X1); WEIGHT(X2)_1; .AGE(X3)

```
                The regression equation is
PERIOD(Y)_1 = 16.2 - 0.231 AGE(X1) - 0.00305 WEIGHT(X2)_1 + 0.0180 M.AGE(X3)
Predictor         Coef     SE Coef          T        P
Constant        16.167       1.261      12.82    0.000
AGE(X1)       -0.23054     0.06377      -3.62    0.000
WEIGHT(X    -0.0030525   0.0006907      -4.42    0.000
M.AGE(X3       0.01795     0.01599       1.12    0.264
S = 1.147        R-Sq = 65.1%     R-Sq(adj) = 64.0%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	235.913	78.638	59.76	0.000
Residual Error	96	126.322	1.316		
Total	99	362.235			

R^2=0.6512

يتبين إن قيم P في الجدول (١) وجود دلالة إحصائية للمتغيرين X1 و X2 في تأثيره على الاختلافات في المتغير المعتمد y عند مستوى معنوية ٠,٠١ في حين لا يوجد أهمية للمتغير X3 ,وأن قيمة F=59.76 وقيمة معامل التحديد R^2=65.12 وانخفاض قيمة MSe =1.٣١٦.

كما اجري تحليل الانحدار المتدرج فكانت كسابقتها أي اختبار X_1 و X_2 كأفضل متغيرين يؤثران على متغير الاستجابة y . كما في جدول (٦) .

الجدول (٦) : نتائج التحليل التدريجي

Stepwise Regression: PERIOD(Y)_1 versus AGE(X1); WEIGHT(X2)_1; .AGE(X3)

```
Backward elimination.  Alpha-to-Remove: 0.05
Response is PERIOD(Y on  3 predictors, with N =  100
Step          1        2
Constant      16.17    16.59
AGE(X1)      -0.231   -0.232
T-Value       -3.62    -3.64
P-Value       0.000    0.000
WEIGHT(X2   -0.00305 -0.00298
T-Value       -4.42    -4.32
P-Value       0.000    0.000
M.AGE(X3      0.018
T-Value        1.12
P-Value       0.264
S              1.15     1.15
R-Sq          65.13    64.67
R-Sq(adj)     64.04    63.94
C-p             4.0      3.3
```

٢- عدم احتواء النموذج على الثابت β_0

أشارت طريقة الانحدار المتدرج الى أهمية المتغيرين X_1 (عمر الخديج) و X_2 (وزن الخديج) ليكونوا في نموذج الانحدار وهذه النتيجة تطابق على ما أشارت التحاليل السابقة .

جدول (٧) يوضح نتائج تحليل الانحدار للمتغيرات التوضيحية الثلاث على متغير الاستجابة y.

الجدول (٧) نتائج تحليل الانحدار للمتغيرات الثلاث في المتغير المعتمد

Regression Analysis: PERIOD(Y)_1 versus AGE(X1); WEIGHT(X2)_1; .AGE(X3)

```
The regression equation is
PERIOD(Y)_1 = 0.335 AGE(X1) - 0.00494 WEIGHT(X2)_1 + 0.0786 M.AGE(X3)
Predictor        Coef     SE Coef          T        P
Noconstant
AGE(X1)       0.33502     0.07541       4.44    0.000
WEIGHT(X2    -0.004943    0.001105      -4.47    0.000
M.AGE(X3      0.07859     0.02503       3.14    0.002

S = 1.879

Analysis of Variance
Source            DF          SS          MS         F        P
Regression         3      777.26      259.09     73.38    0.000
Residual Error    97      342.49        3.53
Total            100     1119.75
```

R^2=0.6941

يتبين ان قيم P في الجدول (٧) وجود دلالة احصائية للمتغيرين X1 و X2 و X3 في تأثيره على الاختلافات في المتغير المعتمد y عند مستوى معنوية ٠,٠١ ، كما يلاحظ ايضاً ارتفاع قيمة F=73.76 وقيمة معامل التحديد R^2=65.12 بعد معالجة البيانات من القيم الشاذة.

ولتحديد أفضل المتغيرات المؤثرة في Y استخدمت طريقة الانحدار المتدرج وكانت النتائج كما في الجدول (٨) .

الجدول(٨) نتائج التحليل التدريجي

Stepwise Regression: PERIOD(Y)_1 versus AGE(X1); WEIGHT(X2)_1; .AGE(X3)

```
Backward elimination.  Alpha-to-Remove: 0.05
Response is PERIOD(Y on  3 predictors, with N =  100
Step          1
No constant
AGE(X1)       0.335
T-Value        4.44
P-Value       0.000

WEIGHT(X    -0.0049
T-Value       -4.47
P-Value       0.000

M.AGE(X3      0.079
T-Value        3.14
P-Value       0.002

S              1.88
C-p             3.0
```

أشارت طريقة الانحدار المتدرج الى أهمية المتغيرين X_1 (عمر الخديج) و X_2 (وزن الخديج) و (x3) ليكونوا في نموذج الانحدار وهذه النتيجة تطابق على ما أشارت اليه النتائج في الجدول (٧) .

جدول مقارنات

من خلال ما تقدم يمكن تلخيص النتائج التي توصلنا اليها كما في جدول (٩) .

الجدول (٩)

	نتائج تحليل الانحدار بوجود القيم الشاذة				نتائج تحليل الانحدار بعد معالجة القيم الشاذة			
	F	R^2	MSe	Step wise	F	R^2	MSe	Step wise
بوجود الثابت β_0 في النموذج	39.86	55.46	٣,٣٢	X_1,X_2	59.76	65.12	1.1471	X_1,X_2
عدم وجود الثابت β_0 في النموذج	4.170	56.01	2.7838	X_1,X_2	73.38	69.41	1.878	X_1,X_2,X_3

الاستنتاجات والتوصيات

من خلال النتائج التي تم التوصل إليها يمكننا أن نبين الاستنتاجات والتوصيات التالية :

١- وجود القيم الشاذة في بعض المتغيرات أثرت سلبيا" على نتائج تحليل الانحدار حيث انها قللت من قيمة F-المحسوبة لمعادلة الانحدار التقديرية وكذلك قيمة معامل التحديد وزادت من قيمة متوسط مربعات الخطأ MSe مقارنة مع نتائج التحليل بعد معالجة القيم الشاذة .

٢- إن المتغيرات المدروسة في بحثنا هذا تحتم منطقيا" على ضرورة مرور مستوى الانحدار من نقطة الأصل أي أن لا تحتوي معادلة لانحدار التقديرية على الثابت β_0 ، حيث يلاحظ عند غياب الثلاث متغيرات التوضيحية التي هي (X1,X2,X3) عندئذ يجب أن تكون قيمة متغير الاستجابة يساوي صفرا" أيضا" وهذا بدوره يحتم عدم إدخال الـ β_0 إلى معادلة الانحدار التقديرية ، وهذا فعلا" ما تم الحصول عليه فقد أعطت مثل هذه المعادلة أفضل النتائج مقارنة بتلك التي احتوت على الثابت β_0 . لذا نوصي بالاهتمام في اتخاذ القرار حول احتواء معادلة الانحدار التقديرية على β_0 من عدمه .

المصادر

١-الصالح ، فرح عبد الغني (٢٠٠٤) " دراسة الأخطاء وتأثيرها على نتائج تحليل الانحدار لمتغيرات المواليد الخدج" ، رسالة ماجستير ، كلية علوم الحاسبات والرياضيات ، جامعة الموصل ، العراق.

2- Behnken , D.W. and Draper , N.R.(1972) "Residuals and their Vriance patterns technometrics" , Vol.14 , No.1. PP101-111.

3- Tukey , J.W. (1977) . "Exploratory Data Analysis" Addison – Wesley , reading , MA.

٤- الجبوري ، شلال حبيب (١٩٩٠) " أهمية طريقة اكتشاف وتقدير القيم الشاذة في حالة الانحدار الخطي البسيط" مجلة كلية الإدارة والاقتصاد ، العدد الثاني ، ص ٣١٦-٣٣٥.

٥- الشيخ ، وفاء سيد (١٩٩٩) " تشخيص القيم المتطرفة في نماذج الانحدار وطرق تقديرها " ، رسالة ماجستير ، كلية الادارة والاقتصاد ، الجامعة المستنصرية ، العراق.

٦- الجبوري، شلال حبيب وناسي ، نبيل جورج (٢٠٠٢) " اسلوب جديد لاكتشاف وتقدير المشاهدات الشاذة لنماذج الانحدار البسيط" مجلة العلوم الإحصائية ، العدد الأول ، جامعة الموصل ، العراق ، ص ٣١-٤٢.

٧- الدليمي ، محمد مناجد عيفان (١٩٩٠) " تحليل الانحدار بالأمثلة" ، مطابع التعليم العالي ، كلية الادارة والاقتصاد ، جامعة بغداد.

8- Barnett, V.D. and Leois , T. (1978), "Outliers in statistical data" . John Wiley and Sons , New York.

٩- دنخا ، دلير صليو ، ١٩٩٦ ، "تحديد القيم الشاذة باستخدام الطرق الاستكشافية ومقارنتها مع الطرق المعلمية" ، رسالة ماجستير مقدمة الى مجلس كلية الإدارة والاقتصاد ، جامعة بغداد .

١٠- الجبوري ، منى حسين (١٩٩٨) "دراسة تحليلية للقيم الشاذة والقيم المفقودة لتصميم المربع اللاتيني وتصميم تام في حالة تكرار مشاهدات العينة " ، رسالة ماجستير مقدمة الى مجلس كلية الإدارة والاقتصاد ، الجامعة المستنصرية ، بغداد ، العراق ".

١١- ألنعيمي ، محمد عبد العال والحمداني ، رفاه شهاب وعبد الرزاق ، كيفان عبد اللطيف (١٩٩١) "نظرية الاقتصاد القياسي" ، مطبعة دار الحكمة للطباعة والنشر ، جامعة الموصل ، العراق .

١٢- دبدوب ، مروان عبد العزيز (١٩٩٨) " تقويم بعض طرق التعرف على تعدد العلاقة الخطية في نماذج الانحدار ، مجلة تنمية الرافدين ، كلية الإدارة والاقتصاد ، جامعة الموصل ، العراق ، مجلد ٢٠ ، العدد ٥٣ ، ص ٣٥٣-٣٦٠.

The effective of Outlier and the intercept on the regression analysis results

ALLA ABD – AL STTAR HAMMODAT

Department of Statistics and Informatics , College of computers Sciences and Mathematics , Mosul University,Mosul ,Iraq

(Received 15 / 2 / 2009 , Accepted 25 / 10 / 2009)

Abstract

Some common problems that apper in the regression analysis are outlier values and cross the regression from the origin . We use the box plot graph method to see whether the data have outliers or not , and use the tirmmed mean method to treat those outliers . The (R^2) and (MSE) are considered as a comparative criteria between the regression analysis results when the data have outliers cross from the origin and if not.