www.mecsj.com\ar

الإخلال بافتراض خلو البيانات من القيم المتطرفة وأثره على استخدام تحليل الانحدار الخطي المتعدد في التنبؤ

إعداد

الخثعمي ، عبد العزيز سعد

جامعة أم القرى، المملكة العربية السعودية

كلية التربية – قسم علم النفس

تخصص احصاء و بحوث

E-MAIL : a210066449@gmail.com

العام **1438 -1439 هـ الموافق 2018م**

الملخص

هدف البحث الحالي لدراسة أثر وجود القيم المتطرفة على نموذج الانحدار الخطي المتعدد و دقة التنبؤ ومعالجتها عبر دراسة أثرها على معلمات النموذج . استخدم الباحث المنهج الوصفي على عينة مكونة من 60 طالبا مأخوذة من مجتمع بلغ 204 طالبا هم طلاب كلية الهندسة للمرحلة الجامعية 1436-1437هـ. تم تطبيق نموذج تحليل الانحدار الخطي المتعدد و بعض الطرق و الاختبارات الخاصة بتشخيص و معالجة القيم المتطرفة. فقد تم الكشف عن القيم المتطرفة بالنظر عبر استخدام صندوق ويسكر Box Plot، كما تم استخدام عناصر القطر الرئيسى للمصفوفة Hat Matrix للكشف عن القيم المتطرفة في المتغيرات المستقلة، و باستخدام بواقي ستيودنت المحذوفة Studentized Deleted Residuals للكشف عنها في المتغير التابع .أما تأثير القيم المتطرفة فتم الكشف عنه باستخدام الطرق التالية : مقياس DFFITS، و قياس الأثر على معاملات الانحدار DFBEATS، و مقياس المسافة كوك COOK'S DISTANCE الخاص بقياس الأثر على كل معاملات الانحدار، و مقياس الأثر على الأخطاء المعيارية COVRATIO. قد تمت معالجة القيم المتطرفة باستخدام طرق الحذف و متوسط البتر مع قياس قدرة النموذج قبل و بعد المعالجة عبر استخدام بعضا من المؤشرات و هي : معامل التحديد R^2 و معامل التحديد المعدل R^2_a و احصاءةF و متوسط الخطأ المطلق Mean Absolute Deviation (MAE)، و جذر متوسط مربع الخطأ Root Mean Square Error (RMSE)، و متوسط الخطأ النسبي المطلق Mean Absolute Percentage Error (MAPE)، بالإضافة إلى معامل التساوي لثايلTheil's Inequality Coefficient . من أهم نتائج الدراسة ما يلي : هنالك تأثير معنوي للقيم المتطرفة على معلمات نموذج الانحدار الخطي المتعدد حيث تعمل على تضخيم قيم تباين الأخطاء MSE، وتقلل من قيمة معامل التحديد R^2 و قيمةF. كما توصل البحث إلى أن معالجة القيم المتطرفة بمتوسط البتر أفضل من حذفها.

أوصت الدراسة بعدة بتوصيات أهمها ضرورة تشخيص وتقليل أثر القيم المتطرفة عند تطبيق نموذج تحليل الانحدار للحصول على نموذج أفضل .

الكلمات المفتاحية *: الانحدار الخطي المتعدد - القيم المتطرفة - حذف القيم المتطرفة - متوسط البتر - مصفوفة القبعة - بواقي ستودنت المحذوفة .*

Abstract

The current research objective is to study the presence of outliers on multiple-linear regression model and itsimpact on prediction's accuracy anditshow they are treated via the model's parameters. The qualitative approach was used in this study . A sample of 60 students selected from population of 204 students of Faculty of Engineering for academic year 1436-1437 AH. Multi-linear regression model and some methods and tests of diagnosis and processing the outliers were applied. Accordingly, outlier values were detected visually using Box and Whisker Plot or Box Plot. Main diagonal elements of the Hat Matrix were also used to detect outlier values in the independent variables, and in addition to that using Studentized Deleted Residuals to detect them in the dependent variable.For the effect of outliers, it was detected using the following methods: Difference in fits (DFFITS) measure, Difference of Beta Values (DFBEATS) measure (effect on regression coefficients), Cook's Distance measure for measuring the effect on all regression coefficients, and Covariance ration (COVRATIO) to measure (effect on standard errors).Outlier values were processed using the methods of deleting and amputation average, with measuring the model leverage pre-and post- processing via using some indicators such as : R Square (R^2), R Square Adjusted and F Test, Mean Absolute Deviation (MAE), Root Mean Square Error (MRSE), and Mean Absolute Percentage Error (MAPE), in addition to Theil's Inequality Coefficient. The research concluded with various findings, the most important ones are : There is a significant impact of outlier values on parameters of the multi-linear regression model, where they amplify the Mean Squared Error (MSE), and reduce R^2 value and F value. The research has concluded also that the outlier values processing by amputation average is better than deleting them. The study recommended that : It is necessary to diagnosis and reduce the outliers effect when applying the regression analysis model to obtain better model.

Keywords: multi-linear regression- outliers - deleting the outliers- amputation average- Hat Matrix- Studentized Deleted Residuals.

1 . المقدمة :

في العصر الحديث ومع تطور العلوم وتقدم الخبرة و المعرفة الإنسانية واستخدام الأساليب الكمية الإحصائية المتقدمة، أصبح التنبؤ Prediction أداة فعالة أكثر علمية و دقة في توقع الأحداث المستقبلية مما ساعد على زيادة استعداد الأفراد، و الهيئات للتغيرات المتوقعة في الظواهر، والمجالات المختلفة و التي من ضمنها المجالات التربوية والنفسية. حيث ذكر (Montgomery, et al, 2008)أن التنبؤ له أهمية تمتد لكثير من المجالات التي تتمثل في الأعمال والصناعة والاقتصاد والبيئة والعلوم الطبية والعلوم الاجتماعية والسياسة "كما أشار إلى أن التنبؤ مهم جدا وخاصة في مجال التنبؤ بالأحداث المستقبلية وبالتالي يعد مدخل حاسم في العديد من أنواع عمليات التخطيط وصنع القرار.

و للتنبؤ أساليب عدة، و يعتبر نموذج الانحدار واحداً من أهمها، والذي يبنى على أن المتغير التابع يعتمد على متغيرات مستقلة تفسر سلوكه، حيث يستخدم الانحدار في دراسة العلاقة بين متغير تابع و متغير مستقل (انحدار خطي بسيط) Simple Linear Regression، أو متغير تابع و متغيرات مستقلة (انحدار خطي متعدد) Multiple Linear Regression و الذي يعتبر من أهم أدوات الاحصاء التي تقوم ببناء نماذج احصائية لتقدير العلاقة بين عدة متغيرات بحيث يعطينا نموذج عبارة عن معادلة رياضية توضح العلاقة بين هذه المتغيرات، و هذا ما تؤكد عليه موسوعة علم النفس التربوي حيث ذكر Salkind& Rasmussen (2008) أن أبحاث العلوم النفسية و التربوية عادة ما تكون مهتمة باختبار العلاقة بين متغير أو أكثر و من ثم استخدام هذه العلاقة في التنبؤ بسلوك مستقبلي والانحدار هو معالج نمذجة هذه العلاقة بين المتغيرات.

وبالتالي فأن تحليل الانحدار المتعدد هو واحد من أهم الموضوعات والأساليب الإحصائية و التي لا غنى عنها للباحثين في المجالات العلمية المختلفة ، ومن ثم فإنه يتعين على المهتمين ببناء النماذج بواسطة الانحدار المتعدد الإلمام الكافي بافتراضات نموذج الانحدار الخطي المتعدد، و هو ما يحاول الباحث طرقه وتوضيحه في هذا البحث .

2. مشكلة البحث وتساؤلاته

تنبع أهمية التنبؤ من كونه محاولة لتوقع الأحداث المستقبلية بالشكل الذي يعمل على زيادة استعداد الأفراد والمنظمات للتغيرات المتوقعة في الظواهر و المجالات المختلفة، و التي من ضمنها المجالات التربوية و التعليمية. وبالتالي يتطلب ذلكأن تكونلأساليب القياس والتنبؤ ، أساليباً علمية كمية وهي تلك التي يزودنا بها الإحصاء كعلم للقياس،و تحديدا قياس التنبؤ بالتغيرات التي على ضوئها تُتخذ قرارات المؤسسة، وهنا تحديدا تبرز أهمية التنبؤ والتي تتمثل في ضمان كفاءة المؤسسة (التعليمية / التربوية في مجال البحث) ومرونتها لتحقيق وظائفها التي يتطلع إليها المجتمع ، بالإضافة إلى معرفة احتياجاتها، كما إن التنبؤ يسهم في الحد من المخاطر والمعوقات التي تواجه المؤسسات التربوية و التعليمية أو العاملين بها و المنتمين إليها.

و أحد أهم وسائل التنبؤ بالتغير في الظواهر هو أسلوب تحليل الانحدار في دراسة العلاقة بين متغير تابع والمتغيرات المؤثرة فيه ، و الانحدار المتعدد هو أحد أوسع الطرق انتشارا في دراسة الظواهر النفسية و التربوية، و ذلك لأن السلوك الإنساني على درجة عالية من التعقيد بحيث تتدخل و تؤثر عليه عدة متغيرات في آن واحد. و لأن استخدام الانحدار الخطي المتعدد يتطلب توفر عدد من الافتراضات والشروط الواجب التي يجب التحقق منها، ومن هذه الشروط شرط خلو البيانات من القيم المتطرفة و التي تمثل نطاق هذا البحث ومجاله. و بصورة أكثر دقة فان البحث الحالي يسعى إلى الاجابة عن التساؤل الرئيس التالي:

هل تتأثر دقة تنبؤ الانحدار المتعدد باحتواء البيانات على قيم متطرفة؟، وهل تتحسن الدقة عند معالجة القيم المتطرفة؟

و الذي يتفرع عنه التساؤلات التالية:

1. هل تختلف قيم احصاءات و مؤشرات دقة تنبؤ الانحدار المتعدد عن القيم المعيارية عند احتواء البيانات على قيم متطرفة؟
2. هل تتحسن قيم مؤشرات دقة تنبؤ الانحدار المتعدد وتصبح ضمن مدى القيم المعيارية بعد معالجة القيم المتطرفة؟

أهداف البحث

1. دراسة تأثير وجود القيم المتطرفة على مؤشرات فاعلية التنبؤ لنموذج الانحدار و تأثيرها على مصداقيته.
2. دراسة وضع نموذج الانحدار الخطي المتعدد ومؤشرات فاعلية التنبؤ قبل معالجة القيم المتطرفة و بعد معالجتها ومقارنة النتائج.

أهمية البحث

تنبثق أهمية هذا البحث مع بروز نموذج الانحدار المتعدد كوسيلة تنبؤ إحصائية متقدمة تعطي وتقدم خدمات للباحثين في شتى ومختلف المجالات الحياتية، و لهذا فإن الباحث يأمل أن تكون لهذا البحث أهمية نظرية وأهمية تطبيقية تتمثل في مايلي :

الأهمية النظرية:

تتمثل الأهمية النظرية للبحث في كونها تتعلق بأحد محددات بناء نموذج الانحدار الخطي المتعدد ألا وهو خلو البيانات من القيم المتطرفة ، وما تأثيرها- إن وجدت - على نموذج الانحدار الخطي المتعدد؟ ، وهل سيكون التأثير إيجابي أم سلبي؟ ، وماذا سيتغير في بنية النموذج وقدرته التنبؤية بعد معالجة القيم المتطرفة؟. كما يأمل الباحث في أن يكون هذا البحث إضافة علمية تعمل على توجيه انتباه الباحثين إلى ضرورة أخذ الاحتياطات اللازمة التي تضمن بناء نموذج انحدار دقيق في تنبؤاته.

الأهمية التطبيقية:

يقدّم هذا البحث جانبا تطبيقيا مهما يتمثل في تناوله لطرق اكتشاف القيم المتطرفة و طرق معالجتها ومن ثم المفاضلة بينها و معرفة أي هذه الطرق أفضل من غيرها.

حدود البحث

الحدود الموضوعية : أثر انتهاك افتراض خلو البيانات من القيم المتطرفة على دقة التنبؤ باستخدام الانحدار الخطي المتعدد مع حلول مقترحة

الحدود المكانية : طلاب كلية الهندسة للمرحلة الجامعية بجامعة أم القرى عام 1436-1437هـ

الحدود الزمانية : العام 1438-1439/2018م .

-أسلوب تحليل الانحدار المتعدد لبناء نموذج الانحدار بحسب متغيرات البحث وهي x_1, x_2, x_3, x_4

- مؤشرات فاعلية التنبؤ المتمثلة في : متوسط الخطأ المطلق، و الجذر التربيعي لمتوسط مربع الخطأ، و متوسط الخطأ النسبي المطلق، و معامل التساوي لثايل.

- طرق الكشف عن القيم المتطرفة تمثلت في: صندوق ويسكر و مصفوفة القبعة للمتغيرات المستقلة وبواقي ستيودنت المحذوفة للمتغير التابع.

- أساليب معالجة القيم المتطرفة تمثلت في طريقتي : الحذف ومتوسط البتر.

فروض البحث :

يسعى البحث الحالي للتحقق من الفرضيتين الرئيسيتين التاليتين من خلال الفرضيات الفرعية المرتبطة بكل فرضية :

الفرضية الأولي : تختلف قيم مؤشرات دقة تنبؤ الانحدار المتعدد عن القيم المعيارية عند احتواء البيانات على قيم متطرفة.

الفرضية الثانية : تتحسن قيم مؤشرات دقة تنبؤ الانحدار المتعدد وتصبح ضمن مدى القيم المعيارية بعد معالجة القيم المتطرفة.

مصطلحات البحث

القيم المتطرفة Outliers: يعرفها إسماعيل (2001: 251) بأنها مجموعة قليلة من المشاهدات تبعد قيمها بصورة كبيرة عن باقي المشاهدات في العينة.

الانحدار الخطي المتعدد Multiple Linear Regression:

يعرف زايد (2011: 286) نماذج الانحدار المتعدد بأنها "نماذج تصف العلاقة بين متغير ما يطلق عليه المتغير التابع و متغير أو أكثر يطلق عليها المتغيرات المستقلة أو المفسرة".

التنبؤ Prediction، كما يعرفه عودة و الخليلي (2000: 460) بأنه طريقة علمية في البحث للحصول على المعرفة ، أي أنه يهتم بتقدير بيانات غير معروفة مبنية على بيانات معروفة و ذات صلة بالظاهرة المدروسة.

إجراءات البحث وطرائقه

يتناول الباحث في هذا الجزء الإجراءات المنهجية التي تتمثل في منهج البحث، مجتمع البحث وعينته وكيفية جمع البيانات والأساليب الإحصائية المستخدمة في تحليل وتفسير بيانات الدراسة بغرض الإجابة عن أسئلتها البحثية وفرضياتها، كما يلي:

منهج البحث

استخدم الباحث المنهج الوصفي الذي يصف ويفسر الظواهر المعاصرة للإجابة على تساؤلات بحثه والمتعلقة بدراسة كيفية اكتشاف القيم المتطرفة ودارسة أثرها على نموذج الانحدار الخطي المتعدد، وذلك لأن الهدف الرئيسي من هذا المنهج هو وصف بيانات وخصائص ماهو قيد الدراسة.و يأتي استخدام المنهج الوصفي كونه المنهج المناسب لدراسة أثر القيم المتطرفة على نموذج الانحدار ثم مقارنة هذا التأثير مع النتائج بعد معالجة القيم المتطرفة بالأسلوب الأمثل .

مجتمع وعينة البحث

تكون مجتمع البحث من طلاب المرحلة الجامعية في كلية الهندسة بجامعة أم القرى للعام الدراسي 1436-1437هـ ، والذين بلغ عددهم 204 طالبا توزعوا في تخصصات الهندسة المدنية و الهندسة الميكانيكية و العمارة الإسلامية ، وقد تم الحصول على بياناتهم عن طريق عمادة تقنية المعلومات ، والمتمثلة في تخصصاتهم ومعدلاتهم الفصلية لأربعة فصول دراسية هي المعدلين الفصليين الأول والثاني من العام الدراسي 1436-1437هـ ، و المعدلين الفصليين الأول والثاني من العام 1437-1438هـ ، والمعدل التراكمي لكل طالب.

و قد تم استخدام برنامج SPSS في اختيار عينة عشوائية من مجتمع البحث تمثلت في عدد 60 مفردة بحيث يضمن الباحث وجود قيم شاذة في العينة بما يحقق أهداف البحث ، كما تم بواسطة برنامج SPSSالحصول على بعض المقاييس الوصفية التي تصف طبيعة بيانات البحث والعينة في جدول (1):

جدول (1) بعض المقاييس الاحصائية الوصفية لبيانات البحث

المتغيرات / المقياس	معدل الفصل الأول عام 1436-1437 X_1	معدل الفصل الثاني عام 1436-1437 X_2	معدل الفصل الأول عام 1437-1438 X_3	معدل الفصل الأول عام 1437-1438 X_4	المعدل التراكمي Y
المتوسط الحسابي	2.98	2.945	2.98	3.15	2.96
الوسيط	3	3.08	3.11	3.18	2.99
الانحراف المعياري	0.57	0.52	0.54	0.49	0.41
أقل قيمة	1.76	1.48	1.72	2.05	1.9
أكبر قيمة	3.97	3.95	3.96	4	3.93

اجراءات البحث

تم الحصول على بيانات العينة المستخدمة في هذا البحث من عمادة تقنية المعلومات بجامعة أم القرى وذلك لتطبيق الأسس النظرية والأدبيات المتعلقة بهذا البحث ، حيث بدأ الباحث في اختيار عينة عشوائية حجمها 60 طالبا ثم بدأ في التحقق من توفر افتراضات نموذج الانحدار الخطي المتعدد في البيانات باستثناء القيم المتطرفة، و من ثم تقدير معالم نموذج الانحدار الخطي المتعدد ، ثم قام الباحث بتقييمها في ضوء القيم المعيارية والمؤشرات التي في التساؤل الأول، تم بعد ذلك الكشف عن القيم المتطرفة بالاساليب المختلفة، وبعد تحديدها تمت معالجتها ومن ثم التحقق من توفر افتراضات نموذج الانحدار الخطي المتعدد في البيانات ومن ثم تقدير معالم النموذج مجددا و تقييمها في ضوء القيم المعيارية السابقة ثم قارن الباحث النتائج قبل المعالجة بها بعد المعالجة.

المعالجة الإحصائية لبيانات البحث :

استخدم الباحث البرامج SPSS , Eviews , Exeal لتحليل بيانات البحث وذلك بغرض الاستفادة من كامل امكانات كل برنامج مما لا يتوفر في غيره حتى يتم إظهار نتائج البحث بصورة وافية . وقد تم استخدام أسلوب الانحدار الخطي المتعدد عبر إدخال كافة المتغيرات دفعة واحدة ومن ثم تطبيق المعالجات الإحصائية التالية :

- معامل التحديد ويتم حسابه عبر المعادلة التالية :

$$R^2 = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST} \quad \text{...(1)}$$

- معامل التحديد المعدل R^2_a ويتم حسابه عبر المعادلة التالية

$$R_a^2 = 1 - \left[\frac{(n-1)}{n-(k+1)}\right]\left(\frac{SSE}{SS_{yy}}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$= 1 - \left[\frac{(n-1)}{n-(k+1)}\right](1-R^2) = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

- جولد فيلد كواندت للكشف عن تجانس تباين حد الخطأ العشوائي عبر حساب F

- اختبار دربن-واتسون (D-W) للاستقلال الذاتي للبواقي بالقانون

$$d = \frac{2\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

- اختبار جارك بيرا و يحسب من القانون التالي

$$Jarque - Bera = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

و لقياس مؤشرات فاعلية التنبؤ تم استخدام المؤشرات والأساليب التالية

- متوسط الخطأ المطلق Mean Absolute Error(MAE):

$$MAE = \frac{\sum_{i=1}^{t}|y - \hat{y}_i|}{n} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

- الجذر التربيعي لمتوسط مربع الخطأ Root Mean Square Error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y}_i)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

- متوسط الخطأ النسبي المطلق mean absolute percentage error (MAPE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

- معامل التساوي لثايل Theil's Inequality Coefficient:

$$U = \sqrt{\frac{\sum_{i=1}^{n-1}(P_i - A_i)^2}{\sum_{i=1}^{n-1} A_i^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

و لاكتشاف القيم المتطرفة تم استخدام الأساليب التالية

- الرسم الصندوقي

- مصفوفة Hat Matrix لاكتشاف القيم المتطرفة في المتغيرات المستقلة
- بواقي ستودنت لاكتشاف القيم المتطرفة في المتغير التابع بالقانون

$$d^{*}=e_i\left(\frac{n-P-2}{RSS(1-h_{ii})-e_i^2}\right)^{1/2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9)$$

و لقياس أثر القيم المتطرفة على القيم المقدرة تم استخدام المقاييس التالية

- DFFITS ، DFBETAS ، مقياس كوك ، COVRATIO

4. النتائج

تناول الباحث دراسة تأثير وجود القيم المتطرفة على مؤشرات فاعلية التنبؤ لنموذج الانحدار الخطي المتعدد وتأثيرها على مصداقيته ، بالإضافة الى دراسة وضع نموذج الانحدار الخطي المتعدد ومؤشرات فاعلية التنبؤ قبل معالجة القيم المتطرفة و بعد معالجتها ومقارنة النتائج. وذلك عبر صياغة عدد من التساؤلات والفروض .

أولا - عرض نتائج الإجابة عن التساؤل الأول

1. هل تختلف قيم احصاءات و مؤشرات دقة تنبؤ الانحدار المتعدد عن القيم المعيارية عند احتواء البيانات على قيم متطرفة؟، واختبار فرضيته التي تنص على " تختلف قيم مؤشرات دقة تنبؤ الانحدار المتعدد عن القيم المعيارية عند احتواء البيانات على قيم متطرفة"

قام الباحث بأجراء تحليل الانحدار المتعدد لتقدير معلمات النموذج وتم استخراج معادلته ، كما سيرد فيما يلي :

تحليل الانحدار المتعدد قبل معالجة القيم المتطرفة :

تقدير معلمات النموذج :

باستخدام برنامج SPSS تم اجراء تحليل الانحدار المتعدد على بيانات العينة وبموجبه تم الحصول على قيم المعلمات β_0 ، β_1 ، β_2 ، β_3 ، β_4، وبالتالي تكون معادلة الانحدار على الصورة التالية :

$$\hat{Y} = 0.257 + 0.248\ X_1 + 0.142\ X_2 + 0.258\ X_3 + 0.246\ X_4$$

و تم الحصول على قيم معامل الارتباط R و معامل التحديد R^2 و معامل التحديد المعدل R^2_a و إحصاءهF و مربع مجموع الاخطاء MSE في جدول (2) التالي :

جدول (2) معاملات النموذج قبل المعالجة

الاحصاءة– المؤشر	قيمتها
معامل الارتباط R	0.96
معامل التحديد R^2	0.935
معامل التحديد المعدل R^2_a	0.93
F	197.101
مربع مجموع الاخطاء MSE	0.012

اختبار المعنوية الكلية للنموذج :

تم اختبار معنوية النموذج المقدر كاملا باستخدام اختبار F(ANOVA)حيث بلغت قيمتها (197.101) ودالة إحصائياً عند مستوى المعنوية (0.01) ، وبناءً على ذلك نرفض الفرض الصفري القائل بأن نموذج الانحدار غير معنوي، وهذا يدل على أن هنالك تأثيرا معنويا من المتغيرات المستقلة على المتغير التابع.

اختبار معنوية معاملات الانحدار β_i::

جدول (3) قيم المعنوية لمعاملات الانحدار

معاملات الانحدار	قيم المعاملات	مستوى المعنوية
β_0	0.257	0.015
β_1	0.248	0.000
β_2	0.142	0.001
β_3	0.258	0.000
β_4	0.246	0.000

ويلاحظ من الجدول أعلاه أن معنوية المعلمات β_0، β_1 و β_2و β_3 و β_4 دالة لأنها جميعا أصغر من 0.05 و هذا يعني أن جميع المتغيرات المستقلة كلا على حده يؤثر على المتغير التابع Y.

اختبار القدرة التفسيرية للنموذج :

للحكم على القدرة التفسيرية للنموذج تم حساب معامل التحديد R^2 ومعامل التحديد المعدل R^2_aمن برنامج SPSSحيث بلغت قيمتها : (0.935) و (0.93) على التوالي ، مما يعني أن المتغيرات المستقلة المضمنة في النموذج تفسر ما نسبته تزيد على 93% من التغيرات في المتغير التابع وأن أقل من 6% من التغيرات ترجع لعوامل أخرى منها الخطأ العشوائي.

اختبار النموذج من الناحية القياسية :

بعد تقدير النموذج واختباره احصائيا ومعرفة قدرته التفسيرية تم اختباره من الناحية القياسية التي توضح سلامة الاختبارات الاحصائية وتحدد درجة الثقة بها و يشمل هذا الجزء اختبار النموذج من عدة نواحي تتمثل في اختبار تجانس تباين الخطأ العشوائي ، واختبار الارتباط الخطي المتعدد بين المتغيرات المستقلة ، و اختبار الاستقلال الذاتي للبواقي وذلك بحسب الاجراءات التالية :

اختبار تجانس تباين حد الخطأ العشوائي: تم استخدام اختبار جولدفيلد - كواندت للكشف عن تجانس تباين حد الخطأ العشوائي واختبار الفرضية الصفرية التي تنص على أن هنالك تجانس في تباين الخطأ كما يلي:

1. تم ترتيب المشاهدات تصاعديا حسب المتغير X_1لأنه الأكثر معنوية وتأثيرا بحسب قيمة Beta =0.339
2. تم استبعاد 20% من المشاهدات في المنتصف
3. تبقى لدينا عدد24 = (n-m)/2 في كل مجموعة
4. باستخدام برنامج SPSS تم حساب مجموع مربعات الخطأ SSE لكل مجموعة فكانت النتيجة كالتالي:-

$SSE_1 = 0.165, SSE_2 = 0.304$

5. نحسب قيمة F:

$$V_1 = V_2 = \frac{(n-d)}{2} - p - 1 = \frac{60 - 12}{2} - 4 - 1 = 19$$

$$F = \frac{0.304/19}{0.165/19} = 1.8$$

وهي قيمة Fالمحسوبة، وبمقارنتها بـ$F_{0.05,19,19}$الجدولية ` 2.12 = نجد أن Fالمحسوبة أقل من Fالجدولية وبموجبه نقبل الفرض الصفري القائل بأن هنالك تجانس أو ثبات في تباين الأخطاء.

اختبار الارتباط الخطي المتعدد بين المتغيرات المستقلة

تم إيجاد قيم معامل تضخم التباين VIF و قيم التسامح Tolerance و قيمة رقم الحالة من SPSS ، وعرضت النتائج في جدول(4) التالي:

جدول (4)نتائج اختبارات الكشف عن الارتباط الخطي المتعدد بين المتغيرات المستقلة للنموذج المقدر قبل معالجة القيم المتطرفة

المتغيرات Variables	الثابت	معدل الفصل الأول عام 1437-1436 X_1	معدل الفصل الثاني عام 1437-1436 X_2	معدل الفصل الأول عام 1438-1437 X_3	معدل الفصل الثاني عام 1438-1437 X_4
قيم معامل التضخم VIF	-	1.931	2.394	3.047	2.211
قيم التسامحTolerance	-	0.518	0.426	0.328	0.452
القيم الكامنة Eigenvalues	4.94	0.019	0.015	0.013	0.006
رقم الحالةCI	1	16.2	18.15	19.28	29.8
مؤشر الحالةCN	$CN = \sqrt{\frac{4.955}{0.006}} = 28.69$				

يلاحظ من الجدول (4) أن قيم VIF لجميع المتغيرات المستقلة أقل من 10 ، كما إن قيم Tolerance لجميع المتغيرات المستقلة أكبر من 0.1 ، كذلك رقم الحالة Condition Number أعطى القيمة CN=28.69 وهي قيمة أقل من 30 إلى 100 ، وهذه جميعها مؤشرات دالة على عدم وجود ارتباط خطي بين المتغيرات .

ثالثا: اختبار الاستقلال الذاتي للبواقي

تم استخدام اختبار (Durbin-Watson) للحكم على مدى وجود استقلال ذاتي بين بواقي النموذج ، ومن خلال برنامج SPSS تم استخراج قيمة D-W = 2.095 وبالرجوع إلى جدول Durbin-Watson بعدد متغيرات مستقلة P=4 و n=60 مشاهدة كانت قيمتي D-W العليا والدنيا d_U = 1.74 و d_L = 1.53 . وحيث أن قيمة D-W المحسوبة = 2.095 وتقع في منطقة القبول فإنا نقبل الفرض الصفري القائل بعدم وجود ارتباط ذاتي، أي أن البواقي مستقلة ذاتيا

اختبار التوزيع الطبيعي للبواقي

أولا: اختبار جارك بيرا Jarque-Bera

تم استخراج قيمة هذا الاختبار بواسطة برنامج EView9 حيث بلغت قيمة اختبار جارك-بيرا قبل معالجة القيم المتطرفة 4.067 عند مستوى المعنوية (0.13)وهي أكبر من مستوى المعنوية المعتمد في هذا البحث وهو 0.05وعليه نقبل الفرض الصفري القائل بأن البواقي تتفق مع معامل الالتواء .

ثانيا: اختبار كولمجروفسميرونوف& اختبار ويلك شابيرو:

في هذا الاختبار تم دراسة توزيع البواقي للنموذج المقدر قبل اكتشاف البيانات المتطرفة ومعالجتها، باستخدام برنامج SPSSحيث تم الحصول على النتائج المعروضة في جدول (5) التالي :

جدول (5)نتيجة اختباري كلومجروفسميرونوف& ويلك شابيرو

اختبار كلموجروفسميرونوف			اختبار شابيرو ويلك		
الاحصاءة	درجات الحرية df	مستوى المعنوية Sig	الاحصاءة	درجات الحرية df	مستوى المعنوية Sig
0.124	60	0.02	0.945	60	0.00

ويلاحظ من الجدول أعلاه أن مستويات الدلالة لكلا الاختبارين أقل من 0.05 وبالتالي فإننا نرفض الفرض الصفري ونقبل الفرض البديل والذي يعني أن البواقي لا تتوزع طبيعيا.

والنتيجة أعلاه تختلف عن اختبار جاك بيرا، ومن الجدير ذكره أن أدبيات الإحصاء تشير إلى أنه ليس بالضرورة أن تتفق جميع الاختبارات على نتيجة واحدة، و إنما يمكن استخدام نتائجها ليدعم بعضها بعضا، وهذا ما سوف نلاحظه لاحقا حيث ستتحسن قيمة اختبار جارك بيرا إيجابا بارتفاعها بعد معالجة القيم الشاذة.

خامسا: اختبار فاعلية التنبؤ للنموذج:

لحساب مؤشرات فاعلية التنبؤ المتمثلة في متوسط الخطأ المطلق Mean Absolute Error(MAE) من الصيغة (2-23) ، و الجذر التربيعي لمتوسط مربع الخطأ Root Mean Square Error من الصيغة (2-24) ، و متوسط الخطأ النسبي المطلق mean absolute percentage error (MAPE) من الصيغة (2-25) ، و معامل التساوي لثايل Theil's Inequality Coefficient من الصيغة (2-26) ، وذلك لمقارنتها لاحقا مع النموذج الجديد بعد كشف القيم المتطرفة ومعالجتها، ولتحقيق هذا الغرض نستخدم برنامج Eview9 للحصول على القيم التالية للمؤشرات المذكورة :

Root Mean Square Error (RSME) = 0.106097

Mean Absolute Error (MAE) = 0.084503

Mean absolute percentage error (MAPE) = 2.8917

Theil's Inequality Coefficient = 0.017748

وكل هذه القيم تمثل مؤشرات مأخوذة من معادلة النموذج الأساسية قبل اجراءات الكشف عن القيم المتطرفة وطرق حلها. وهذه النتائج أجابت عن التساؤل الأول حيث نلاحظ اختلاف قيم المؤشرات المختلفة عن القيم المعيارية لكل مؤشر .

ثانيا ـ عرض نتائج الإجابة عن التساؤل الثاني

للإجابة عن هذا التساؤل والذي ينص على " هل تتحسن قيم مؤشرات دقة تنبؤ الانحدار المتعدد وتصبح ضمن مدى القيم المعيارية بعد معالجة القيم المتطرفة؟" واختبار فرضه الذي ينص على " تتحسن قيم مؤشرات دقة تنبؤ الانحدار المتعدد وتصبح ضمن مدى القيم المعيارية بعد معالجة القيم المتطرفة" تم تشخيص القيم الشاذة في المتغيرات المستقلة والمتغير التابع، وتم الكشف عن تأثيرها ومعالجتها.

تشخيص القيم المتطرفة في المتغيرات المستقلة والمتغير التابع و أثرها ومعالجتها:

استخدام صندوق ويسكر Pox Plot في اكتشاف القيم المتطرفة:

وهو عبارة عن رسم صندوق يشتمل على حدود للقيم الطبيعية فاذا تخطت القيم هذه الحدود تعتبر قيم شاذة ويعتمد هذا الصندوق في رسمه علي الربيعات المتمثلة في الوسيط (الربيع الثاني) Q_2 والربيع الأدنى (الربيع الاول) Q_1 والربيع الأعلى (الربيع الثالث) Q_3 ويتم اضافه قيمة الي الربيع الثالث وهي (Q_3-Q_1) وطرح هذه القيمة من الربيع الادنى وبذلك نكون كونا الصندوق، ويمكن الحصول على رسم المتغيرات كلا على حدة مباشرة من SPSS كما يلي : معدل الفصل الأول 1436-1437 (X_1) ، و معدل الفصل الثاني 1436-1437 (X_2)،ومعدل الفصل الأول 1437-1438 (X_3) ، و معدل الفصل الثاني 1437-1438 (X_4) ، والمعدل التراكمي (Y).

و تبين من الأشكال اكتشاف قيمة متطرفة واحدة هي القيمة رقم 39 في المتغير المستقل$\mathbf{X_2}$وهذا يدل على ضعف صندوق ويسكر في اكتشاف القيم المتطرفة، ولهذا سنحتاج إلى طرق تشخيص أخرى مساندة لهذا الأسلوب.

استخدام مصفوفة Hat Matrix في اكتشاف القيم المتطرفة في المتغيرات المستقلة:

بواسطة برنامج EView تم استخراج قيم مصفوفة القبعة Hat Matrix ثم عن طريق بعض المعالجات والعمليات الحسابية بواسطة برنامج Exel وفقا لمعيار بيلسلي المقدرة قيمته ب (0.166667)، حيث تم إيجاد الفروق بينه وبين قيم مصفوفة القبعة ، يتضح وجود (5) من القيم المتطرفة في المتغيرات المستقلة كما موضح بالجدول رقم (7) في ملاحق الدراسة الرئيسية: وبذلك تكون القيم المتطرفة في المتغير المستقل هي الحالات 13 ، 38 ، 39 ، 47 ، 49 .

استخدام بواقي ستودنت لاكتشاف القيم المتطرفة في المتغير التابع:

تم إيجاد القيم المتطرفة في المتغير المتغير التابع باستخدام بواقي ستودنت بالصيغة عن طريق برنامج Eview 9، كما أشرنا إليها في ملاحق الدراسة الرئيسية جدول رقم (8)، حيث تمت مقارنة القيمة المطلقة لبواقي ستودنت مع قيمة tبايجاد الفرق بينهما ثم تحديد القيمة الموجبة والتي مثلت الحالة المتطرفة، و هي الحالة 34 .

قياس أثر القيم المتطرفة على القيم المقدرة قبل المعالجة:

تمت دراسة تأثير هذه القيم باستخدام مقاييس DFFITS، DFBETAS، مقياس كوك، COVRATIO، ومن خلال التطبيق تبين أن القيم المتطرفة تؤثر على المقاييس كما في جدول رقم (9) المرفق في جداول الدراسة الرئيسية . كما يتضح النتائج بالجدول كثرة نتائج المقاييس المتأثرة بالقيم المتطرفة، حيث نلاحظ عمق التأثير للحالات الحالات 13 ، 38 ، 39 ، 47 ، 49 وبالأخص الحالات 34 و 38 و 47 والتي سنشاهد عمق تأثيرها بعد المعالجة لاحقا في تحسن المعاملات والمؤشرات المختلفة.

معالجة القيم المتطرفة:

تم معالجة هذه القيم بطريقتين: أولا إما بحذف القيمة المتطرفة، أو استخدام متوسط البتر

أولا ـ معالجة القيم المتطرفة بواسطة الحذف

www.mecsj.com\ar

تم معالجة القيم المتطرفة بواسطة الحذف لكل من عبارة من العبارات الموضحة أعلاه وجاءت النتائج كما هو مبين بالجدول رقم (7) أدناهـ ، حيث بينا ف يه المفاضلة بين نتائج النماذج المقدرة بعد الحذف :

جدول (7)المقارنة بين النماذج المقدرة بعد حذف القيم المتطرفة

الحالة المحذوفة	β_0	β_1	β_2	β_3	β_4	F	R^2	MSE
قبل الحذف	0.257	0.248	0.142	0.258	0.246	197.101	0.93	0.012
حذف مشاهدة13	0.250	0.248	0.145	0.253	0.250	191.752	0.934	0.012
حذف مشاهدة38	0.251	0.259	0.133	0.293	0.211	198.448	0.936	0.012
حذف مشاهدة39	0244	0.246	0.148	0.258	0.246	185.073	0.932	0.012
حذف مشاهدة47	0.257	0.248	0.142	0.258	0.246	197.101	0.935	0.012
حذف مشاهدة49	0.244	0.243	0.153	0.251	0.250	193.515	0.935	0.012
حذف مشاهدة34	0.233	0.262	0.136	0.257	0.249	223.381	0.94.	0.011
حذف مشاهدة34و38و47	0.216	0.286	0.141	0.290	0.194	228.237	0.95	0.011

ومن النتائج بالجدول أعلاه يتضح أنه وباستخدام طريقة الحذف كان حذف المشاهدة 34 مؤثر بشكل كبير حيث أصبحت قيمةF = 223.381 و كان R^2 = 0.94 و MSE = 0.011 .

أما عندما تم حذف المشاهدات 34 و 38 و 47 معا كان التأثير عميقا بشكل واضح حيث أصبحت قيمةF = 228.237 و R^2 = 0.95 و MSE = 0.011 .

ونستنتج من هذا أن حذف المشاهدات 34 و 38 و 47 كان الأكثر تأثيرا وأدى إلى تحسن كبير في نتائج النموذج وذلك بتقليل قيمة MSE والذي كان قبل الحذف يساوي 0.012 وأصبح بعد الحذف يساوي 0.011 ، وهو فرق قد يبدو ضئيلا ولكنه في الحقيقة فرق جوهري لأن هذا البحث يبحث في معدلات طلابية يؤثر فيها الجزء من مئة.

كذلك أدى حذف المشاهدات المذكورة إلى رفع قيمة F حيث كانت قبل الحذف تساوي 197.101 واصبحت بعد الحذف تساوي 228.237 ، بالاضافة إلى رفع قيمةR^2 من 0.93 إلى 0.95.

ومما تقدم نستنتج أن أفضل النماذج المقدرة بعد معالجة القيم المتطرفة بطريقة الحذف هو النموذج الممثلة معاملاته في جدول (7) كما جاءت ذي المعادلة التالية:

$$\hat{Y} = 0.216 + 0.286\ X_1 + 0.141\ X_2 + 0.290\ X_3 + 0.194\ X_4$$

ثانيا ـ معالجة القيم المتطرفة بواسطة متوسط البتر

تم معالجة القيم المتطرفة بواسطة البتر لكل من عبارة من العبارات الموضحة في السابق وجاءت النتائج كما هو مبين بالجدول رقم (8) أدناهـ ، حيث بينا فيه المفاضلة بين نتائج النماذج المقدرة بعد البتر :

جدول (8)المقارنة بين النماذج المقدرة بعد معالجة القيم المتطرفة بمتوسط البتر

الحالة المحذوفة	β_0	β_1	β_2	β_3	β_4	F	R^2	MSE
قبل المعالجة	0.257	0.248	0.142	0.258	0.246	197.101	0.93	0.012
معالجة مشاهدة13	0.250	0.248	0.145	0.253	0.250	195.213	0.934	0.012
معالجة مشاهدة38	0.251	0.259	0.133	0.293	0.211	202.058	0.936	0.012
معالجة مشاهدة39	0.244	0.246	0.148	0.258	0.246	188.421	0.932	0.012
معالجة مشاهدة47	0.250	0.275	0.153	0.253	0.234	199.369	0.935	0.012
معالجة مشاهدة49	0.244	0.243	0.153	0.251	0.250	197.012	0.935	0.012
معالجة مشاهدة34	0.233	0.262	0.136	0.257	0.249	227.302	0.95	0.011
معالجة مشاهدة34و38و47	0.216	0.286	0.141	0.290	0.193	241.049	0.95	0.010

ومن النتائج بالجدول أعلاه نلاحظ أنه وباستخدام طريقة متوسط البتر كانت معالجة المشاهدة 34 مؤثرة بشكل كبير حيث أصبحت قيمةF = 227.302 و كان R^2 = 0.95 و MSE = 0.011 .

أما عندما تم حذف المشاهدات 34 و 38 و 47 معا كان التأثير عميقا بشكل جلي حيث أصبحت قيمةF = 241.049 و R^2 = 0.95 و MSE = 0.010 .

ونستنتج من هذا أن معالجة المشاهدات 34 و 38 و 47 كان الأكثر تأثيرا وأدى إلى تحسن كبير في نتائج النموذج وذلك بتقليل قيمة MSE والذي كان قبل المعالجة يساوي 0.012 وأصبح بعدها يساوي 0.010 ، وهو فرق قد يبدو ضئيلا ولكنه في الحقيقة فرق جوهري لأن هذا البحث يبحث في معدلات طلابية يؤثر فيها الجزء من مئة.

كذلك أدى حذف المشاهدات المذكورة إلى رفع قيمة F حيث كانت قبل المعالجة تساوي 197.101 واصبحت بعد المعالجة تساوي 241.049 ، بالاضافة إلى رفع قيمةR^2 من 0.93 إلى 0.95

ومما تقدم نستنتج أن أفضل النماذج المقدرة بعد معالجة القيم المتطرفة بطريقة متوسط البتر هو النموذج التالي:

$$\hat{Y} = 0.216 + 0.286\ X_1 + 0.141\ X_2 + 0.290\ X_3 + 0.193\ X_4$$

جدول (9) أفضل نموذج مقدر بعد معالجة القيم المتطرفة بمتوسط البتر

الاحصاءة	قيمتها
معامل الارتباط R	0.97
معامل التحديد R^2	0.95

معامل التحديد المعدل R^2_a	0.942
F	241.049
مربع مجموع الاخطاء MSE	0.010

وتم عرض نتائج المقارنة بين نتائج الجداول السابقة لاختيار أفضل نموذج في جدول (10):

جدول (10) المقارنة بين النماذج لاختيار أفضل نموذج

الحالة	β0	β1	β2	β3	β4	F	R^2	MSE
قبل المعالجة بالحذف او متوسط البتر	**0.257**	**0.248**	**0.142**	**0.258**	**0.246**	**197.101**	**0.93**	**0.012**
حذف مشاهدة34و38و47	**0.216**	**0.286**	**0.141**	**0.290**	**0.194**	**228.237**	**0.95**	**0.011**
البتر مشاهدة34و38و47	**0.216**	**0.286**	**0.141**	**0.290**	**0.193**	**241.049**	**0.95**	**0.010**

ومن الجدول السابق يتضح أن أفضل نموذج يمكن استخدامه للتنبؤ بقيم المتغير التابع هو النموذج الذي تم الحصول عليه بعد معالجة القيم المتطرفة بطريقة متوسط البتر للمشاهدات رقم 34 و 38 و 47 كما يلي:

$$\hat{Y} = 0.216 + 0.286\ X_1 + 0.141\ X_2 + 0.290\ X_3 + 0.193\ X_4$$

وقيم معاملاته في جدول(28) السابق ، حيث نلاحظ في النموذج المذكور تسجيل أعلى قيمة لإحصاءة F = 241.049 ، كما يلاحظ ارتفاع قيمة R^2 = 095 ، وانخفاض قيمة MSE = 0.010

تشخيص أفضل نموذج تم تقديره:

بعد أن تم اختيار أفضل نموذج تم تقديره وقبل استخدامه للتنبؤ بقيم المتغير التابع لابد من القيام باختبار هذا النموذج من الناحية الاحصائية والقياسية ، وذلك للتأكد من أن معالجة المشاهدات الشاذة لم تخلق أي مشكلات إحصائية أو قياسية جديدة في النموذج المقدّر ، وللتأكد من ملاءمة النموذج قياسيا و احصائيا .

اختبار معنوية النموذج احصائيا:

إن مدى ملائمة نموذج الانحدار واستخدامه للتنبؤ يعتمد على أهمية (معنوية) المقدرات β_0، β_1، β_2، β_3، β_4 ويتم لهذا الغرض استخدام اختبار ت والذي يستخدم عندما يكون تباين المتغير التابع مجهولا ويمكن الحصول على قيمه عبر SPSS كما هي في جدول(11)

جدول(11) نتائج اختبار t لأفضل نموذج مقدر

	قيم المعاملات Beta	قيم اختبار"ت" المحسوبة	مستوى المعنوية .Sig

β_0	-	2.306	0.025
β_1	0.372	8.399	0.000
β_2	0.167	3.522	0.001
β_3	0.364	6.22	0.000
β_4	0.225	4.081	0.000

ويتضح من الجدول أعلاه أن كل معلمات النموذج معنوية حيث أنها أقل من مستوى المعنوية 0.05 المعتمد في هذا البحث بما في ذلك معنوية الحد الثابت β_0.

اختبار تجانس تباين حد الخطأ العشوائي:

باستخدام اختبار جولدفيلد - كواندت للكشف عن تجانس تباين حد الخطأ العشوائيوتطبيق الخطوات السابق ذكرها عبر ترتيب المشاهدات تصاعديا حسب المتغير X_2 لأنه الأكثر معنوية وتأثيرا بحسب قيمة بيتا، ومن ثم استبعاد 20% من المشاهدات في المنتصف ومن ثم حساب مجموع مربعات الخطأ SSE للمجموعتين تم الحصول على

$SSE_1 = 0.145$ ، $SSE_2 = 0.164$

$$F = (0.145/19)/(0.164/19) = 0.993$$

وهي قيمة Fالمحسوبة، وبمقارنتها بـ$F_{19,19,0.05}$الجدولية 2.12 = نجد أن Fالمحسوبة أقل من Fالجدولية وبموجبه نقبل الفرض الصفري القائل بأن هنالك تجانس أو ثبات في تباين الأخطاء

اختبار الارتباط الخطي المتعدد بين المتغيرات المستقلة:

باستخدام برنامج SPSS تم حساب قيم التسامح و قيم عامل التضخم وذلك لاكتشاف وجود الارتباط الخطي المتعدد، وكانت النتائج كما في جدول (12)

جدول(12) اكتشاف وجود الارتباط الخطي المتعدد

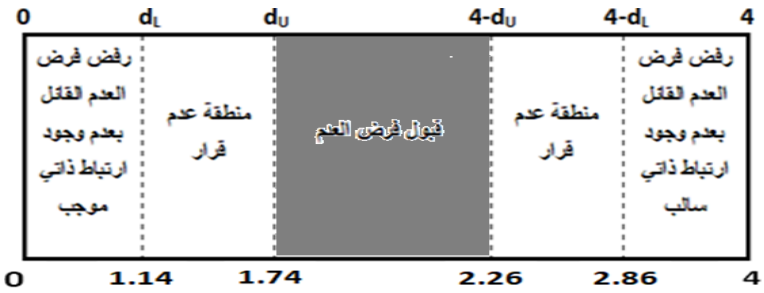
المتغيرات Variables	الثابت	معدل الفصل الأول عام 1437-1436 X_1	معدل الفصل الثاني عام 1437-1436 X_2	معدل الفصل الأول عام 1438-1437 X_3	معدل الفصل الثاني عام 1438-1437 X_4
قيم عامل التضخم VIF	-	2.005	2.294	3.494	3.097
قيم التسامح Tolerance	-	0.499	0.436	0.286	0.323
القيم الكامنة Eigenvalue	4.955	0.016	0.014	0.010	0.006
رقم الحالة CI	1	16.2	18.15	19.28	29.8
مؤشر الحالة CN	$CN = \sqrt{\frac{4.955}{0.006}} = 28.7$				

يتضح من الجدول السابق (12) أعلاه أن قيم VIF لجميع المتغيرات المستقلة أقل من 10 وبالتالي لا يوجد ارتباط خطي متعدد بين المتغيرات المستقلة، كما إن قيم Tolerance لجميع المتغيرات المستقلة أكبر من 0.1 وهذا مؤشر أيضا و دليل على عدم وجود الارتباط الخطي المتعدد بين المتغيرات المستقلة ، كذلك رقم الحالة Condition Number أعطى القيمة CN=28.7 وهي قيمة أقل من 30 إلى 100 وبالتالي فإن جميع هذه المؤشرات تشير إلى عدم وجود ارتباط خطي .

اختبار الاستقلال الذاتي للبواقي:

باستخدام اختبار (Durbin-Watson) للحكم على مدى وجود استقلال ذاتي بين بواقي النموذج، ومن خلال برنامج SPSS تم استخراج قيمة D-W = 2.11 وبالرجوع إلى جدول Durbin-Watson بعدد متغيرات مستقلة P=4 و n=50 مشاهدة كانت قيمتي D-W العليا والدنيا $d_U = 1.67$ و $d_L = 1.42$ ومن ثم ستكون مناطق القبول والرفض كما في جدول (32) التالي:

جدول (13) قرار دربن واتسن



وحيث أن قيمة D-W = 2.11 المحسوبة تقع داخل المنطقة المظللة على الشكل أعلاه فإنا نقبل الفرض الصفري القائل بعدم وجود ارتباط ذاتي، أي أن البواقي مستقلة ذاتيا

اختبار التوزيع الطبيعي للبواقي:

بتطبيق اختبار جارك بيرا على النموذج المقدر نلاحظ أن قيمة الاختبار هي 3.7913 وأن مستوى المعنوية لها = 0.150 وهو أكبر من مستوى المعنوية 0.05 المعتمد في هذا البحث، وبموجبه نقبل الفرض الصفري القائل بأن البواقي تتوزع طبيعيا، وتجدر الملاحظة هنا بأن القيمة السابقة لهذا الاختبار قبل معالجة القيم المتطرفة كانت 4.067 وهي أكبر من القيمة بعد المعالجة وكانت بمستوى معنوية = 0.13 قبل المعالجة وهو أصغر منه بعد المعالجة وهذا يعطي مؤشر على أن قيمة الاختبار تتحسن بعد المعالجة حيث تصغر قيمته ويتضخم مستوى المعنوية ، وهذا يعني أن القيم المتطرفة تؤثر في توزيع البواقي طبيعيا.

اختبار القدرة التنبؤية للنموذج:

مجددا وعبر Eviews تم إيجاد قيم مؤشرات فاعلية التنبؤ التالية

Root Mean Square Error (RSME) = 0.09

Mean Absolute Error (MAE) = 0.07

Mean absolute percentage error (MAPE) = 2.5

Theil's Inequality Coefficient = 0.01

وبالمقارنة مع المؤشرات قبل المعالجة يلاحظ أن المؤشرات زاد اقترابها من الصفر بعد معالجة القيم المتطرفة وهذا مؤشر على تحسن النموذج

جدول(14)المقارنة بين مؤشرات فاعلية التنبؤ قبل وبعد المعالجة بمتوسط البتر

مؤشر فاعلية التنبؤ	قبل المعالجة	بعد المعالجة
الجذر التربيعي لمتوسط مربع الخطأ RMSE	0.106	0.09
متوسط الخطأ المطلق MAE	0.084	0.07
متوسط الخطأ النسبي المطلق MAPE	2.98	2.5
معامل التساوي لثايل	0.02	0.01

وهذا يدل على تحسن فاعلية النموذج بعد معالجة القيم المتطرفة حيث يلاحظ انخفاض في قيم المؤشرات نحو المعدلات المعيارية بعد المعالجة عنها قبل المعالجة.

قياس أثر القيم المتطرفة على القيم المقدرة بعد المعالجة بمتوسط البتر:

نلاحظ عدم ظهور قيم شاذة في المتغير التابع بعد المعالجة كما نلاحظ انخفاض عدد القيم المتطرفة في المتغيرات المستقلة إلى ثلاثة قيم فقط وهي 2 ، 16 ، 49 وهي قيم غير مؤثرة بحسب ما تم اجراءه من حسابات قبلية على النموذج قبل المعالجة ، بينما اختفت الشواذ من المتغير التابع.

هذا بدوره أدى إلى انخفاض عدد القيم المقدرة المتأثرة إلى عدد أقل منه قبل المعالجة ، ونلاحظ هذا من المقارنة بين مجموع حالات التأثر في الجدولين ، بالإضافة إلى أنه وكما تقدم ليست عملية التأثير بتلك الحدة لأن الشواذ المذكورة ليست عميقة التأثير.

جدول(15)المقارنة بين حالات التأثر بحسب المقياس قبل وبعد المعالجة

النموذج	القيم الشاذة في المتغير المستقل	القيم الشاذة في المتغير التابع	مقياس DFFITS لدراسة تأثير القيم الشاذة على القيم المقدرة	مقياس كوك Cook لدراسة تأثير القيم الشاذة على معاملات الانحدار	مقياس COVRATIO لدراسة تأثير القيم الشاذة على الأخطاء المعيارية	مقياس DFBETAS لدراسة تأثير القيم الشاذة على معاملات الانحدار كلا على حدة				
						β_0	β_1	β_2	β_3	β_4
قبل المعالجة	٥	١	٣	٣	٨	٤	٥	٣	٣	٤
بعد المعالجة	٣	٠	١	١	٥	٤	٣	٢	٣	٣

يوضح الجدول رقم (15) أعلاه انخفاض عدد القيم المتأثرة بالقيم المتطرفة بعد المعالجة ، وهذا يدل كذلك على تأثير القيم المتطرفة على المعاملات المختلفة للنموذج وكيف يقل هذا التأثير بعد المعالجة ، فمن الجدول نشاهد على سبيل المثال أن تأثير مقياس COVRATIO الذي يدرس تأثير القيم الشاذة على الأخطاء المعيارية قد أعطى قراءة ل 8 حالات قبل المعالجة وانخفضت بعد المعالجة إلى 5 حالات وهكذا في باقي المقاييس.

ثالثا –اختبار فروض البحث :

باستعراض كل ما تقدم من أرقام ومعالجات خلص الباحث إلى الإجابات التالية على فروض البحث :

أولا : اختلاف قيم مؤشرات دقة تنبؤ الانحدار المتعدد عن القيم المعيارية عند احتواء البيانات على قيم متطرفة :

أ - تختلف قيمة إحصاءةF عن القيمة المعيارية عند احتواء البيانات على قيم متطرفة.

ب - تختلف قيمة معامل التحديد R^2 عن القيمة المعيارية عند احتواء البيانات على قيم متطرفة.

ج - تختلف قيمة معامل التحديد المعدل R^2_a عن القيمة المعيارية عند احتواء البيانات على قيم متطرفة.

د - تختلف قيمة إحصاءةMSE عن القيمة المعيارية عند احتواء البيانات على قيم متطرفة.

هـ - تختلف قيمة معامل التساوي لثايلTheil's Inequality Coefficient عن القيمة المعيارية $\boldsymbol{U \rightarrow 0}$ عند احتواء البيانات على قيم متطرفة.

و - تختلف قيم الجذر التربيعي لمتوسط مربع الخطأ Root Mean Square Errorعن القيمة المعيارية **RMSE < 0.05** عند احتواء البيانات على قيم متطرفة.

ز - تختلف قيم متوسط الخطأ المطلق Mean Absolute Errorعن القيمة المعيارية **MAE** $\rightarrow$ **0** عند احتواء البيانات على قيم متطرفة

ثانيا : تتحسن قيم مؤشرات دقة تنبؤ الانحدار المتعدد وتصبح ضمن مدى القيم المعيارية بعد معالجة القيم المتطرفة :

أ - تتحسن قيمة إحصاءةF وتصبح ضمن مدى القيم المعيارية بعد معالجة القيم المتطرفة

ب - تتحسن قيم مؤشر معامل التحديد R^2 وتصبح ضمن مدى القيم المعيارية بعد معالجة القيم المتطرفة

ج - تتحسن قيم مؤشر معامل التحديد المعدل R^2_a وتصبح ضمن مدى القيم المعيارية بعد معالجة القيم المتطرفة

د - تتحسن قيم إحصاءةMSE وتصبح ضمن مدى القيم المعيارية بعد معالجة القيم المتطرفة

هـ - تتحسن قيمة مؤشر معامل التساوي لثايلTheil's Inequality Coefficient وتصبح ضمن مدى القيم المعيارية $\boldsymbol{U \rightarrow 0}$ بعد معالجة القيم المتطرفة

و - تتحسن قيمة الجذر التربيعي لمتوسط مربع الخطأ Root Mean Square Errorوتصبح ضمن مدى القيمة المعيارية $RMSE < 0.05$ بعد معالجة القيم المتطرفة

ي - تتحسن قيمة متوسط الخطأ المطلق Mean Absolute Error وتصبح ضمن مدى القيمة المعيارية $MAE \rightarrow 0$ بعد معالجة القيم المتطرفة

ويوضح جدول (36) التالي المقارنة بين المؤشرات قبل وبعد المعالجة باستخدام الطريقة الأفضل للمعالجة التي توصل لها هذا البحث ،وهي متوسط البتر

جدول (16) مقارنة بين احصاءات ومؤشرات النموذج قبل وبعد المعالجة

الاحصاءة– المؤشر	قبل المعالجة	بعد المعالجة بمتوسط البتر
معامل التحديد R^2	0.93	0.95
معامل التحديد المعدل R^2_a	0.93	0.942
F	197.101	241.049
مربع مجموع الاخطاء MSE	0.012	0.010
الجذر التربيعي لمتوسط مربع الخطأ RMSE	0.106	0.09
متوسط الخطأ المطلق MAE	0.084	0.07

متوسط الخطأ النسبي المطلق MAPE	2.89	2.5
معامل التساوي لثايل	0.02	0.01

ومنه نلاحظ ازدياد قيمة معامل التحديد R^2 و معامل التحديد المعدل R^2_a وقيمة F بعد المعالجة عنها قبل المعالجة ، كما نلاحظ انخفاض مربع مجموع الاخطاء MSE بعد المعالجة عنه قبل المعالجة ، كما إن كافة مؤشرات فاعلية التنبؤ المستخدمة والمتمثلة في الجذر التربيعي لمتوسط مربع الخطأ RMSE و متوسط الخطأ المطلق MAE و معامل التساوي لثايل قد تحسنت قيمها واقتربت من القيم المعيارية بعد المعالجة عنها قبل المعالجة ، ومن الجدير ذكره أنه قد يبدو للناظر بأن الفروقات طفيفة ولا تذكر ، لكنه في الحقيقة فرق جوهري لأن هذا البحث يبحث في معدلات طلابية يؤثر فيها الجزء من عشرة وكذلك الجزء من مئة.

5.مناقشة النتائج

النتائج التي خلصت إليها الدراسة الحالية تتفق مع نتائج الدراسات السابقة التي خلص إليها كلا من جيلبرت(1992)Gilbert في كون القيم المتطرفة ذات تأثير على خصائص الإنحدار ، كما يتفق أيضا مع دراسة دبدوب و يونس (2006) التي خلصت إلى أن القيم المتطرفة تؤثر على تركيبة النموذج ونتائج تحليل الانحدار ، كما يتفق كذلك مع ما قدمه الغنام (2007) والذي توصل إلى أن معاملات الإنحدار تتأثر بالمشاهدات الغير مألوفة (القيم المتطرفة) ، كما اتفق مع ما قدمه تشوي (2009)Choi في أن وجود القيم المتطرفة يمكن أن يعطي نتائج مضللة ، كما أضاف هذا البحث الكشف عن الشواذ بطريقة بواقي ستودنتبالاضافة إلى ما أوصت به دراسة حسن و رضا (2011) التي استخدمت طريقة البوكس بلوت و مصفوفة القبعة H . بينما خلص هذا البحث إلى نتيجة مختلفة نسبيا عن ما توصل إليه يحي(2009) و قويدر (2016) في بحثيهما كلا على حدة واللذين خلصا إلى نتيجة متقاربة هي أن حذف المشاهدات المتطرفة أثر في دقة معلمات النموذج بانخفاض متوسط مربعات الخطأ وعامل تضخم التباين وارتفاع قيمة F ومعامل التحديد R^2 وتغيير إشارات بعض المعلمات المقدرة والوصول إلى نموذج انحدار يمثل بيانات الدراسة خير تمثيل لدى يحي (2009) ، و في أن حذف القيم المتطرفة قد أثر بشكل إيجابي على نتائج التحليل وذلك من حيث تقليل قيمة الخطأ المعياري في التقدير و زيادة قيمة معامل التحديد لدى قويدر(2016) ، غير أن البحث الحالي توصل إلى أن طريقة متوسط البتر أفضل من طريقة حذف المشاهدات المتطرفة وتثبت هذا النتائج المتوصل إليها والمقارنة المعقودة بين الطريقتين التي قدمها هذا البحث، كما تختلف مع أمانثونج و سوتي(2009)Ampanthong & Suwattee في أن جميع المقاييس التي تم استخدامها لقياس أثر القيم المتطرفة في المعاملات قد تفاوتت في اكتشاف الأثر بشكل متقارب .

الاستنتاجات :

وبالعودة إلى النتائج والأرقام المعروضة آنفا يستنتج الباحث ما يلي :

1. وجود تأثير معنوي من المتغيرات المستقلة مجتمعة على المتغير التابع وهذا يتفق مع الدراسات السابقة

2. أظهر الرسم الصندوقي Pox Plot قدرة ضعيفة على اكتشاف القيم المتطرفة في متغيرات البحث ويظهر للباحث أن السبب في ضعف هذا الأسلوب هو كونه أسلوب بصري بالدرجة الأولى ولعدم شمول العمليات الحسابية فيه كل بيانات العينة وهو ما لم تأخذه بعض الدراسات السابقة في الاعتبار.

3.أظهرت عناصر القطر الرئيسي للمصفوفة Hat Matrix و بواقي ستودنت المحذوفة قدرات أكبر في اكتشاف القيم المتطرفة في المتغيرات المستقلة والمتغير التابع على التوالي ، وذلك لأنها تعتمد طرق حسابية تأخذ في الاعتبار جميع مفردات بيانات العينة بدون استثناء .

4. للقيم المتطرفة تأثير معنوي على معلمات نموذج الانحدار الخطي المتعدد حيث تعمل على تضخيم قيم تباين الأخطاء MSE وتقلل من قيمة R^2 و قيمة F وهذا يعني أن القيم المتطرفة تقلل من جودة النموذج وتزيد من أخطاء بياناته.

5.لاحظ الباحث في بعض العينات التي أجرى عليها بحثا إضافيا أن وجود القيم المتطرفة يعطي نتائج مظللة فيما يخص الارتباط الخطي حيث أظهر التحليل الإحصائي للعينة محل البحث عدم وجود ارتباط خطي في النموذج قبل المعالجة ، بخلاف ما أظهره فحص النموذج المقدر بعد المعالجة والذي أظهرت نتائجه وجود ارتباط خطي ، وهو مالم تشر له أيا من الدراسات السابقة .

6.تتحسن مؤشرات فاعلية التنبؤ بعد معالجة القيم المتطرفة عنها قبل المعالجة وهذا يتفق مع الدراسات السابقة

7. يقل كثيرا حجم التأثير بعد معالجة القيم المتطرفة ويتضح هذا من خلال مقاييس الأثر المختلفة والتي توضح قلة عدد المفردات المتأثرة بعد المعالجة عنها قبل المعالجة .

8. توصل البحث إلى أن معالجة القيم المتطرفة بمتوسط البتر أفضل من حذفها ويظهر للباحث أن السبب هنا هو أن استبدال القيم المتطرفة بقيم متوسطات البتر تحفظ حجم العينة كعدد مفردات مما يزيد قيمة F و R^2 ، بخلاف الحال عند حذف القيم المتطرفة التي تقلل مفردات العينة وبالتالي تقلل قيم F و R^2.

6. الخلاصة

تناول الباحث في هذه الورقة البحثية الإخلال بافتراض خلو البيانات من القيم المتطرفة وأثر ذلك على استخدام الانحدار الخطي المتعدد في التبنؤ. ونظراً لأهمية التنبؤ بالأحداث المستقبلية في كافة المجالات التربوية والاجتماعية والاقتصادية في حياة الناس وأثر ذلك في اتخاذ القرارات الصحيحة وعمليات التخطيط ، فقد جاءت هذه الدراسة لمناقشة مدى أهمية التخلص من القيم المتطرفة في البيانات المتعلقة بالأحداث المستقبلية والتوقعات بشأنها. تم في هذا البحث استخدام العديد من الأساليب الإحصائية من أجل التخلص من ظاهرة القيم المتطرفة في البينانات، حيث بينت نتائج الدراسة مدى دقة التنبؤ باستخدام تحليل الانحدار بعد معالجة القيم المتطرفة، حيث تحسنت قوة المؤشرات التفسيرية لنموذج الانحدار المتعدد.

7.التوصيات

اعتمادا على النتائج التي تم التوصل إليها في هذا البحث يمكن التوصية بما يلي :

1. ضرورة استخدام الطرق المختلفة للكشف عن وجود القيم المتطرفة قبل استخدام نموذج الإنحدار.
2. ضرورة تقليل أثر القيم المتطرفة عند تطبيق نموذج تحليل الانحدار للحصول على نموذج أكثر دقة و فاعلية.
3. يفضل استخدام طريقة متوسط البتر لمعالجة القيم المتطرفة في بيانات نموذج الانحدار الخطي المتعدد.
4. عدم الاعتماد على الرسم الصندوقي Pox Plot في تشخيص القيم المتطرفة كلياً وإنما يكتفى الاستئناس به كنظرة مبدئية لاستشفاف وجود قيم شاذة من عدمه
5. استخدام عناصر القطر الرئيسي للمصفوفة Hat Matrix و بواقي ستودنت المحذوفة Studentized Deleted Residuals في اكتشاف القيم المتطرفة في المتغيرات المستقلة و المتغير التابع
6. تقديم المزيد من الدراسات حول موضوع القيم المتطرفة وتأثيراتها المختلفة في بيانات نموذج الانحدار الخطي المتعدد و النماذج الاحصائية الأخرى .

8.البحوث المقترحة

بالنظر إلى حدود البحث ونتائجه فإن الباحث يقترح البحوث التالية :

1. اجراء البحث الحالي باستخدام طرق أخرى لاكتشاف القيم المتطرفة مثل مقياس مسافة مهلانوبيس (Mahalanobis) distance و مقارنته مع الطرق المختلفة : الرسم الصندوقي Pox Plot و مصفوفة القبعة (Hat Matrix)للمتغيرات المستقلة و بواقي ستيودينت المحذوفة Studentized Deleted Residuals للمتغير التابع.

المصادر و المراجع:

إسماعيل، محمد عبدالرحمن (2001). تحليل الانحدار الخطي. الرياض: معهد الإدارة العامة.

الغنام ، محمد طه (2007). المشاهدات غير المألوفة في نموذج الانحدار. تكريت: مجلة تكريت للعلوم الادارية والاقتصادية.

دبدوب، مروان عبدالعزيز؛ يونس، فرح عبدالغني (2006). تأثير القيم الشاذة على نتائج تحليل الانحدار مع التطبيق على المواليد الخدج. الموصل: جامعة الموصل ، مجلة علوم الرافدين .

حسن، تارا احمد ؛ رضا، مهدي صابر (2011). استخدام الانحدار الحصين لإيجاد أنسب نموذج لتمثيل بيانات الأنواء الجوية في مدينة أربيل خلال الفترة 1998-2010. أربيل: مجلة العلوم الاقتصادية والادارية.

يحي، مزاحم محمد (2009). المشاهدات غير العادية في الإنحدار الخطي المتعدد وبعض طرائق تشخيصها مع التطبيق. الموصل: المجلة العراقية للعلوم الاحصائية.

قويدر، عمر قاسم (2016). تحديد القيم الشاذة في بيانات تيمس TIMSS وأثر أسلوب التعامل معها في نتائج التحليلات الإحصائية. رسالة ماجستير غير منشورة. إربد: جامعة اليرموك.

زايد، مصطفى (2011). موسوعة الإحصاء. القاهرة: مطابع الدار الهندسية.

عودة ، أحمد سليمان ؛ الخليلي ، خليل يوسف (2000). الاحصاء للباحث في التربية والعلوم الإنسانية،ربد: دار الأمل.

المراجع الأجنبية

Ampanthong, P., &Suwattee, P. (2009). A comparative Study of Outlier Detection Procedures in Multiple Linear Regression. Hong : Proceedings of the International MultiConference of Engineers and Computer Scientists , Vol.1.1 , IMECS.

Choi, S. (2009). The effect of outliers on regression analysis. Chicago: Regime Type and Foreign direct Investment Quarterly Journal of Political Science , (4) : 153-65.

Montgomery, D., Jennings, C., &Kulahci, M. (2008). Introduction to Time Series Analysis and Forecasting. New Jersey: John Wiley & Som. Inc.

Salkind, N. J., & Rasmussen, K. (2008). Encyclopedia of Educational Psychology. Sage Publications, Inc: Los Angeles.