التنبؤ باستخدام دالة انحدار شبه معلمية (دراسة تطبيقية مقارنة)

رائد قراحسن * مصطفى مظهر رنة**

(الإيداع: 5 آيلول 2019 ، القبول: 8 كانون الأول 2019)

الملخص

تم تطوير العديد من نماذج الانحدار لمُعالجة مسائل التنبؤ، كطريقة المربعات الصغرى، وانحدار النواة، ونماذج الشَّبكات العُصبونيّة وانحدار العملية الغاوصية وكان آخرها نماذج الانحدار شبه المعلمية ، فموضوع تحليل النماذج شبه المعلمية والذي يدمج النماذج المعلمية والنماذج اللامعلمية يلقى اهتماماً واضحاً في معظم الدراسات التي تأخذ طابعاً أكثر تقدماً في عملية التحليل الإحصائي الدقيق الذي يهدف إلى الحصول على مقدرات ذات مستوى عالٍ من الكفاءة.

قُمنا في هذا البحث باقتراح طريقة جديدة شبه معلمية لتحسين التنبُؤ عن طريق دمج نماذج الانحدار المعلمية المتمثلة بطريقة انحدار المربعات الصغرى مع نماذج الانحدار اللامعلمية والمتمثلة بطريقة انحدار العملية الغاوصية، وتأتي الميزة الكبيرة لهذه النماذج في كونها تحتوي على كل الميزات الإيجابية التي يتضمنها النموذج المعلمي واللامعلمي ولوضوح التفاعل بين مكوناتها المعلمية واللامعلمية والتي لاقت قبولاً واسعاً في الدراسات الطبية والاقتصادية والاجتماعية والعلمية الحديثة وذلك بسب حلها لمشكلة السلوك غير المفهوم لبعض المتغيرات الداخلة في الدراسة من جهة وللمرونة العالية التي تتمتع بها هذه النماذج من جهة أخرى.

وتم التحقق من جودة الطريقة المقترحة عبر تطبيقها على بيانات واقعية ومولدة باستخدام أسلوب المحاكاة. كما تم مقارنة هذه الطريقة مع طريقة انحدار المربعات الصغرى وانحدار العملية الغاوصية باستخدام مقاييس دقة التنبؤ (MSE، RMSE، MAPE)، بهدف الوصول لأفضل طريقة لتحسين دقة التنبؤ.

ودلّت نتائج المقارنة أن الطريقة المقترحة تعطي أفضل دقة تنبؤ وأفضل نتائج وذلك لتكرار عدد الأفضلية بالاعتماد على أصغر قيمة من قيم مقاييس الأخطاء المستخدمة وبسبب قدرة منحني الانحدار الممثل لها على ملاءمة وتمثيل البيانات بشكل أفضل.

الكلمات المفتاحية: انحدار المربعات الصغرى ، انحدار العملية الغاوصية، انحدار شبه المعلمي، مقاييس دقة التنبؤ.

* طالب دراسات عليا (دكتوراه)-قسم الإحصاء الرياضي-كلية العلوم-جامعة حلب

**أستاذ مساعد-قسم الإحصاء الرياضي-كلية العلوم-جامعة حلب

prediction using semi parametric regression function (Comparative Applied Study)

Raed Kara Hasan * Moustafa Mazhar Rene **

(Received: 5 September 2019, Accepted: 8 December 2019)

Abstract

Several regression models have been developed to address prediction issues, such as the least-squares method, the kernel regression, neural network models, and Gaussian process regression, the most recent semi parametric regression models, semi parametric methods combined parametric methods and nonparametric methods .It is important in most of studies which take in their nature more progress in the procedure of accurate statistical analysis which aim at getting estimators efficient. In this study, we proposed semi parametric new method to improve prediction by combining the parametric regression models represented by the least squares regression method with the non-parametric regression models represented by the Gaussian process regression. The great advantage of these models is that they contain all the positive features contained in the teacher and non-teacher model and the clarity of the interaction between the components of the teacher and non-teachers, which received wide acceptance in modern medical, economic, social and scientific studies, because of its solution to the problem of incomprehensible behavior of some variables included in the study on the one hand and for flexibility Highly enjoyed by these models on the other. The quality of the proposed method was verified by applying it to realistic and generated data using simulation. This method was also compared with the least squares regression method and Gaussian process regression using the prediction accuracy measures (MSE, RMSE, MAPE) in order to reach the best way to improve the accuracy of the prediction.The comparison that the proposed method gives the best predictive accuracy and better results in order to replicate the number of preference based on the smallest value of the values of the error measures used , because of the ability of the regression curve which ideals have an appropriate and better data representation.

Keywords: least squares Regression, Gaussian Process regression, semi parametric regression, the measurements of prediction error explanation.

*Postgraduate Student (PhD)-Dept. of Mathematical Statistics -Faculty of Science-University of Aleppo

**Assistant Professor-Dept. of Mathematical Statistics-Faculty of Science-University of Aleppo

1- مقدمة: Introduction

يعد تحليل الانحدار من أكثر الطرائق الإحصائية استعمالاً حيث يقوم ببناء نموذج إحصائي للظاهرة المدروسة لفهم العلاقة بين المتغير التابع Y ومجموعة المتغيرات المستقلة X_i من خلال تكوين صيغة رياضية معينة تدعى بنموذج الانحدار التي تقوم بتحديد طبيعة واتجاه العلاقة بين المتغيرات والتي تستخدم أيضاً في التنبؤ، ونظراً لاختلاف وتنوع الظواهر في الواقع العملي فقد وجد أكثر من نوع لنماذج الانحدار تبعاً لطبيعة الظاهرة المدروسة .

فالنوع الأول من النماذج يدعى بنماذج الانحدار المعلمي الذي يعتمد في صياغته على مجموعة من المعالم المجهولة في النموذج والتي يتم تقديرها باستخدام عدة طرائق منها طريقة المربعات الصغرى والإمكانية العظمى وغيرها من الطرائق إلا أن هذا النوع من النماذج لا يعبر عن الظاهرة المدروسة نظراً لسلوك بعض المتغيرات الداخلة في عملية التحليل سلوكاً معلمياً وبعضها الأخر سلوكاً لامعلمياً .

أما النوع الثاني فيدعى بنماذج الانحدار اللامعلمية والذي لا يتقيد بشروط صارمة من حيث التوزيع الخاص لمتغير التابع والخطأ ، ولا يتقيد بدالة معينة تفسر العلاقة بين المتغيرات المستقلة والمتغير التابع ويمتاز هذا النوع بمرونة في التعامل مع البيانات وكذلك سهولة تفسير نتائجه ، إلا أن هذا النوع من النماذج يعاني من مشكلة تعدد الأبعاد والتي تحصل عند زيادة عدد المتغيرات المستقلة في عملية التحليل مما يؤدي إلى تناقص دقة التقدير.

أما النوع الأخير من نماذج الانحدار فيدعى بنماذج الانحدار شبه المعلمية والذي يعد ثمرة التكامل بين نموذجي الانحدار المعلمي واللامعلمي ويمتاز هذا النوع بالمرونة العالية في التطبيق بالمقارنة مع النماذج المعلمية وكذلك فإنه يتخلص من مشكلة تعدد الأبعاد التي تظهر في النماذج اللامعلمية (Ruppert وزملائه ،2003؛ Akkus،2011).

يعدّ أول ظهور لمصطلح شبه معلمي Semiparametric لعام 1980 من قبل الباحثين santner,Brown في مجال الإحصاءات الحيوية Biometric وفي عام 1981 تم استعمال هذا المصطلح من قبل الباحث Finnas وزملائه في مجال الرياضيات والديمغرافية ، وفي عام 1988 قدم الباحث speckman دراسة عرض فيها استعمال ممهد النواة في تقدير النماذج الخطية الجزئية (Speckman ،1988) .

وفي عام 2003 قدم الباحث Millimet وزملاؤه دراسة تضمنت إجراء مقارنة بين نموذج شبه معلمي ونموذج معلمي عند دراستهم مشكلة تلوث الهواء في الولايات المتحدة (Millimet وزملائه ،2003).

وقد حدد Hardle وزملائه في عام 2004 تعريف الانحدار شبه المعلمي بأنه نموذج يحتوي على قسمين أحدهما معلمي بأبعاد نهائية والأخر لا معلمي بأبعاد لا نهائية (Hardle وزملائه،2004).

وفي عام 2011 قام الباحث Aydin بتقديم بحثاً درس فيه طرائق مختلفة لتقدير معلمة التمهيد لنموذج الانحدار الخطي الجزئي شبه المعلمي عند تقدير هذا النموذج باستخدام شرائح التمهيد التكعيبية (Aydin ،2011).

2- أهداف البحث وأهميته:

يتمثل الهدف الرئيسي للبحث في استخدام وتطوير بعض النماذج شبه المعلمية لمعالجة الأبعاد الكبيرة وتحسين عملية التنبؤ عبر اقتراح طريقة جديدة ومقارنتها مع طرائق الانحدار المعلمية واللامعلمية بهدف الوصول لأفضل طريقة لتحسين دقة التنبؤ.

تأتي أهمية البحث من كونه يسلط الضوء على أهمية تطبيق بعض النماذج شبه المعلمية في معالجة مشكلة تعدد الأبعاد التي تظهر في البيانات الحيوية والطبية وتصنيفها كطريقة بديلة للطرائق الإحصائية التقليدية التي تعالج الموضوع نفسه.

3- موادُّ وطرائقُ البحث: Materials and Methods

3-1- فرضية البحث ومشكلته :

يلاحظ في الجوانب التطبيقية للنماذج المعلمية أن أغلب البيانات تتكون من أكثر من متغير مستقل يؤثر على المتغير التابع علماً أن هذا النوع من النماذج له عيوب مختلفة منها المعرفة المسبقة بتوزيع البيانات المدروسة كما أنه في بعض الأحيان قد لا يمثل الدالة قيد الدراسة تمثيلاً كاملاً وذلك لكون بعض المتغيرات المدروسة تسلك سلوكاً معلمياً وبعضها الأخر يسلك سلوكاً لامعلمياً ، وأن عملية اختيار المتغيرات الخاصة بالدراسة قد يتم وفق رؤى معينة حول خاصة معينة ولكن عملية النمذجة قد لا تأخذ بعين الاعتبار مشاكل أخرى كمشاكل تعدد الأبعاد التي تحصل عند زيادة عدد المتغيرات المدروسة في النماذج اللامعلمية والتي تعتبر عملية معقدة حيث بزيادتها تزيد درجة التعقيد للنموذج المدروس .

ومن هنا تتمثل فرضية البحث في إيجاد طريقة جديدة للتنبؤ ومقارنتها مع بعض طرائق الانحدار المعلمية واللامعلمية.

وتكمن مشكلة البحث في استخدام بعض النماذج شبه المعلمية لمعالجة مشكلة الأبعاد الكبيرة التي تظهر عند زيادة عدد المتغيرات المدروسة ونظراً لكون النموذج شبه المعلمي يمتلك مميزات كلا النموذجين المعلمي واللامعلمي لذا فإنه يعد نموذج وسط بين النماذج المعلمية واللامعلمية.

3-2- أنواع نماذج الانحدار:

3-2-1- نماذج الانحدار المعلمية: parametric Regression Models :

تعتبر أساليب تحليل الانحدار من أهم وأقوى أساليب التحليل الإحصائي الذي يُقيم العلاقات بين مجموعة من المتغيرات بغرض الوصول إلى صيغة تصف هذه العلاقات التي تمكننا من التنبؤ عن حصول تغير واحد أو أكثر في ضوء التغيرات الأخرى التي تتعلق بها، أي أن تحليل الانحدار طريق لتوقع نتيجة معينة اعتماداً على متحول أو عدة متحولات مستقلة.

حيث أننا في تحليل الانحدار نجري توافقاً بين النموذج التنبؤي والبيانات المتوفرة لدينا أي أننا سنستخدم البيانات لتقدير نموذج يمكنه أن يصف الظاهرة بشكل جيد، ونستخدم هذا النموذج لنتوقع قيماً للمتحول التابع اعتماداً على متحول أو أكثر من المتحولات المستقلة (التنبؤية)، هذا ويمكننا التنبؤ بأية بيانات اعتماداً على المعادلة العامة التالية:

$$\text{Outcome}_i = \text{model}_i + \text{error}_\text{i} \qquad (1)$$

وهذا يعني أن النتيجة يمكننا الحصول عليها باستخدام نموذج ملائم لبيانات مع إضافة نوع من الخطأ، تتخذ شكل المعادلة وفقاً لنوع العلاقة التي نشاهدها من واقع البيانات الإحصائية الخاصة بهذه المتغيرات والتي يجب أن تتصف بالدقة وذلك حتى يلائم النموذج طبيعة الظاهرة (Izenman، 2008؛ Nielsen، 2009).

تدعى طريقة تحليل علاقة الانحدار معلمية إذا افترضت شكل موصوف بشكل كامل بواسطة مجموعة منتهية من المعالم.

المثال النموذجي للنموذج المعلمي هو معادلة الانحدار كثيرة الحدود عندما تكون المعالم معاملات للمتغيرات المستقلة.

ومن أشهر أنواع هذه النماذج نموذج الانحدار المعلمي الخطي و يوصف نموذج الانحدار المعلمي الخطي بشكل عام وفق الصيغة:

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_i \qquad (2)$$

حيث $\boldsymbol{X}$ و $\boldsymbol{Y}$ تمثل متغيرات عشوائية و $\boldsymbol{f}(\boldsymbol{X_i}, \boldsymbol{\beta})$ دالة خطية للمتغيرات المستقلة $\boldsymbol{X_i}$ ولمعالم مجهولة $\boldsymbol{\beta}$ و $\boldsymbol{\varepsilon}_i$: الأخطاء العشوائية تتوزع وفق التوزيع الطبيعي وتكون مستقلة بمتوسط صفر وتباين ثابت σ^2 أي أن: $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2)$
(Hastie وزملائه، 2009؛ Nielsen، 2009).

3-2-2- نماذج الانحدار اللامعلمية: Non parametric Regression Models:

هي أسلوب إحصائي مغاير لمفهوم الانحدار المعلمي ويتفق معه بالهدف النهائي وهو الحصول على أفضل تقدير لمنحني الانحدار حيث في بعض الحالات لا يستطيع نموذج الانحدار الخطي المعلمي تفسير السلوك الفردي في منحني التوزيع. كما أن الطرائق المعلمية لتقدير منحني الانحدار ليست قادرة دائماً على الحصول على معلومات كافية ، حيث تفترض الطرائق المعلمية توزيع المتغيرات المدروسة معلومة التوزيع، وبما أن هذا الافتراض لا يتحقق في أغلب التطبيقات العملية لأنه لا يأخذ في الاعتبار التأثير اللاخطي للمتغيرات المستقلة أو عدم تجانس التباين، وكذلك كون توزيع الأخطاء ليس توزيعاً طبيعياً وإنما قد يكون ثنائي المنوال Bimodal وهذه النماذج اللاخطية تتصف عادة بكثرة صيغها وعدم محدوديتها مما يولد مشكلة أخرى هي مسألة اختيار الصيغة الأكثر ملاءمة والتي قد تسبب في إدخال الباحث في مسألة تجريب النماذج والصيغ الواحدة تلو الأخرى، لذا كان من المناسب إيجاد طريقة جديدة تأخذ بالاعتبار هذا التأثير، تمثلت باللجوء إلى استخدام الطرائق اللامعلمية والتي تم اقتراحها من قبل الباحث (Jacob Wolfowitz) عام 1942 حيث أن هذه الطرائق لا تضع قيوداً أو افتراضات أو صيغاً خاصة على الدوال ، فلو كانت لدينا ظاهرة من الظواهر المختلفة ولا توجد هناك أي فرضيات تحكم العلاقة بين المتغير التابع y والمتغيرات المستقلة t_i ولا يمكن تحديد أي علاقة سواء كانت خطية أو غير خطية بينها عندها ستكون العلاقة بين المتغير التابع y والمتغيرات المستقلة t_i تسمى بالانحدار اللامعلمي والذي يأخذ النموذج التالي:

$$y_i = \mathbf{g}(\mathbf{t_i}) + \boldsymbol{\varepsilon}_i \ ; i = 1, \dots, \boldsymbol{n} \qquad (3)$$

حيث أن $\mathbf{g}(\mathbf{t_i})$: تمثل دالة التمهيد أو الانحدار اللامعلمية وهي عبارة عن دالة مجهولة يتم تقديرها بالطرائق اللامعلمية. (Ruppert وزملائه، 2003؛ Hardle ، 2004 ؛Pérez وزملائه، 2008) .

3-2-3- نماذج الانحدار شبه المعلمية Semi parametric Regression Models :

إن الانحدار شبه المعلمي هو أسلوب إحصائي يحقق الخصائص العامة للانحدار المعلمي واللامعلمي ويتفق معهما في الغاية نفسها وهي الحصول على أفضل منحني للبيانات ويقترب أو يطابق منحني المتغير التابع بالدمج بين أساليب التقدير المعلمية واللامعلمية وتختلف طرائق تقدير معالم نموذج الانحدار شبه المعلمي باختلاف الصيغة الرياضية ولكنها بصورة عامة تكون بأسلوبين رئيسيين: الأول وهو الأكثر استخداماً من قبل الباحثين ويعني تقدير الجزء المعلمي في المرحلة الأولى بأي طريقة من طرائق التقدير المعلمية المعروفة وبعدها في المرحلة الثانية يتم تقدير الجزء اللامعلمي بأي طريقة من طرائق التقدير اللامعلمية بالاعتماد على تقديرات المرحلة الأولى أما الأسلوب الثاني وهو معاكس للأسلوب الأول حيث يتم تقدير الجزء اللامعلمي في المرحلة الأولى ، وفي المرحلة الثانية يتم تقدير الجزء المعلمي بالاعتماد على تقديرات المرحلة الأولى . إن اختيار الأسلوب الأول (معلمي ولا معلمي) يوفر للباحثين من الناحية المبدئية إمكانية تقدير عدد كبير من المنحنيات المتوافقة مع منحني المتغير التابع y، واختيار الأسلوب الثاني (لامعلمي ومعلمي) سيوفر عدد كبير من النماذج المتوافقة وبالتالي عدد كبير من المنحنيات التقديرية لهذه النماذج .

يعد نموذج الانحدار الخطي الجزئي (Partial Linear Regression Model) من أشهر النماذج شبه المعلمية وهو النموذج الكلاسيكي المعبر عن مفهوم الانحدار شبه المعلمي وذلك لوضوح حالة التفاعل بين المكون المعلمي والمكون اللامعلمي ويرد عند الكثير من الباحثين باسم نموذج انحدار شبه معلمي بسيط ولهذا السبب فإنه كان محط اهتمام الكثير من الباحثين والذي انعكس على الأساليب المتعددة لتقدير معالمه، حيث تم اقتراحه من قبل الباحثان Speckman & Robinson في عام 1988، وهو من النماذج التي تعتمد على متغيرات خطية معلمية وأخرى غير خطية لا معلمية ، إذ أن هذه المتغيرات الخطية واللاخطية تؤثر في المتغير التابع y ويعد هذا النموذج حالة خاصة من النماذج التجميعية (Additive Models) وكذلك يتميز بميزة وهي إمكانية تجنب مشكلة الأبعاد والتي تحدث في النماذج اللامعلمية عند زيادة

عدد المتغيرات المستقلة لذلك يكون أفضل من النماذج اللامعلمية ومن ناحية أخرى هو أكثر مرونة من النماذج المعلمية الخطية القياسية لأنها تقلل من الافتراضات الخطية المفروضة على هذه النماذج.

وسبب تسميته خطي لأنه يتضمن جزئين جزء معلمي خطي وجزء لامعلمي وتربط هذه الأجزاء مع بعضها بعلاقة تجميعية والتي تستند على تقسيم النموذج العام إلى جزئيين معلمي ولا معلمي باعتبار البيانات للجزء الأول لها نموذج معلمي بمعالم مجهولة يتم تقديرها بالطرائق المعلمية في حين للجزء الثاني تعتبر المتغيرات المستقلة متغيرات مستمرة ذات صيغة مجهولة وتمهيدية يتم تقديرها بالطرائق اللامعلمية مع الاشارة إلى أن متغيرات الجزء الأول تكون مستمرة أو متقطعة أو ثنائية (Speckman ، 1988 ؛Ruppert وزملائه ، 2003) والصيغة العامة لهذا النوع من النماذج:

$$y_i = X_i'\beta + g(t_i) + \varepsilon_i \ ; i = 1, \dots, n \qquad (4)$$

ويمكن التعبير عن النموذج الموصوف بالمعادلة السابقة بشكل مصفوفات كما يلي:

$$Y = X\beta + g + \varepsilon \qquad (5)$$

حيث أن:

Y : متجه عمود للمتغير التابع من الدرجة $(n\times 1)$.

X : مصفوفة المتغيرات المستقلة $X_0, X_1, \cdots, X_k$ (المعلمية) من الدرجة $n\times (k+1)$ حيث يحوي قيم الواحد لمثيل المعامل الثابت.

β : متجه المعلمات المجهولة من الدرجة $((k+1) \times 1)$ يحوي على معالم نموذج الانحدار المجهولة $\beta_0, \beta_1, \cdots, \beta_k$ المراد تقديرها.

g: متجه دالة تمهيدية مجهولة من الدرجة $(n\times 1)$.

ε: متجه الأخطاء العشوائية من الدرجة $(n\times 1)$ وتكون مستقلة بمتوسط صفر وتباين σ^2 (Akkus، 2011).

3-3- طرائق تقدير دالة الانحدار:

3-3-1- التقدير باستخدام انحدار المربعات الصغرى:

يُعرَّف نموذج الانحدار الخطي بأنه نموذج انحدار خطي المعالم (كل معلمة من معالمه غير مضروبة أو مقسومة على معلمة أخرى) والآن من أجل الحالة العامة سنفترض أنه لدينا n مشاهدة من المتغير التابع y ولدينا p من المتغيرات المستقلة $x_j \ ; j = 1, \dots, p$ تسمى أيضاً بالمتغيرات المستقلة أو التنبؤية (predictor) أو المتغيرات المنحدرة (regressors). عندما يكون لدينا أكثر من متغير مستقل واحد فإننا نكون بصدد ما يدعى بتحليل الانحدار المتعدد (multiple regression analysis). هنا تنحصر مهامنا بما يلي: 1- تقدير معاملات النموذج β. 2- التنبؤ بالقيمة المتوقعة لـ y مع الإفتراض بأن y تابعة للمعاملات β وليس للمتغيرات المستقلة x_j (والتي نفترض أنها ثوابت كونها جاءت من مجموعة بيانات التدريب). وبفرض أن $n > p + 1$ (أي أن عدد مشاهدات مجموعة بيانات التدريب أكبر من عدد معالم نموذج الانحدار المراد تقديره بما في ذلك المعامل الثابت) من أجل المشاهدة i لدينا

$$\boldsymbol{y}_i = f_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{\boldsymbol{p}}; \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \dots, \boldsymbol{x}_{1\boldsymbol{p}}) + \boldsymbol{\varepsilon}_i$$

$$\boldsymbol{y}_i = \boldsymbol{f}_i(\boldsymbol{\beta}; \boldsymbol{x}_i) + \boldsymbol{\varepsilon}_i \qquad (6)$$

بحيث أن $\beta = [\beta_0, \beta_1, \dots, \beta_p]^T$ و $x_i = [1, x_{i1}, x_{i2}, \dots, x_{ip}]^T$ و $\boldsymbol{\varepsilon}_i$ هي الفرق بين القيمة المتوقعة والقيمة الحقيقية للمشاهدة i.

من الآن وصاعداً سوف نُضمّن β_0 داخل النموذج، وبكتابة كل الـ n معادلة بالشكل المصفوفي نحصل على العلاقة التالية:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{x}_{11} & \boldsymbol{x}_{12} & \cdots & \boldsymbol{x}_{1p} \\ 1 & \boldsymbol{x}_{21} & \boldsymbol{x}_{22} & \cdots & \boldsymbol{x}_{2p} \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ 1 & \boldsymbol{x}_{n1} & \boldsymbol{x}_{n2} & \cdots & \boldsymbol{x}_{np} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

أو يمكن كتابتها بالشكل

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (7)$$

بحيث أن X مصفوفة المدخلات لها الحجم $(n, p+1)$، Y متجه المخرجات له الحجم $(n, 1)$ ، و β متجه معالم نموذج الانحدار له الحجم $(p+1, 1)$، و $\boldsymbol{\varepsilon}$ متجه الأخطاء أو البواقي وله الحجم $(n, 1)$ ولها توقع معدوم $E(\boldsymbol{\varepsilon}) = 0$

والآن لتقدير معالم النموذج لدينا عدة طرق لكن أشهرها هي طريقة المربعات الصغرى الاعتيادية (OLS).

تفترض طريقة المربعات الصغرى الاعتيادية (Ordinary Least squares) بأن مصفوفة التغاير أو التباين لـ Y متناسبة مع المصفوفة الواحدية أي أن جميع البواقي لها نفس التباين $\mathrm{Var}(Y) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$ وهي مستقلة عن بعضها البعض. في هذه الطريقة نأخذ المعاملاتβ ليكون نصف مجموع مربعات البواقي (الأخطاء) أصغرياً

$$\begin{aligned} \boldsymbol{\varepsilon} &= \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta})^2 = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) \qquad (8) \\ &= \frac{1}{2}(\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}) \\ &= \frac{1}{2}(\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}) \end{aligned}$$

بحيث تكون أعمدة المصفوفة X مستقلة خطياً والمشتق الثاني $\frac{d^2(\varepsilon)}{d\beta d\beta^T} = X^TX$ موجب تماماً. وبهذا يكون لـ $\boldsymbol{\varepsilon}$ قيمة أصغرية.إن المصفوفة X^TX هي مصفوفة متناظرة أي أن $X^TX = (X^TX)^T$ ولها الحجم$(p+1, p+1)$ وبالاشتقاق بالنسبة لـ β مع الفرض بأن قيمة هذا المشتق معدوم ($\frac{d(\varepsilon)}{d(\beta)} = 0$) يكون الحل كالتالي

$$\frac{\boldsymbol{d(\varepsilon)}}{\boldsymbol{d(\beta)}} = -\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{X}^T\boldsymbol{X\beta} \qquad (9)$$

ومنه يكون: $\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$ ويصبح مقدر المربعات الصغرى المطلوب كما يلي:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad (10)$$

(Nielsen وزملائه، 2009) .

3-3-2- التقدير باستخدام انحدار العملية الغاوصية:

يستخدم انحدار العملية الغاوصية (Gaussian Process Regression) أو اختصاراً (GPR) في تقنيات التعلم الآلي. قُدِّمَت طَّريقة العملية الغاوصية كأداة للانحدار(Regression) في مجال التعلم الآلي، لأول مرّة من قبل العالمين Rasmussen وWilliams عام 1996 حيث قاموا بوصف تحسين المعلمات في دالة التغاير والتي كانت مستوحاة من استخدام العملية الغاوصية مع الشبكات العصبونية، وقد تم استخدامها في تطبيقات مختلفة مثل التنبؤ بالنفوذية الجلدية من المواد الكيميائية والتنبؤ بتركيز الأوزون في الهواء (Bishop،2007؛Rasmussen وWilliams، 2006).

ليكن لدينا g=$(g_1(.), g_2(.), g_3(.), \ldots, g_d(.))^T$ مُتَّجه ذو d بُعد من الدوال عندئذ تسمى العملية العشوائية $\{g(x) : x \in \chi\}$ بعملية غاوص (بحيث أن χ هو فضاء المدخلات) إذا كان مُتَّجه المتغيرات العشوائية $X_1, X_2, \ldots\ldots, X_d$ يتوزع وفق التوزيع الطبيعي المتعدد بمتوسط $\boldsymbol{\mu}$ و مصفوفة تغاير$\mathbf{K}$، تُعرف عملية غاوص كتوزيع على الدوال$P(g(x))$ بحيث أن$g(x)$ هي دالة معرفة على فضاء المدخلات χ كما يلي: $g: \chi \to \mathbb{R}$

أي أن العملية الغاوصية هي مجموعة من المتغيرات العشوائية المستمرة محدودة الأبعاد والتي كل منها يخضع للتوزيع الطبيعي وتكون جميع توزيعاتها هي توزيعات طبيعية، وتعتبرعملية غاوص (GP) من أهم تقنيات التعلم الآلي (Rasmussen وWilliams،2006؛ Liu وزملاؤه، 2017).

لتكن لدينا $\mu(x)$ دالة متوسط و $\mathrm{k}(x,\mathrm{x}')$ دالة تغاير معرفتان كما يلي:

$$\mu(x) = E[g(x)]$$

$$\mathrm{k}(x,\mathrm{x}') = Cov\big(g(x), g(\mathrm{x}')\big) = E\big[\big(g(x)-\mu(x)\big)\big(g(\mathrm{x}')-\mu(\mathrm{x}')\big)\big]$$

بحيث $x, x' \in \chi$ عندئذٍ العملية الغاوصية (GP) تأخذ الشكل التالي:

$$\begin{bmatrix} g(x_1) \\ \vdots \\ g(x_d) \end{bmatrix} \sim N_d\left(\begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_d) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_d) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_d) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_d,x_1) & k(x_d,x_2) & \cdots & k(x_d,x_d) \end{bmatrix}\right) \qquad (11)$$

ونرمز لذلك بالرمز:

$$P\big(g(x)\big) = \mathcal{GP}(\,\mu(x), k(x,\mathrm{x}')) \qquad (12)$$

نسمي الدالة$k(x,x')$ بدالة التغاير أو دالة النواة (نواة التغاير) وهي دالة موجبة محدودة ولها عدة أنواع (Bishop،2007) ، ليكن Y متغير تابع وX متغيرات عشوائية ذو d بُعد، يعطى نموذج الانحدار اللامعلمي وفق العلاقة:

$$\boldsymbol{y} = g(\mathrm{x}) + \boldsymbol{\varepsilon} \qquad , \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma^2}) \qquad (13)$$

بحيث أن: $g(x)$ هي دالة مجهولة أما في الانحدار المعلمي تكون معلومة، تعاني الطرائق اللامعلمية من مشكلة تعدد الأبعاد (curse of dimensionality) عندما يتم تطبيقها مع المتغيرات المتعددة (أي عندما تكون d كبيرة)، لقد تم تطوير مجموعة متنوعة من النماذج البديلة للتغلب على هذه المشكلة منها نموذج انحدار العملية الغاوصية (GPR) .

إن نموذج انحدار العملية الغاوصية هو نموذج لامعلمي، وهذا يعني بأنه لا يفترض شكل معين للدالة المدروسة ولكن يتم تحديد شكل العلاقة بين المدخلات والأهداف بالكامل من خلال البيانات التي قد تتضمن عدد غير محدود من الدوال، وتكون الدالة الأساسية التي تنتج البيانات مجهولة ولكن يتم توليد التنبؤات من خلال مجموعة من الدوال التي تخضع لتوزيع غاوص في فضاء الدوال، ويعتبر نموذج انحدار العملية الغاوصية من أحدث طرائق التنبؤ، وهو من نماذج بايز الاحتمالية، ففي معظم طرائق انحدار بايز يتم إيجاد معلومات مسبقة عن معلمات النموذج، وبعد ذلك يتم وضع شروط على البيانات لإعطاء معلمات النموذج اللاحق(البعدي)، حيث يمكن صياغة هذه المعلومات المسبقة بشكل توزيع احتمالي يسمى التوزيع القبلي و يحدد نموذج بايز المعلمات المجهولة للنموذج القبلي بينما يحدد نموذج عملية غاوص علاقات الدوال القبلية مباشرة بين مدخلات الاختبار ومدخلات ومخرجات التدريب (Rasmussen وWilliams،2006؛ Liu وزملاؤه، 2017). لنفترض لدينا مجموعة من البيانات $\{(x_i, y_i)\}_{i=1}^{n}$ بحيث تشير $x_i \in \mathbb{R}^d$ إلى المدخلات والتي لها d بُعد وتشير $y_i \in \mathbb{R}$ إلى القيم الحقيقية للنواتج و n إلى عدد البيانات،عندئذٍ يأخذ نموذج انحدار العملية الغاوصية (GPR) الشكل التالي:

$$\boldsymbol{y_i} = g(\boldsymbol{x_i}) + \boldsymbol{\varepsilon_i} \quad ; \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{n} \quad , \boldsymbol{\varepsilon_i} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma^2}) \qquad (14)$$

بحيث أن: $g(x) \sim \mathcal{GP}(\,\mu(x), k(x,\mathrm{x}'))$ و $\mathcal{GP}(\,\mu(x), k(x,\mathrm{x}'))$ هي عملية غاوص القبلية (Gaussian process prior) مع دالة متوسط $\mu(x)$ ودالة تغاير $k(x,\mathrm{x}')$ وبالتالي يعطى نموذج انحدار العملية الغاوصية وفق العلاقة:

$$\boldsymbol{y} = \mathcal{GP}(\,\mu(x), k(x,\mathrm{x}')) + \sigma_n^2\,\delta(x,\mathrm{x}') \qquad (15)$$

بحيث أن: $\delta(x, \mathrm{x}')$ دالة دلتا كرونكير (Kronecker delta) و$\delta(x, \mathrm{x}') = 0$ عندما $x \neq \mathrm{x}'$

و$\delta(x, \mathrm{x}') = 1$ عندما $x = x'$ وσ_n^2 تباين الضجيج العشوائي ومن الشائع أيضاً أن نفترض $\mu(x)$ =0 أي (دالة المتوسط للعملية الغاوصية القبلية معدومة)عندئذٍ يأخذ نموذج انحدار العملية الغاوصية الشكل التالي:

$$\boldsymbol{y} \sim \mathcal{GP}(\,0, k(x, \mathrm{x}')) + \sigma_n^2\,\delta(x, \mathrm{x}') \qquad (16)$$

تم تصميم مجموعة متنوعة من دوال النواة، وسيتم في هذا البحث استخدام دالة النواة الغاوصية والموضحة وفق العلاقة الآتية:

$$k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \qquad (17)$$

بحيث أن: $\|x - x'\| = \sqrt{(x - x')^T(x - x')}$ تشير إلى طويلة الشعاع $(x - x')$ أو نظيم الفرق بين قيمتين

x, x' و σ معامل دالة نواة غاوص (Rasmussen وWilliams،2006).

3-3-3- التقدير باستخدام الانحدار شبه المعلمي:(الطريقة المُقترَحة لتحسين التنبؤ)

يملك كل من نموذجي انحدار المربعات الصغرى الاعتيادية (OLS) وانحدار العملية الغاوصية GPRإمكانات وخواص مُختلفة عند وصف سلوك وسمات منحني الانحدار ضمن الأنماط الخطية وغير الخطية، لذا فإن النموذج المُقترح في هذا البحث يتُكوَّن من مُركبات كلا النموذجين بحيث نستطيع باستخدام النموذج المقترح نمذجة الأنماط المختلفة لنموذج الانحدار وتحسين مُجمل سلوك التنبُؤ.

إن الهدف الرئيسي من استخدام هذه الطريقة المقترحة يتمثل بمحاولة تمثيل النموذج للبيانات بالشكل الصحيح أو حتى قريب من الصحة ومحاولة الابتعاد عن عدم تمثيل المجتمع تمثيلاً غير أمثل ، حيث أن هذه الطريقة تعتمد على كون نموذج الانحدار من النوع المدمج بين كلاً من النموذج المعلمي ذو صيغة معروفة ومجهولة المعالم ضمن الأنماط الخطية ونموذج لا معلمي لدالة انحدار مجهولة الصيغة ضمن الأنماط غير الخطية، وبالتالي يُمكن التعبير عن y_i (مجموعة البيانات الأصلية) كما يلي:

$$\mathbf{y_i} = (\mathbf{1} - \mathbf{u}).\boldsymbol{f}(\boldsymbol{X_i}, \boldsymbol{\beta}) + \mathbf{u}.\mathbf{g}(\mathbf{t_i}) + \boldsymbol{\varepsilon}_\mathbf{i} \qquad ; \mathbf{i} = \mathbf{1}, \dots, \mathbf{n} \qquad (\mathbf{18})$$

حيث تشير: $\mathrm{f}(\mathrm{X_i}, \beta)$ إلى دالة الانحدار المعلمي بمعالم مجهولة وصيغة معلومة وهي دالة خطية.

في حين تشير: $\mathrm{g}(\mathrm{t_i})$ إلى دالة الانحدار اللامعلمية وهي دالة غير خطية .

أما $\mathbf{u}$ فتشير إلى معلمة الدمج بحيث أن: $0 < \mathbf{u} < \mathbf{1}$.

ويتم تقدير $\hat{y}_i$ من خلال مجموعة البيانات المدروسة بثلاث مراحل:

أولاً: يتم تقدير قيم $\boldsymbol{f}(\boldsymbol{X_i}, \boldsymbol{\beta})$ باستخدام طريقة انحدار المربعات الصغرى الاعتيادية (OLS) (أي أننا سنقوم بالتنبُؤ بالبيانات الأصلية التي لدينا باستخدام نماذج انحدار المربعات الصغرى الاعتيادية وذلك فقط من أجل المتغيرات المستقلة الخطية)، وبعدها يتم تقدير معلمة الدمج $\mathbf{u}$ باستخدام طريقة انحدار المربعات الصغرى الاعتيادية (OLS).

ثانياً: يتم تقدير دالة الانحدار اللا معلمية $\mathbf{g}(\mathbf{t_i})$ باستخدام طريقة انحدار العملية الغاوصية GPR (أي أننا سنقوم بالتنبُؤ بالبيانات التي لدينا باستخدام نماذج انحدار العملية الغاوصية وذلك فقط من أجل المتغيرات المستقلة غير الخطية)

ثالثاً: يتم جمع التقديرين اللذان حصلنا عليهما باستخدام نموذجي التنبُؤ المُستخدمين، وبالتالي فإن النموذج المقترح للتنبُؤ بالبيانات هو:

$$\hat{y}_i = (\mathbf{1} - \hat{u}).\boldsymbol{f}(\boldsymbol{X_i}, \widehat{\boldsymbol{\beta}}) + \hat{u}.\,\hat{g}(\mathbf{t_i}) \qquad ; \mathbf{i} = \mathbf{1}, \dots, \mathbf{n} \qquad (19)$$

وبتعويض العلاقتين (10) و(16) في العلاقة (19) تنتج لدينا العلاقة التالية التي تمثل النموذج المقترح التالي:

$$\hat{y}_i = (1 - \hat{u})\,(X^T X)^{-1} X^T Y + \hat{u}\,[\mathcal{GP}(\,0, k(x, \mathrm{x}'))+ \sigma_n^2\, \delta(x, \mathrm{x}')] \qquad (20)$$

بحيث $k(x, \mathrm{x}')$, $\delta(x, \mathrm{x}')$, σ_n^2 قد تم شرحها ضمن الفقرة(**5−2−2−**) مع الإشارة إلى أن$\hat{u}$ يرمز إلى مقدر المربعات الصغرى لمعلمة الدمج وعندما$\hat{u}$(0=)فإن المقدر $\hat{y}_i$ المعطى وفق العلاقة (20) سوف يمثل المقدر المعلمي لدالة الانحدار المعطى وفق العلاقة(10) ،أما عندما ($\hat{u}$=1) فإن المقدر $\hat{y}_i$ سوف يمثل المقدر اللامعلمي لدالة الانحدار المعطى وفق العلاقة(16).

3−4− الجانب التطبيقي:

بغرض اختبار أداء الطريقة المُقترحة في تحسين التنبؤ قُمنا بتطبيق الطريقة المُقترحة شبه المعلمية وطريقتي انحدار المربعات الصغرى وانحدار العملية الغاوصية على مجموعتي بيانات واقعية ومولدة.

3−4−1−البيانات الواقعية:

جمعت البيانات من سجلات أبقار الفريزيان (Friesian) العائدة إلى محطة أبقار جب رملة الواقعة في منطقة الغاب التابعة للمؤسسة العامة للمباقر في محافظة حماة والبالغ عددها **40** بقرة فريزيان ، حيث كانت الأبقار في حظائر مغلقة ، وتتم الحلابة والتغذية داخل الحظائر، وتخضع لنفس الظروف من التغذية والخدمة والرعاية وكان نظام التغذية ثابتاً في المحطة، وعدد مرات الحلابة مرتين يومياً (صباحاً ومساءاً) وكان نظام التلقيح المتبع في المحطة هو التلقيح الاصطناعي.

تمثل مجموعة البيانات الواقعية كمية إنتاج الحليب الفعلي/كغ/ وطول موسم الحليب /يوم/ وطول المدة من الولادة إلى أول تلقيح /يوم/ ومدة الحياة الانتاجية /سنة/ و فترة الجفاف /يوم/ ومتوسط عدد مرات التلقيح ودليل المثابرة على الإنتاج.

يهدف نموذج التَّنبُؤ المراد بناءه إلى تقدير كمية إنتاج الحليب الفعلي للأبقار عن طريق معرفة طول موسم الحليب و طول المدة من الولادة إلى أول تلقيح ومدة الحياة الانتاجية وفترة الجفاف ومتوسط عدد مرات التلقيح ودليل المثابرة على الإنتاج.

بحيث تمثل كمية إنتاج الحليب الفعلي المتغير التابعY، بينما يمثل طول موسم الحليب المتغير الأول المستقل X_1 ويمثل طول المدة من الولادة إلى أول تلقيح المتغير الثاني المستقل $\boldsymbol{X_2}$ ، ومدة الحياة الانتاجية تمثل المتغير الثالث المستقل $\boldsymbol{X_3}$، و فترة الجفاف تمثل المتغير الرابع المستقل $\boldsymbol{X_4}$ ، ويمثل متوسط عدد مرات التلقيح المتغير الخامس المستقل $\boldsymbol{X_5}$، **ويمثل** دليل المثابرة على الإنتاج المتغير السادس المستقل $\boldsymbol{X_6}$.

ومن خلال تمثيل العلاقة بين المتغيرات المستقلة المدروسة لاحظنا بأن المتغيرين المستقلين X_1 وX_2 خطيين وبقية المتغيرات غير خطية حيث أننا في هذا البحث نعالج مشكلة وجود حالة من البيانات فيها جزء خطي وجزء غير خطي ونقوم بتقدير الجزء الخطي بالاعتماد على الطريقة المعلمية والجزء الغير خطي نقوم بتقديره بالاعتماد على الطريقة اللامعلمية

3−4−2−البيانات المولدة:

يوجد العديد من المشاكل التي يصعب وضعها في قالب رياضي سهل الحل وذلك بسبب تعدد وكثرة المتغيرات والقيود فيها، لذلك تستخدم طريقة المحاكاة لإيجاد الحل الأمثل لهذه الحالات وتقوم طريقة المحاكاة على إيجاد الوسيلة التي يستطيع بها الباحث دراسة المشكلة وتحليلها على الرغم من وجود الصعوبات في التعبير عنها بنموذج رياضي، وحتى يتم إجراء المحاكاة لأي نظام لابد أن تتوفر لدينا معلومات كافية عن أجزاء النظام وخصائصه حتى نستطيع فهم النظام والتنبؤ بالطريقة التي يعمل بها (Law وزملائه،2000).

إن التنوع الكبير في النماذج ناتج من تنوع الظواهر التي تمثلها وقد تم اختيار نموذج من بعض النماذج التي تناسب الطرائق المستخدمة في هذا البحث، حيث تم استخدام تجارب المحاكاة في توليد مجموعة من البيانات العشوائية باستخدام دالة معطاة وفق العلاقة (2000، Schimek ؛ Wang وزملائه، 2004) :

$$Y = 1 - X_1 + X_2 - 3\boldsymbol{X_3}^2 + 3\boldsymbol{X_4}^3 \qquad (21)$$

وتم إضافة ثلاثة حالات مختلفة للضجيج العشوائي (تشويش) وفق التوزيع الطبيعي بتوقع رياضي معدوم وبانحراف معياري قدره ($\sigma 0.1=$ ، $\sigma 0.5=$، $\sigma 1=$) وعند الحالات المختلفة لحجوم العينات (n=10, n=50, n=100, n=200) وبتكرار قدره ($\boldsymbol{L = 500}$).

3-4-3- منهجية اختبار الطريقة المُقترَحة:

تعتبر الدّقة معياراً لاختيار النّموذج الأمثل للتَّنبؤ، ويُقصد بالدّقة قُدرة نموذج التَّنبؤ على إعادة إنتاج البيانات الأصل للعينة المدروسة، ومنه فإن الاختيار المُناسب لمقياس دقة التَّنبؤ يؤثر إيجاباً في تحديد فاعلية نموذج التنبؤ المُستخدم وتعمل مقاييس دقّة التَّنبُؤ القياسيّة بشكل عام على مفهوم الفرق بين القيم الأصلية والقيم المُتوقَّعة أو المُتنبئ بها، وهو ما ندعوه بخطَأ التَّنبؤ، وكُلما كان مقدار الفرق قليلاً كانت دالة التنبؤ أفضل وأدقّ، يوجد العديد من مقاييس دقّة التنبؤ، وعادةً لا يتم الاعتماد على مقياس واحد في عمليّةِ ضبط نموذج التنبؤ. تعتبر المقاييس التّالية: MSE, RMSE،MAPE من أفضل مقاييس المقارنة بين نماذج تنبؤ مختلفة تم بناؤها باستخدام نفس مجموعة بيانات التدريب (Hyndman و Koehler، 2006). وقد اعتمدنا في بحثنا هذا على هذه المقاييس كونها تلائم طبيعة البحث ويتم حسابهم كما يلي:

$MSE = mean(e_i^2)$ متوسط مربعات الأخطاء Mean squared error

$RMSE = \sqrt{MSE}$ جذر متوسط مربعات الأخطاء Root mean squared error

$MAPE = mean(|e_i/y_i|)$ متوسط القيم المطلقة للأخطاء النسبية *Mean absolute percentage error*

بحيث يحسب خطأ التنبؤ من العلاقة: $e_i = y_i - \hat{y}_i$ (يُدعى خطأ التنبؤ في بعض المراجع بالبواقي residuals) (Hardle وزملاؤه ، 2004).

وبهدف تطبيق الطرائق الثلاثة تم كتابة برنامج نصي باللغة البرمجية **R (لغة برمجية إحصائية)**، حيث تستطيع هذه اللغة القيام بالعديد من تحليلات البيانات بحيث يتم تنظيم هذه التحليلات ضمن ما يسمى بالحزم Packages مما يعني قدرة الباحثين على تطوير البرامج المختلفة الأمر الذي ساهم بانتشار استخدامها في المجالات الأكاديمية (Cotton، 2013 ؛Matloff، 2011).

لتنفيذ هذا البحث تم الاستفادة من الحزمة kernlab بهدف تطبيق انحدار العملية الغاوصية، في المرحلة الأولى قُمنا بتطبيق نموذج انحدار المربعات الصغرى على مجموعتي البيانات الواقعية والمولدة وتم تقدير المعالم المجهولة و حساب قيم مقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE) للحكم على جودة أداء كل طريقة من طرائق التنبؤ.

وفي المرحلة الثانية قُمنا بتطبيق نموذج انحدار العملية الغاوصية على مجموعتي البيانات الواقعية والمولدة وتم الاعتماد على دالة النواة الغاوصية بالعلاقة (17) كدالة نواة وتم حساب قيم مقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE) للحكم على جودة أداء كل طريقة من طرائق التنبؤ.

وفي المرحلة الثالثة قُمنا بتطبيق النموذج المقترح على مجموعتي البيانات الواقعية والمولدة وتم اعتماد على نفس قيم المعاملات المحسوبة في المرحلتين السابقتين وتم أيضاً حساب قيم مقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE)

ولكن بالنسبة لطريقة انحدار المربعات الصغرى أخذنا المتغيرات المستقلة الخطية فقط في كلا مجموعتي البيانات وأما طريقة انحدار العملية الغاوصية أخذنا المتغيرات المستقلة غير الخطية.

4- النتائج:

سنقوم الآن بعرض النتائج التطبيقية على كل من مجموعتي البيانات:

بعد التطبيق العملي ظهرت لدينا النتائج الموضحة بالجداول والأشكال كما يلي:

الجدول رقم (1) قيم المعاملات المستخدمة ضمن مجموعة البيانات الواقعية

معاملات انحدار المربعات الصغرى							قيمة معلمة الدمج u	معامل انحدار العملية الغاوصية σ_n	معامل دالة النواة الغاوصية σ
b6	b5	b4	b3	b2	b1	b0			
46.3166	377.1494	-2.4632	-91.8679	2.0191	2.9253	0.376	0.675	0.559560847	0.1379639

المصدر: من إعداد الباحث اعتماداً على مخرجات اللغة البرمجية R

بعد تطبيق الطريقة المُقترحة وطريقة انحدار العملية الغاوصية وطريقة انحدار المربعات الصغرى وفق المعاملات الموضحة بالجدول رقم (1) على مجموعة بيانات أبقار الفريزيان الواقعية وحساب قيم مقاييس الأخطاء لكل طريقة ظهرت لدينا النتائج التالية:

الجدول رقم (2): نتائج تطبيق الطرائق الثلاثة على مجموعة البيانات الواقعية

الطريقة المقترحة			طريقة انحدار العملية الغاوصية			طريقة انحدار المربعات الصغرى		
MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE
0.1198439	1.002935	1.005879	0.135126	1.163154	1.352927	0.1264899	1.086605	1.180709

المصدر: من إعداد الباحث اعتماداً على مخرجات اللغة البرمجية R

وبإعادة الحسابات بالنسبة لمجموعة البيانات المولدة باستخدام تجارب المحاكاة وعند الحالات المختلفة لحجوم العينات (n=10, n=50, n=100, n=200) والحالات الثلاثة المختلفة للضجيج العشوائي(σ=0.1 ، σ=0.5، σ=1) ظهرت لدينا النتائج التالية:

الجدول رقم (3) قيم المعاملات المستخدمة ضمن مجموعة البيانات المولدة

حجم العينة n	الانحراف المعياري σ	قيمة معلمة الدمج u	معامل دالة النواة الغاوصية σ	معامل انحدار العملية الغاوصية σ_n	معاملات انحدار المربعات الصغرى			
					b1	b2	b3	b4
10	0.1	0.989	0.6225617	0.132431673	-2.016	3.078	-3.063	3.002
	0.5	0.647	0.6548663	0.029609287	-2.079	3.390	-3.317	3.009
	1	0.577	0.6225617	0.393339303	-2.157	3.781	-3.634	3.019
50	0.1	0.9989	0.5748172	0.050811268	-2.058	3.008	-2.983	3.084
	0.5	0.979	0.574817213	0.205103008	-2.290	3.042	-2.916	3.418
	1	0.879	0.57481721377	0.363951507	-2.580	3.085	-2.833	3.837
100	0.1	0.9892	0.5692009	0.121105095	-1.999	3.027	-3.008	2.972
	0.5	0.997	0.56920092061	0.192488998	-1.994	3.137	-3.041	2.861
	1	0.9987	0.5692009206113	0.386797548	-1.989	3.273	-3.081	2.722
200	0.1	0.8657	0.5664197	0.086746695	-2.032	3.002	-2.992	3.034
	0.5	0.998	0.5664197344075	0.184551292	-2.159	3.008	-2.959	3.169
	1	0.898	0.566419734448	0.396750565	-2.318	3.016	-2.919	3.338

المصدر: من إعداد الباحث اعتماداً على مخرجات اللغة البرمجية R

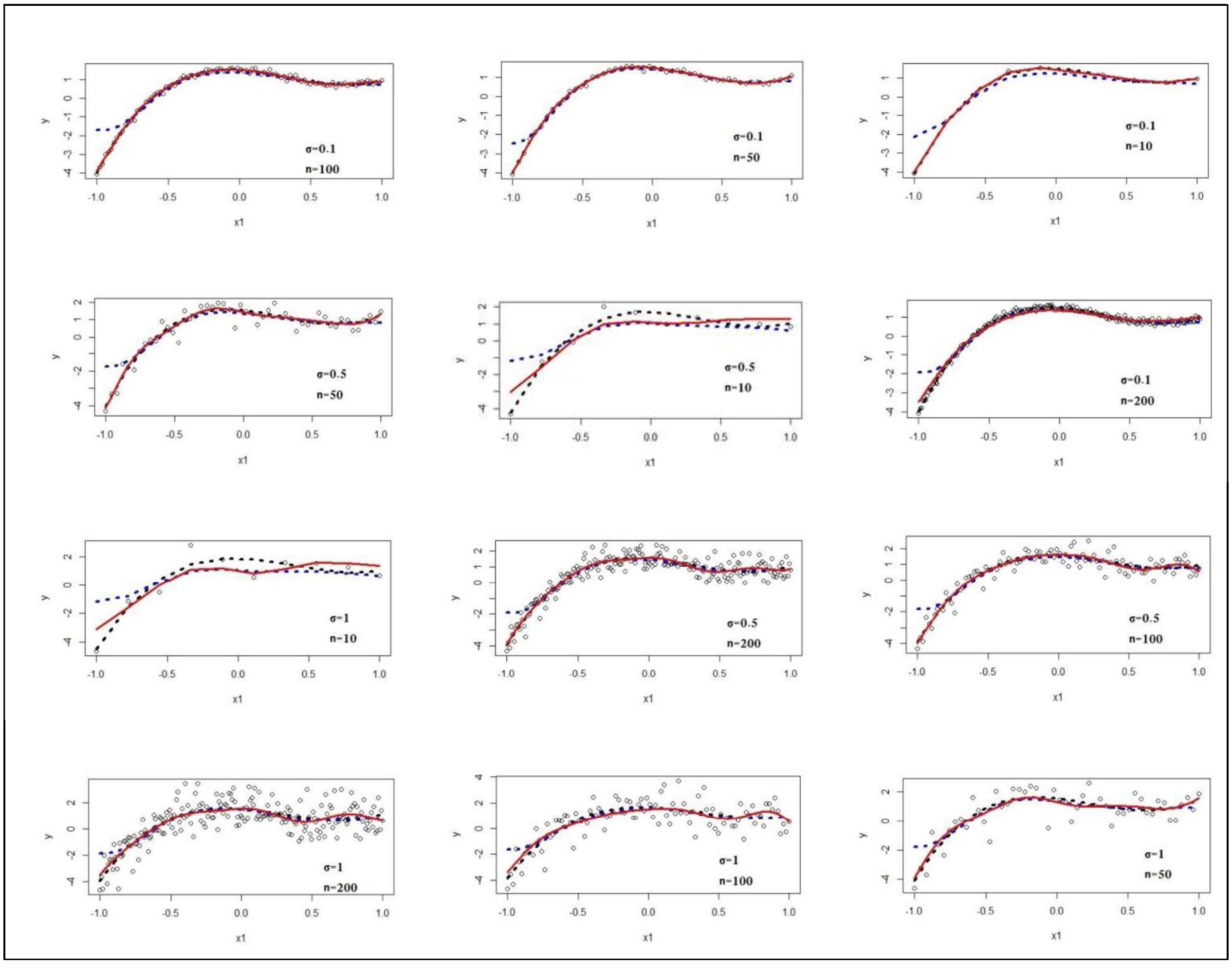
بعد تطبيق الطرائق الثلاثة وفق المعاملات الموضحة بالجدول رقم (3) على مجموعة البيانات المولدة باستخدام تجارب المحاكاة وعند الحالات المختلفة لحجوم العينات (n=10, n=50, n=100, n=200) والحالات الثلاثة المختلفة للضجيج العشوائي(σ=0.1 ، σ=0.5، σ=1) وحساب قيم مقاييس الأخطاء لكل طريقة ظهرت لدينا النتائج التالية:

الجدول رقم (4): نتائج تطبيق الطرائق الثلاثة على مجموعة البيانات المولدة

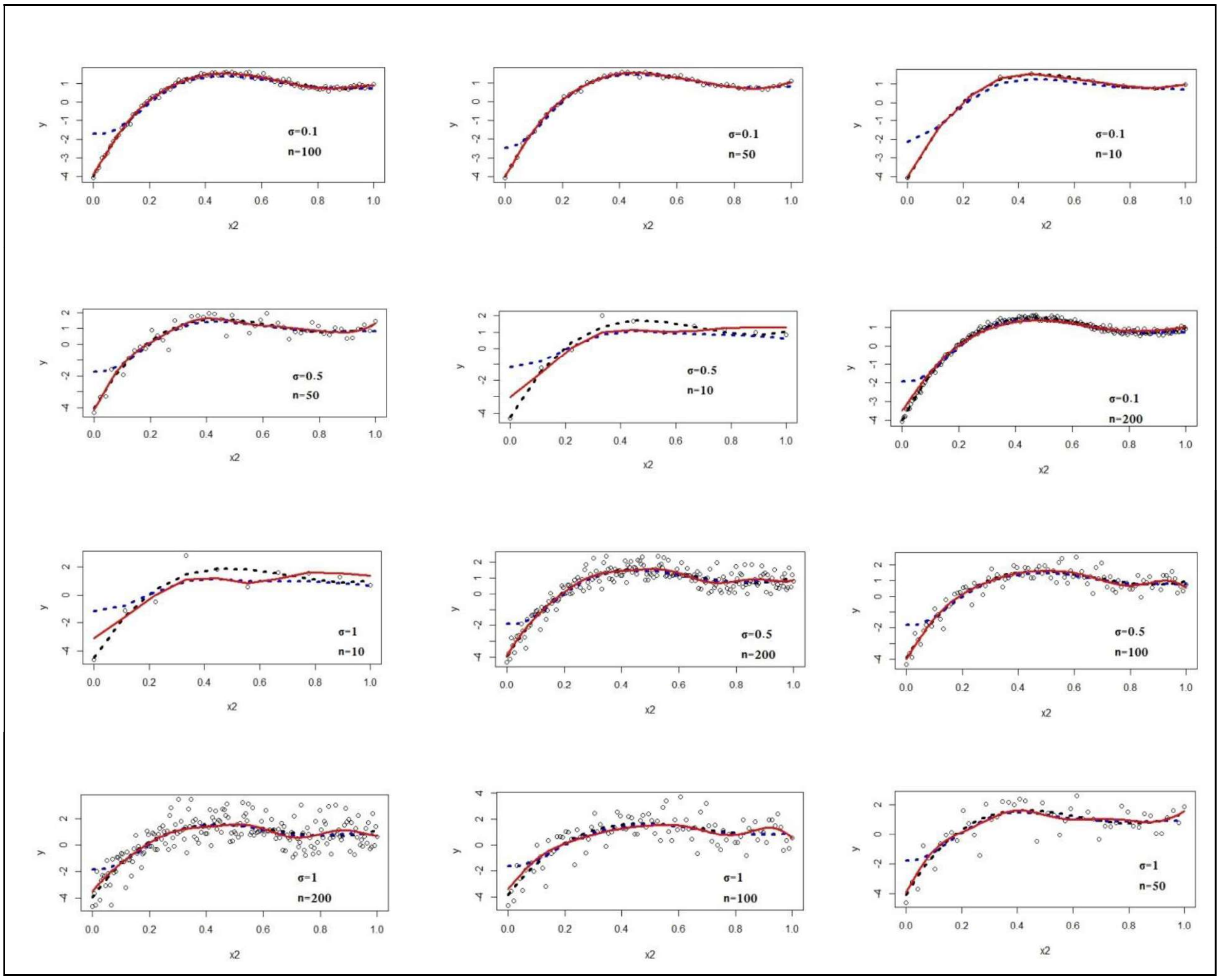
		طريقة انحدار المربعات الصغرى			طريقة انحدار العملية الغاوصية			الطريقة المقترحة		
حجم العينة n	**الانحراف المعياري σ**	**MSE**	**RMSE**	**MAPE**	**MSE**	**RMSE**	**MAPE**	**MSE**	**RMSE**	**MAPE**
10	**0.1**	0.00471979	0.068700	0.07128219	0.400048	0.632493	0.209931	**0.002842**	**0.053318**	**0.05665**
	0.5	0.1179948	0.343503	0.7287599	1.222484	1.10566	0.580972	**0.359109**	**0.599257**	**0.40222**
	1	0.4719791	0.687007	0.5790334	1.751465	1.323429	0.53271	**0.688230**	**0.829596**	**0.46033**
50	**0.1**	0.00656648	0.081033	0.07524147	0.091179	0.301959	0.131244	**0.006545**	**0.080905**	**0.07421**
	0.5	0.1641621	0.405169	0.4510113	0.403561	0.635265	0.478150	**0.153032**	**0.391194**	**0.42555**
	1	0.6566486	0.810338	3.128822	0.908080	0.952932	3.070814	**0.641103**	**0.800689**	**2.81080**
100	**0.1**	0.00796041	0.089221	0.1079193	0.203568	0.451186	0.268915	**0.007754**	**0.088062**	**0.09431**
	0.5	0.1990102	0.446105	0.7356406	0.359619	0.599682	0.736764	**0.184351**	**0.429361**	**0.68489**
	1	0.796041	0.892211	1.960669	0.953636	0.976543	1.762315	**0.736444**	**0.858163**	**1.69462**
200	**0.1**	0.00851910	0.092299	0.09052244	0.140584	0.374945	0.177091	**0.027292**	**0.165205**	**0.13139**
	0.5	0.2129776	0.461495	2.58647	0.330668	0.575038	2.294292	**0.200364**	**0.447621**	**2.71725**
	1	0.8519106	0.92299	1.716151	0.94982	0.974605	1.620708	**0.827497**	**0.909669**	**1.69805**

المصدر: من إعداد الباحث اعتماداً على مخرجات اللغة البرمجية R

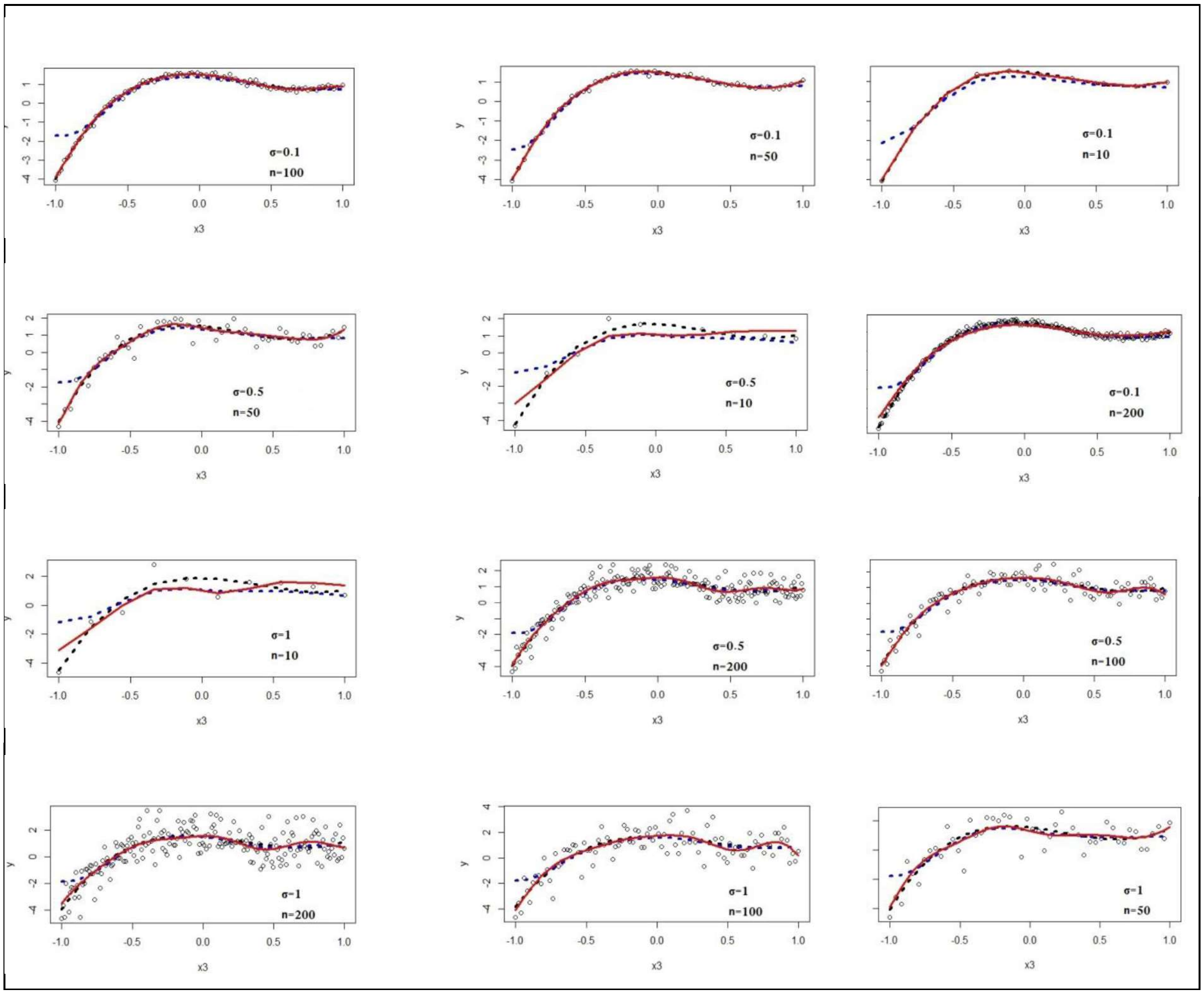
سنقوم الآن باستعراض منحنيات الانحدار الناتجة وفق الطرائق الثلاثة في حالة البيانات المولدة باستخدام المتغير التابع Yوالمتغيرات المستقلة X_1 X_2, X_3, X_4 , مع إضافة أربع حالات مختلفة لحجوم العينات (n=10, n=50, n=100, n=200) وثلاث حالات مختلفة للضجيج العشوائي(σ=0.1 ، σ =0.5، σ =1) حيث تم الرسم باستخدام اللغة البرمجية **R** باستخدام التعليمتين plot وlines.



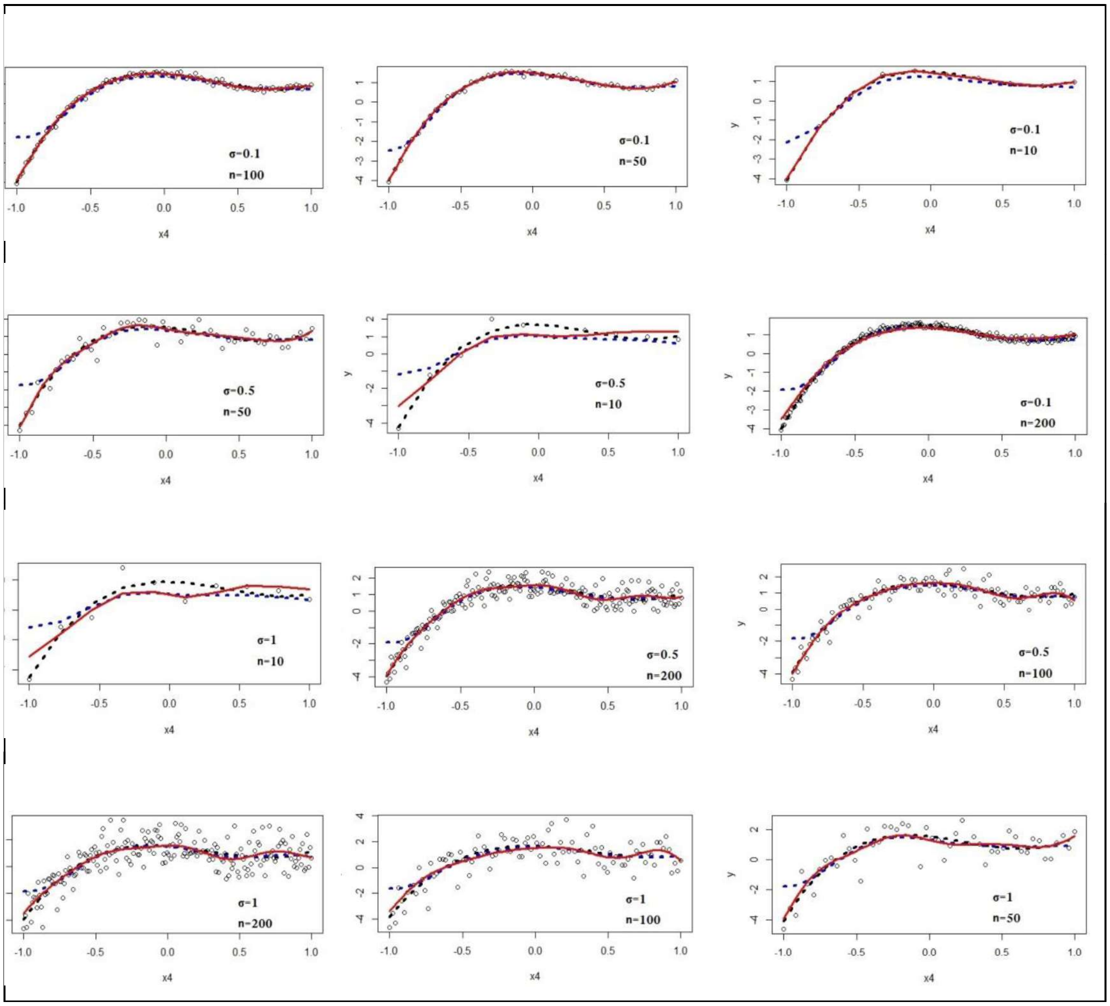
الشكل رقم (1): مقارنة منحنيات الانحدار بالطرائق الثلاثة في حالة البيانات المولدة باستخدام المتغير التابع Y والمتغير المستقل X_1 مع إضافة أربع حالات مختلفة لحجوم العينات و ثلاث حالات مختلفة للضجيج العشوائي



الشكل رقم (2): مقارنة منحنيات الانحدار بالطرائق الثلاثة في حالة البيانات المولدة باستخدام المتغير التابع Y والمتغير المستقل X_2 مع إضافة أربع حالات مختلفة لحجوم العينات و ثلاث حالات مختلفة للضجيج العشوائي



الشكل رقم (3): مقارنة خطوط منحنيات بالطرائق الثلاثة في حالة البيانات المولدة باستخدام المتغير التابع Y والمتغير المستقل X_3 مع إضافة أربع حالات مختلفة لحجوم العينات و ثلاث حالات مختلفة للضجيج العشوائي



الشكل رقم (4): مقارنة منحنيات الانحدار بالطرائق الثلاثة في حالة البيانات المولدة باستخدام المتغير التابع Y والمتغير المستقل X_4 مع إضافة أربع حالات مختلفة لحجوم العينات و ثلاث حالات مختلفة للضجيج العشوائي

5- المناقشة:

يُظهر الجدول رقم (2) من اليسار إلى اليمين اسم الطريقة المستخدمة وقيم مقاييس الأخطاء الثلاثة (RMSE،MSE ، MAPE) على مجموعة البيانات الواقعية . نلاحظ من الجدول رقم (2) بأنه كان للطريقة المقترحة قيماً أصغر لمقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE) من قيم مقاييس الأخطاء الناتجة عن التطبيق المفرد لطريقتي انحدار المربعات الصغرى وانحدار العملية الغاوصية على مجموعة البيانات الواقعية.

يُظهر الجدول رقم (4) من اليسار إلى اليمين اسم الطريقة المستخدمة وقيم مقاييس الأخطاء الثلاثة (RMSE،MSE ، MAPE) على مجموعة البيانات المولدة، إضافة لأربع حالات مختلفة لحجوم العينات (n=10, n=50, n=100, n=200) وثلاث حالات مختلفة للضجيج العشوائي(σ=0.1 ، σ=0.5، σ=1).

نلاحظ من الجدول رقم (4) تفوق الطريقة المقترحة شبه المعلمية على طريقتي انحدار المربعات الصغرى و انحدار العملية الغاوصية لأنها حققت قيماً أصغر لمقاييس الأخطاء الثلاثة (MAPE، RMSE،MSE) عند تطبيقها على مجموعة البيانات المولدة في الحالات الثلاث المختلفة للضجيج العشوائي وحجوم العينات المختلفة ونلاحظ أيضاً أنه كلما ازدادت قيمة الضجيج العشوائي كلما ازدادت قيمة مقياس الخطأ المستخدم.

توضح الأشكال (1,2,3,4) مقارنة منحنيات الانحدار المحسوبة بالطرائق الثلاثة (انحدار المربعات الصغرى بالخط المتقطع باللون الأسود وانحدار العملية الغاوصية بالخط المنقط باللون الأزرق والطريقة المقترحة شبه المعلمية بالخط المتواصل باللون الأحمر) في حالة البيانات المولدة باستخدام المتغير التابع Yوالمتغيرات المستقلة X_1 X_2, X_3, X_4 , مع إضافة أربع حالات مختلفة لحجوم العينات (n=10, n=50, n=100, n=200) وثلاث حالات مختلفة للضجيج العشوائي (σ=0.1 ، σ=0.5 ، σ=1).

نلاحظ من الأشكال (1,2,3,4) بأن منحني انحدار الطريقة المقترحة باللون الأحمر كان هو الأقرب إلى شكل الدالة الأصلية من المنحنيات المحسوبة باستخدام طريقتي انحدار المربعات الصغرى بالخط المتقطع باللون الأسود وانحدار العملية الغاوصية بالخط المنقط باللون الأزرق ونلاحظ ابتعاد منحنيات الانحدار المحسوبة باستخدام طريقتي انحدار المربعات الصغرى وانحدار العملية الغاوصية عن الدالة الأصلية في حالة بيانات لها تشويش مرتفع بشكل أكبر الأمر الذي يمكن تبريره بتأثر هاتين الطريقتين بزيادة مستوى التشويش ، وأما في حالة التشويش المنخفض للبيانات تقترب منحنيات هاتين الطريقتين من منحني الدالة الأصلية.

كما نلاحظ أيضاً بأن النموذج المقترح يمتلك أفضل تمثيل للبيانات وهذا يتوافق مع قيم الأخطاء الموضحة بالجدول رقم (4).

6- الاستنتاجات: Conclusions

من خلال ما ذكر سابقاً وما سُجل من نتائج نورد ما يلي:

1- أظهرت الدراسة التي أجريناها على مجموعتي البيانات الواقعية والمولدة تفوق الطريقة المقترحة على طريقتي انحدار المربعات الصغرى وانحدار العملية الغاوصية، وذلك من خلال تحقيقها لأصغر قيمة من قيم مقاييس الأخطاء الثلاثة (MSE ،RMSE ،MAPE).

2- أظهرت نتائج مجموعة البيانات المولدة باستخدام تجارب المحاكاة بأن أفضل طريقة للتنبؤ هي طريقة الانحدار شبه المعلمي المقترحة وذلك في الحالات الثلاثة المختلفة للضجيج العشوائي وحجوم العينات المختلفة.

3- من خلال نتائج هذه الدراسة تبيّن بأن النموذج المُقترح قد أعطى دقة تنبُؤ أفضل من دقة التنبُؤ التي حصلنا عليها باستعمال نماذج التنبُؤ المُفردة لطريقتي انحدار المربعات الصغرى وانحدار العملية الغاوصية وذلك لتكرار عدد الأفضلية بالاعتماد على أصغر قيمة من قيم مقاييس الأخطاء المستخدمة وبسبب قدرة منحني الانحدار الممثل لها على ملاءمة وتمثيل البيانات بشكل أفضل.

4- إن المرونة التي يوفرها النموذج شبه المعلمي في توصيف البيانات بصورة عامة تكون كبيرة جداً مقارنة بالنموذج المعلمي الخطي والنموذج اللامعلمي اللاخطي ويزداد الأمر وضوحاً عندما تكون المشكلة أو الظاهرة المدروسة تحوي على متغيرات كثيرة.

7- التوصيات: Recommendations

1- مما سبق نوصي باستخدام الطريقة المقترحة على مجموعات بيانات أُخرى كونها أعطت أفضل النتائج وكفاءة ومرونة عالية في التطبيق.

2- نوصي الباحثين بإجراء عمليات دمج بين طرائق الانحدار الأخرى وتطويرها وتطبيقها في مجالات العلوم المختلفة.

3- إجراء دراسات مستقبلية حول تحسين دقة التنبؤ باستخدام طرائق أخرى غير الطريقة المقترحة في هذا البحث مثل استخدام طريقة بايز في حالة تقدير الجزء المعلمي وطريقة انحدار متجه الدعم في حالة تقدير الجزء اللامعلمي.

8- المراجع: References:

1- **Akkus,** O.,**(2011)**-Xplore Package For The Popular Parametric And Semi parametric Single Index Models .Journel of science, vol.24, No.4, pp. 753-762.

2- **Aydin,D.,(2011)**-'' Partially Linear Models Based on Smoothing Spline Estimated by Different Selection Methods: A Simulation Study" Department of Statistics, Faculty of Arts and Sciences, Muğla University.

3- **Bishop, C. M., (2007).**Pattern Recognition and Machine Learning. Springer.

4- **Cherkassky, V., Ma, Y., (2004).**Practical Selection of SVM Parameters and Noise Estimation for SVM Regression, Neural Networks, 17, 113-126.

5- **Cotton, R., (2013).** Learning R, O'Reilly Media, Inc., United States of America, 377.

6- **Hardle, W., Muller, M., Sperlich, S.,and Werwatz A., (2004)** .Nonparametric and Semiparametric Models, Springer, Berlin, 301.

7- **Hastie, T., Tibshirani, R., Friedman, J., (2009).** The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer, 2th ED, Berlin, 764.

8- **Hyndman, R., J., Koehler, A. B., (2006).** Another Look at Measures of Forecast Accuracy, International Journal of Forecasting, 22, 679–688.

9- **Izenman, A.J., (2008)**- Multivariate Statistical Techniques: Regression,Classification and manifold learning. New York: Springer.

10- **Law ,Kelton, Mcgraw, Hill., (2000)**- Simulation Modeling and Analysis .3rd edition.

11- **Liu, Y., Keller, Y., Song, PH., Bond, J., Jiang, G., (2017).**Prediction of concrete corrosion in sewers with hybrid Gaussian processes regression model. RSC Advances, 7, 30894–30903.

12- **Matloff, N., (2011).**The Art of R Programming, Malloy Incorporated, United States of America, 373.

13- **Millimet, D., List, J., and Stengos, T., (2003)**- The environmental Kuznets curve. Real progress or misspecified models, Rev Econ Stat (85)4 .pp1038-1047,

14- **Nielsen, A.,**(2009) .Least Squares Adjustment Linear and Nonlinear Weighted Regression Analysis Informatics and Mathematical Modelling.Technical University, Denmark.

15- **Pérez G. A.;Vieu PH.,(2008)** -Nonparametric time series prediction: A semi-functional partial linear modeling. Journal of Multivariate Analysis, 99, 834 – 857.

16- **Rasmussen, C. E., Williams C. K. I., (2006).**Gaussian Processes for Machine Learning. MIT, Press.

17- **Ruppert, D., Wand, M.P., Carroll, R.J., (2003)** -Semiparametric Regression.Cambridgeuniversity Press,New York.

18- **Schimek,M.G.,(2000)**-" Estimation and inference in partially linear models with smoothing splines" Journal of Statistical Planning and Inference(91) ,PP 525_540.

19- **SPECKMAN, P., (1988)**- Kernel Smoothing in partially Linear Models. Journal of Royal Statistical Soc. 50, No.3,pp. 413-436.

20- **Wang,Q.,Linton,O.,and Härdle,W.,(2004)**-" Semiparametric Regression Analysis with Missing Response at Random " the institute for fiscal studies department of economics, UCL ,cemmap working paper CWP11/03.