أسلوب مقترح في تقدير القيم المفقودة في نموذج الانحدار المتعدد اللامعلمي

أ.م.د. قتيبة نبيل نايف القزاز/ كلية الادارة والاقتصاد/ جامعة بغداد

المستخلص

في هذا البحث سوف نقدم أسلوب مقترح في تقدير القيم المفقودة لمشاهدات المتغيرات التوضحية لنموذج الانحدار المتعدد اللامعلمي ومقارنتها مع طريقة التعويض بالوسط الحسابي، أن أساس فكرة هذا الأسلوب أستندت الى كيفية توظيف العلاقة السببية بين المتغيرات في ايجاد تقدير كفوء للقيمة المفقودة، معتمدين في ذالك على استعمال تقدير Kernel والمتمثل بمقدر Nadary - Watson وعلى طريقة المربعات الصغرى للعبور الشرعي LSCV في تقدير المعلمة التمهيدية، ومستخدمين اسلوب المحاكاة في المقارنة بين الطريقتين.

المصطلحات الرئيسية للبحث/ الانحدار المتعدد اللامعلمي- المشاهدات المفقودة- آلية الفقدان- نمط الفقدان- مقدر Nadary – Watson - المربعات الصغرى للعبور الشرعي

مجلة العلوم
الاقتصادية والإدارية
المجلد ٢٢ العدد ٨٩
الصفحات ٣٩٦-٤٠٦

المقدمة

من المشاكل التي تواجه البيانات ولاسيما بيانـــات نماذج الانحـــدار المتعدد اللامعلمي، وجود فقدان في مشاهدات المتغيرات التوضيحية ($X^{'s}$) بشكل انماط مختلفة للفقدان وبسبب آليات مختلف للفقدان أيضاً. لقد شهدت مشكلة الفقدان اهتماماً واسعاً في بداية السبعينات لما شهدته هذه الفترة من تطور ملحوظ في اجهزة الحاسوب وبرامجها، ولقد اسهم العديد من الباحثين من امثال **Rubin , Demprster & Laird** الذين يعدون من الرواد الأوئل، في تطوير وايجاد طرائق كفوءة في تقدير القيم المفقودة في شتى بيانات البحث العلمي خلال الاربعين سنة المنصرمة. كما ان هناك العديد من طرائق تقدير القيم المفقودة وكل طريق لها شروطها في الأستعمال وبحسب نمط الفقدان وآليتهِ، ومن الطرائق المتداولة في تقدير القيم المفقودة لمشاهدات المتغيرات التوضيحية $X^{'s}$ لنموذج الانحدار المتعدد اللامعلمي وعند آلية فقدان من نوع " فقدان البيانات تماماً بشكل عشوائي" **Missing Complete At Random (MCAR)**، هي طريقة التعويض بالوسط الحسابي اذ تُعد هذه الطريقة من طرائق التعويض الاحادية **Single Imputation**، ولم يُعاب على هذه الطريقة، ضعف كفائتها عند نسب فقدان عالية من جهة، ومن جهة اخرى ضعف العلاقة السببية بين المتغيرات التي تؤدي الى فَشل هذه الطريقة في التقدير أيضاً.

وفي ضوء هذه المشكلتين التي تواجه طريقة التعويض بالوسط الحسابي، انبثقة فكرة هذا البحث بتوظيف اسلوب جديد ضمن هذه الطريقة لغرض زيادة كفائتِها، والمتمثل بأستعمال طريقة تقدير **Kernel** في تقدير القيم المفقودة في المتغيرات التوضيحية $X^{'s}$.

سيتم في هذا البحث عرض طريقة المتوسط الشرطي في تقدير القيم المفقودة لمشاهدات المتغيرات التوضيحية $X^{'s}$ ومن ثم عرض مفصل لكيفية توظيف تقدير **Kernel** والمتمثل بمقدر **Nadary - Watson** في تقدير القيمة المفقودة لمشاهدات المتغيرات المستقلة، ومن ثم استعراض طريقة المربعات الصغرى للعبور الشرعي **Least Squared Cross Validation (LSCV)** في تقدير المعلمة التمهيدية في مقدر **Nadary – Watson**، ومن ثم تتم المقارنة بين الطريقتين بأستعمال المحاكاة.

الجانب النظري : في بعض الظواهر التي يتم دراستها والتي يمكن تمثيلها بمنحنى الانحدار لوصف العلاقة بين المتغيرات التوضيحية $X^{'s}$ ومتغير الأستجابة **Y** والتي تتم الحصول عليها من عينة حجمها **n**، فأن هذه العلاقة (علاقة تأثير) بين هذه المتغيرات والتي يمكن وصفها بالنموذج الآتي:

$$Y_i = g(X_{1i}, X_{2i}, X_{3i}, \dots, X_{ki}) + \varepsilon_i \quad ; \quad i = 1, 2, 3, \dots, n \qquad \dots(1)$$

وان الصيغة (1) تمثل نموذج الانحدار المتعدد اللامعلمي **Nonparametric Multiple Regression** وان **K** تمثل عدد المتغيرات المستقلة وأن $g(X_{1i}, X_{2i}, X_{3i}, \dots, X_{ki})$ تمثل دالة الأنحدار المجهولة والتي يتم افتراضها كدالة مستمرة، وأن ε_i يمثل الخطأ العشوائي بمتوسط مساواً للصفر وتباين مقدارة σ^2.

أن مشكلة المشاهدات المفقودة في المتغيرات التوضيحية من المشاكل التي تواجه الباحثين في شتى المجالات، لذلك فان هذه المشكلة لها تاثير كبير في نتائج معادلة الانحدار التقديرية وعليه يجب تقدير هذه القيم قبل البداء في تقدير معادلة الانحدار، ولغرض دراسة اساليب تقدير المشاهدات المفقودة في اي متغير مستقل (أو متغيرات مستقلة) سنقوم اولا باستعراض آليات وانماط الفقدان ومن ثم نستعرض اسلوب الوسط الحسابي في تقدير القيم المفقودة وبعد ذلك سيتم توضيح آلية توظيف مقدر **Kernel** في التقدير مع توضيح طريقة المربعات الصغرى للعبور الشرعي **LSCV** في تقدير معلمة التمهيد.

القيم المفقودة : أن فقدان المشاهدات في أية ظاهرة معينة من الأمور المسلم بها ويعود ذلك لأسباب عديدة ومختلفة عمدية كانت أم غير عمدية ولكن بصورة عامة هناك آليات وأنماط لهذه المشكلة، ولغرض التطرق لهذه الآليات والأنماط نفترض لدينا المتغيرات المستقلة $X^{'s}$ مع المتغير المعتمد **Y** وكما يأتي:(القزاز، صفحة ٨ - ١١)[1]

$$Y \quad X_1, X_2, X_3, \dots, X_k$$

أن آلية الفقدن تتمحور على ثلاثة أنواع هي:

١.الآلية الأولى: فقدان البيانات تماماً بشكل عشوائي **Missing Complete at Random** ويرمز لها **MCAR**، ويعني ان سبب فقدان أية مشاهدة لمتغير مستقل معين يكون مستقل عن القيمة المفقودة نفسها وكذلك عن قيم اي متغير مستقل أخر (أو متغيرات مستقلة أخرى).

٢.الآلية الثانية: فقدان البيانات بشكل عشوائي **Missing At Random** ويرمز لها **MAR**، هنا سبب فقدان اي مشاهدة لمتغير مستقل معين يعتمد على مشاهدات المتغير المستقل الأخر (أو المتغيرات المستقلة الأخرى) ولكن مستقل عن المشاهدة المفقودة نفسها.

٣.الآلية الثالثة: فقدان البيانات بشكل غير عشوائي **Missing At Not Random** ويرمز لها **Not MAR**، هذا يعني ان سبب فقدان اية مشاهدة لمتغير مستقل معين يعتمد على مشاهدات المتغير المستقل الأخر (أو المتغيرات المستقلة الأخرى) وكذلك يعتمد على القيمة المفقودة نفسها.

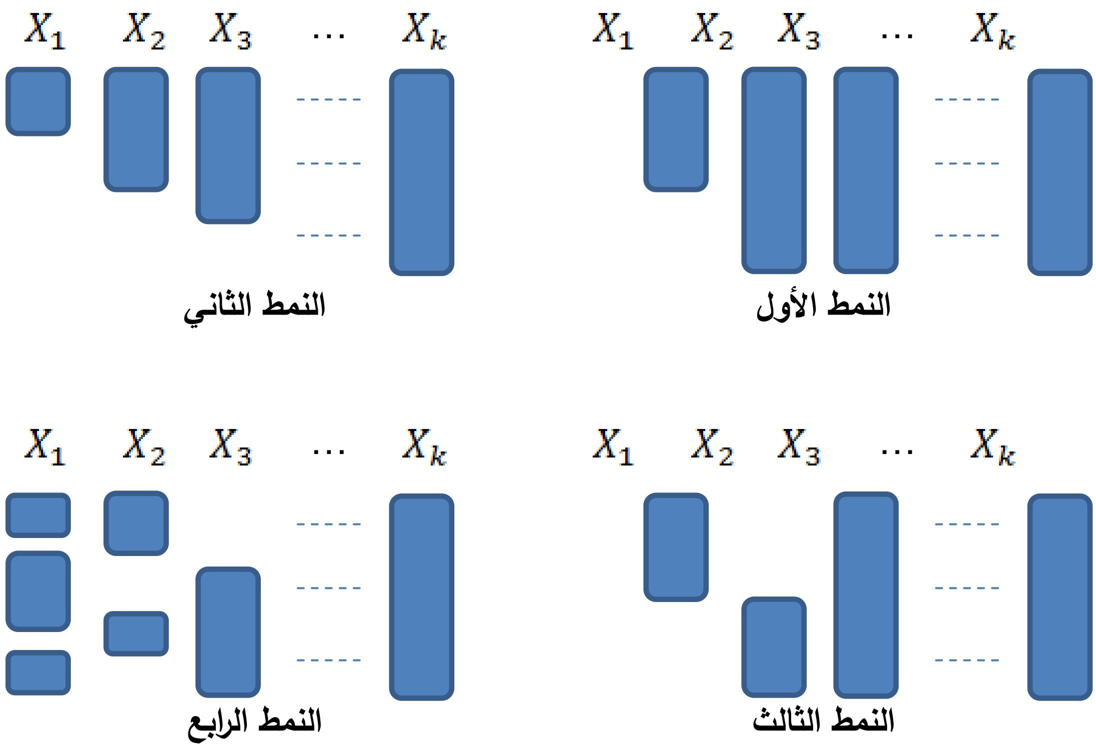
أما الأنماط الخاصة بالفقدان فهي أربعة انواع وكما يأتي:

النمط الأول: فقدان مشاهدات متغير مستقل واحد، في هذه الحالة يكون الفقدان في مشاهدات أحد المتغيرات المستقلة وباقي المتغيرات تكون تامة المشاهدات.

النمط الثاني: فقدان المشاهدات بشكل مرتب لبعض المتغيرات المستقلة، في هذا النمط يكون ترتيب المشاهدات المفقودة لبعض المتغيرات المستقلة بشكل تصاعدي أو تنازلي.

النمط الثالث: فقدان المشاهدات بشكل عدم تطابقها لمتغيرين مستقلين، في هذا النمط يكون الفقدان في متغيرين فقط أذ يكون هناك فقدان في مشاهدات المتغير الأول يقابلها مشاهدات تامة في المتغير الثاني، كذلك فان المشاهدات المفقودة في المتغير الثاني يقابلها مشاهدات تامة في المتغير الأول.

النمط الرابع: النمط العام، في هذا النمط لا يمكن وضع شكل معين للمشاهدات المفقودة بل تكون بشكل عشوائي ومبعثرة.



شكل (١) يوضح أنماط الفقدان في المتغيرات المستقلة

طرائق تقدير القيم المفقودة: سوف نستعرض طريقتين في تقدير القيم المفقودة لمشاهدات المتغيرات التوضيحية $X^{'s}$ والمتمثلة بطريقة التعويض بالوسط الحسابي وطريقة التعويض بتقدير **Kernel** كأسلوب مقترح وتحت آلية فقدان من نوع **MCAR** .

أولا: التعويض بالوسط الحسابي: تعد هذه الطريقة من اقدم طرائق التعويض الأحادي **Single Imputed** وابسطها اذ يتم حساب متوسط مشاهدات المتغير الذي يعاني من فقدان في مشاهداته، ويتم تعويض هذا المتوسط بدلاً عن القيمة المفقودة، لو فرضنا أن المتغير المستقل X_j يعاني فقدان في بعض مشاهدته فانه يمكن حساب متوسط مشاهدات هذا المتغير بحسب الصيغة الآتية:

$$\bar{X}_j = \frac{\sum_{r=1}^{m} X_{jr}}{m_j} \;\; ; \;\; r = 1,2,3,\dots,m_j \;\; ; \;\; j = 1,2,3,\dots,k \qquad \dots(2)$$

حيث أن m_j يمثل عدد المشاهدات في المتغير التوضيحي X_j بعد استبعاد عدد المشاهدات المفقودة، اي ان m_j هي مجموعة جزئية من المشاهدات الكلية **n** اي بصورة اخرى $m_j \subset n$ ان طريقة التعويض بالمتوسط تعد من اكفاء الطرائق في التقدير ولاسيما في حالة تجانس بيانات العينة وضعف العلاقة السببية بين المتغيرات التوضيحية في معادلة الانحدار، وكذلك تبرز كفائتها في حالة نسب الفقدان الصغيرة، ولكن تفشل هذه الطريقة في حالة نسب الفقدان الكبيرة وكذلك عند وجود علاقة سببية بين المتغيرات التوضيحية.

ثانياً: التعويض باستعمال تقدير **Kernel**: في عام ١٩٩٢ قدم الباحث **Little** [5] بحثاً وضح فيه اسلوب تقدير المشاهدات المفقودة لأحد المتغيرات التوضيحية في أنموذج الانحدار المتعدد المعلمي باستعمال اسلوب التعويض بالمتوسط الشرطي الذي يتم الحصول عليه من خلال بناء نموذج انحدار بين المتغير التوضيحي الذي تعاني مشاهداتهُ من فقدان مع بقية المتغيرات التوضيحية وتوصل الى كفاءة هذا المقدر تقل عند نسب الفقدان العالية وكذلك عند ضعف العلاقة السببية بين المتغيرات التوضيحية. كذالك في عام ٢٠٠٨ قدم الباحثان قتيبة ومناف [2] بحثاً حول تقدير القيم المفقودة في مشاهدات متغير الاستجابة **Y** لنموذج الانحدار البسيط اللامعلمي بالاعتماد على تعويض مقدر قاعدة **Kernel** الاحادي بدلاً عن القيمة المفقودة، اذا اظهر هذا الاسلوب كفاءة عالية في التقدير.

ومما سبق فأن فكرة الاسلوب المقترح في هذا البحث هو الحصول على مقدرات للقيم المفقودة تكون ذات كفاءة عالية في حالة نسب الفقدان الكبيرة وفي حالة التشتت الكبير في مشاهدات المتغيرات وكما يأتي:

لو فرضنا انه لدينا متغيرين توضيحيين هما $X_1 \,, X_2$ يؤثران في المتغير المعتمد **Y** فان نموذج الانحدار المتعدد اللامعلمي هو :

$$Y_i = g(X_{1i}\,, X_{2i}) + \varepsilon_i \quad ; \quad i = 1\,,2\,,3\,,\dots,n \qquad \dots(3)$$

ولو فرضنا ان المتغير X_2 يعاني من فقدان في بعض مشاهداته، ولغرض تقدير هذه المشاهدات يتم بناء نموذج انحدار لامعلمي للمتغير X_2 مع المتغير X_1 وبحسب الصيغة الآتية:

$$X_{2i} = g(X_{1i}\,) + \varepsilon_i \quad ; \quad i = 1\,,2\,,3\,,\dots,n \qquad \dots(4)$$

حيث أن ε_i يمثل الخطأ العشوائي المجهول لأستجابة X_{2i} كذلك ان الصيغة (4) سوف تحتوي على حالة أستجابة وعدم استجابة للمتغير X_2 والتي يرمز لها بالرمز γ بحيث ان :

$$\gamma_i = \begin{cases} 1 & if\ X_2 \text{ is rsponse} \\ 0 & if\ X_2 \text{ is not response} \end{cases}$$

وعليه فان المعادلة التقديرية للصيغة (4) يمكن الحصول عليه كما يأتي :

$$\hat{X}_{2i} = \hat{g}(X_{1i}\,) + \hat{\varepsilon}_i \quad ; \quad i = 1\,,2\,,3\,,\dots,n \qquad \dots(5)$$

بحيث أن $\hat{\varepsilon}_i$ هي الفرق بين $g(X_{1i})$ و $\hat{g}(X_{1i})$ أما $\hat{g}(X_{1i})$ تتم الحصول عليها بحسب قاعدة Kernel (مقدر Nadary – Watson) وكما يأتي:

$$\hat{g}(X_{1i}) = \frac{\sum_{i=1}^{n} \gamma_i X_{2i}\ k\left(\frac{X_1 - X_{1i}}{h}\right)}{\sum_{i=1}^{n} \gamma_i\ k\left(\frac{X_1 - X_{1i}}{h}\right) + n^{-2}} \qquad \dots (6)$$

وعلية فان الصيغة (5) تحتوي على مشاهدات تامة للمتغير X_{2i}، بعد ذلك يتم استبدال قيمة $\hat{X}_{2i}$ بدل القيمة المفقودة في المتغير X_{2i} فيصبح لدنيا متغير جديد بمشاهدات تامة نرمز له $\tilde{X}_{2i}$ ومن ثم يتم اعادة كتابة الصيغة (5) كما يلي:

$$\tilde{X}_{2i} = g(X_{1i}) + \hat{\varepsilon}_i \quad ; \quad i = 1, 2, 3, \dots, n \qquad \dots (7)$$

ان السبب في اعادة بناء الصيغة (7) هو لتمهيد البيانات التي تم تقديرها بدل القيم المفقودة في المتغير X_{2i}، لغرض الحصول على مقدارات اكثر كفاءة بالاعتماد على العلاقة السببية بين المتغيرات، وباستعمال مقدر Nadary – Watson مرة ثانية وللمشاهدات التامة للمتغيرين $\tilde{X}_{2i}$ و X_{1i} وبحسب الصيغة الآتية:

$$\tilde{g}(X_{1i}) = \frac{\sum_{i=1}^{n} \tilde{X}_{2i}\ k\left(\frac{X_1 - X_{1i}}{h}\right)}{\sum_{i=1}^{n}\ k\left(\frac{X_1 - X_{1i}}{h}\right) + n^{-2}} \qquad \dots (8)$$

نحصل على تقدير جديد للمتغير $\tilde{X}_{2i}$ هو :

$$\hat{\tilde{X}}_{2i} = \hat{g}(X_{1i}) \quad ; \quad i = 1, 2, 3, \dots, n \qquad \dots (9)$$

ومن ثم نقوم باستبدال $\hat{\tilde{X}}_{2i}$ بدلاً عن القيمة المفقودة في X_{2i}، تجدر الاشارة هنا أن $k(\,.\,)$ في الصيغة (6) و (8) تمثل دالة لبية والتي يمكن الحصول عليها من دالة الكثافة الطبيعية القياسية او ما يطلق عليها بـ Gaussian Kernel وبحسب الصيغة الآتية:

$$k(z) = \sqrt{2\pi}\ \exp\left(-\frac{z^2}{2}\right) \qquad \dots (10)$$

أما h في الصيغة (6) و (8) فتمثل المعلمة التمهيدية أو معلمة عرض الحزمة (Bandwidth)، وهناك طرائق مختلفة في تقديره فقد استعمل قتيبة ومناف [2] صيغة المصدر الطبيعي وكما يلي :

$$h = 1.06\ \sigma\ n^{-1/5} \qquad \dots (11)$$

في بحثنا هذا سوف نستعمل طريقة تقدير اكثر كفاءة من صيغة المصدر الطبيعي وهي طريقة المربعات الصغرى للعبور الشرعي Least Squared Cross Validation (LSCV).

طريقة المربعات الصغرى للعبور الشرعي LSCV : ان أساس فكرة طريقة المربعات الصغرى تعتمد على تصغير تكامل مربعات الخطاء بين دالة القيم الحقيقة والتقديرية اي أن: [3] [Horne & Garton : p.641-642]

$$\int \left(\hat{f}(X_i) - f(X_i)dx\right)^2 dx = \int \hat{f}^2(X_i)dx - 2\int \hat{f}(X_i)f(X_i)dx + \int f^2(X_i)dx$$

وفي عام ١٩٨٢ توصل الباحث Silverman الى صيغة تقريبية للصيغة السابقة وهي:

$$LSCV(h) = \int \hat{f}^2(X_i)dx - 2n^{-1}\sum \hat{f}_{-i}(X_i) \qquad \dots .(12)$$

اذ تمثل الصيغة (12) مجموعة قيم للمعلمة h بطريقة المربعات الصغرى للعبور الشرعي.

وأن $\hat{f}_{-i}$ هي تقدير دالة الكثافة عند حذف المشاهدة i . في عام ١٩٩٢ توصل Worton الى قيم للمعلمة h في الصيغة (12) وكما يلي:[4]

$$LSCV(h) = \frac{1}{\pi h^2 n} + \frac{1}{4\pi h^2 n^2} \times \sum_{i=1}^{n}\sum_{j=1}^{n}\left(exp\left[\frac{-d_{ij}^2}{4h^2}\right] - 4exp\left[\frac{-d_{ij}^2}{2h^2}\right]\right) \quad ...(13)$$

حيث ان d_{ij}^2 تمثل المسافة بين الموقعين i^{th} و j^{th} للمشاهدات أما قيمة h فتحسب من الصيغة (11)، وللحصول على قيمة h المثلى من الصيغة (13) يكون بحسب الصيغة الاتية:

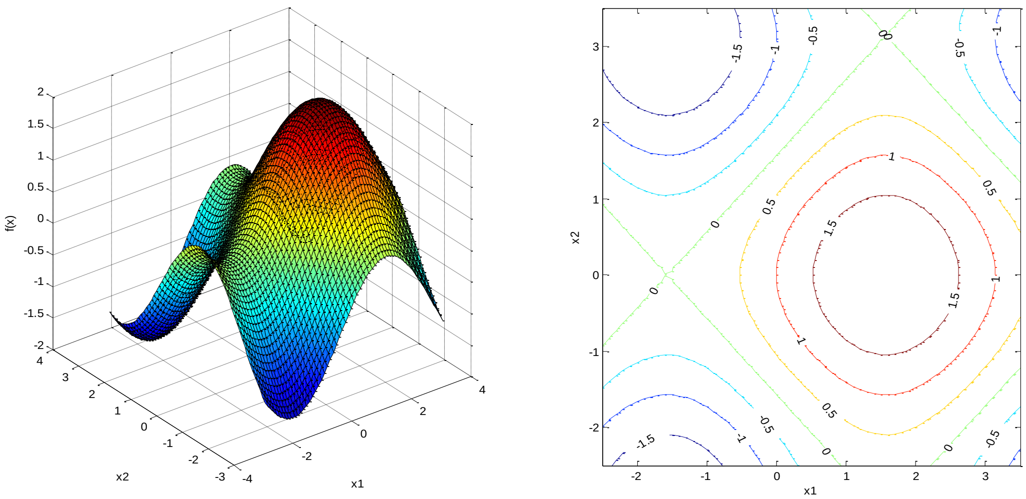
$$\hat{h}_{LSCV} = argmin_h \, LSCV(h) \quad ...(14)$$

الجانب التجريبي : لغرض معرفة كفاءة الأسلوب المقترح والمتمثل بتعويض مقدر Nadary - Watson عن القيم المفقودة بالنسبة لطريقة التعويض بالوسط الحسابي وبيان تأثر هذه الطرائق بنسب الفقدان وأختلاف حجوم العينة و تغير قيم التباينات والأرتباط تم اللجوء الى اسلوب المحاكاة وكما يأتي:

أولاً: تم استعمال نموذج الانحدار المتعدد اللامعلمي الاتي:

$$Y_i = sin(X_{1i}) + cos(X_{2i}) + \varepsilon_i \quad ...(15)$$

والشكل رقم (2) يمثل الرسم البياني لهذا النموذج.



شكل (٢) يوضح الشكل البياني لنموذج الانحدار المتعدد اللامعلمي في الصيغة (15)

ثانياً: توليد توزيع طبيعي ثنائي المتغيرات باستخدام دالة التوليد (mvnrnd) المتوفرة في البرنامج الجاهز Mat lab، وتم استخدام تباينات مختلفة للمتغيرات المستخدمة وهي على التوالي :

$$X\sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right) \; ; \; X\sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}2 & \rho\sqrt{6}\\ \rho\sqrt{6} & 3\end{pmatrix}\right) \; ; \; X\sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}3 & \rho\sqrt{15}\\ \rho\sqrt{15} & 5\end{pmatrix}\right)$$

وقد تم استعمال قيم مختلفة للارتباطات هي $\rho = 0.20\,,0.50,0.90$ كذلك تم استخدام حجوم مختلفة للعينات $n = 15\,,40\,,100$

<u>ثالثاً:</u> تم توليد مصفوفة الفقدان تحت شرط الية فقدان من نوع MCAR [1] وللنمط الاول وبنسب فقدان $(10\%\ ,\ 20\%\ ,\ 30\%\ ,\ 40\%)$ ، وتم الفقدان بالنسبة للمتغير الثاني.

<u>رابعاً:</u> ولغرض المقارنة تم استعمال معيار متوسط مربعات الخطأ MSE.

والجدول الاتي يوضح نتائج المحاكاة على وفق المعطيات المذكورة انفاً، كذلك تم عرض بعض الاشكال لعدد من الحالات المستخدمة.

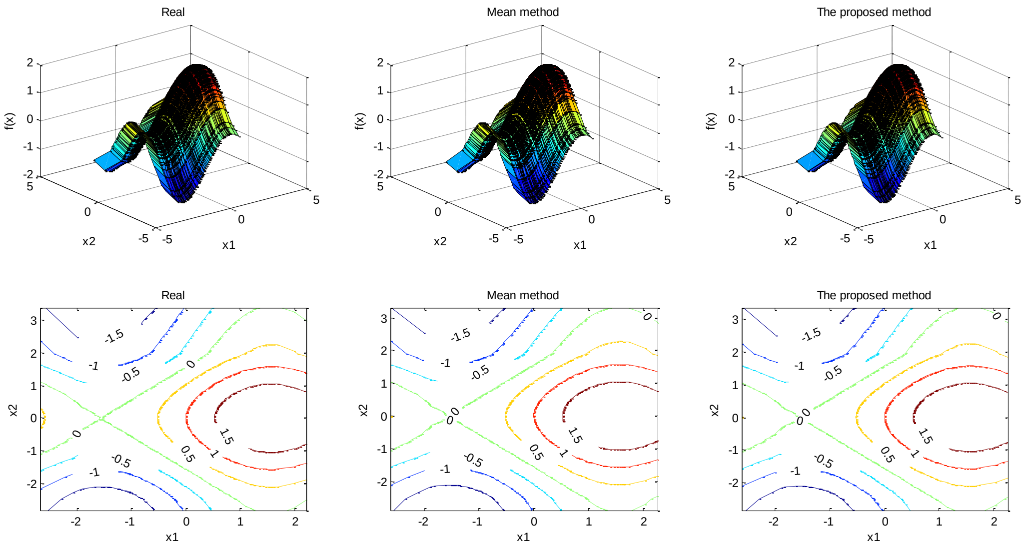
جدول (١)

يبين قيمة متوسط مربعات الخطأ لتقدير دالة الانحدار المتعدد اللامعلمي في الصيغة (15) وحسب النمط الأول للقيم المفقودة وقيم التباينات وحجوم العينات المستخدمة ونسب الفقدان (العدد الناتج مضروب في ١٠٠٠٠)

Missing		10%									20%								
n	ρ	0.2			0.5			0.9			0.2			0.5			0.9		
	σ_1^2	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
	σ_2^2	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5
15	Mean Method	453	1224	1521	442	1248	1502	421	1288	1493	645	1800	2190	644	1795	2129	689	1797	2188
	Proposed method	515	1136	1202	432	941	1126	105	360	564	770	1672	1922	595	1397	1727	168	510	879
40	Mean Method	365	986	1231	347	975	1228	350	995	1228	647	1973	2456	686	1982	2454	680	1946	2387
	Proposed method	402	846	974	298	685	875	84	258	394	763	1626	1897	600	1430	1767	163	522	816
100	Mean Method	359	1020	1296	357	1011	1296	343	1061	1299	686	2097	2590	689	2025	2594	686	2058	2579
	Proposed method	399	828	958	308	706	877	79	267	414	767	1641	1913	607	1425	1753	160	514	803
Missing		30%									40%								
n	ρ	0.2			0.5			0.9			0.2			0.5			0.9		
	σ_1^2	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
	σ_2^2	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5
15	Mean Method	1100	2954	3560	1066	2970	3441	1085	2959	3441	1279	3456	4240	1301	3407	4132	1276	3482	4228
	Proposed method	1293	2705	3269	989	2389	2900	264	2361	2765	1477	3325	3741	1214	2788	3426	330	1052	1626
40	Mean Method	1049	2950	3607	1016	2911	3596	1022	2895	3564	1357	3893	4734	1323	3879	4752	1371	3838	4739
	Proposed method	1159	2440	2840	917	2076	2625	240	756	1211	1531	3285	3812	1189	2820	3480	318	1036	1619
100	Mean Method	1045	3077	3848	1069	3088	3810	1052	3091	3836	1398	4050	5038	1380	4087	5073	1386	4033	5042
	Proposed method	1159	2485	2871	935	2120	2579	243	776	1221	1527	3295	3770	1239	2849	3521	328	1031	1601

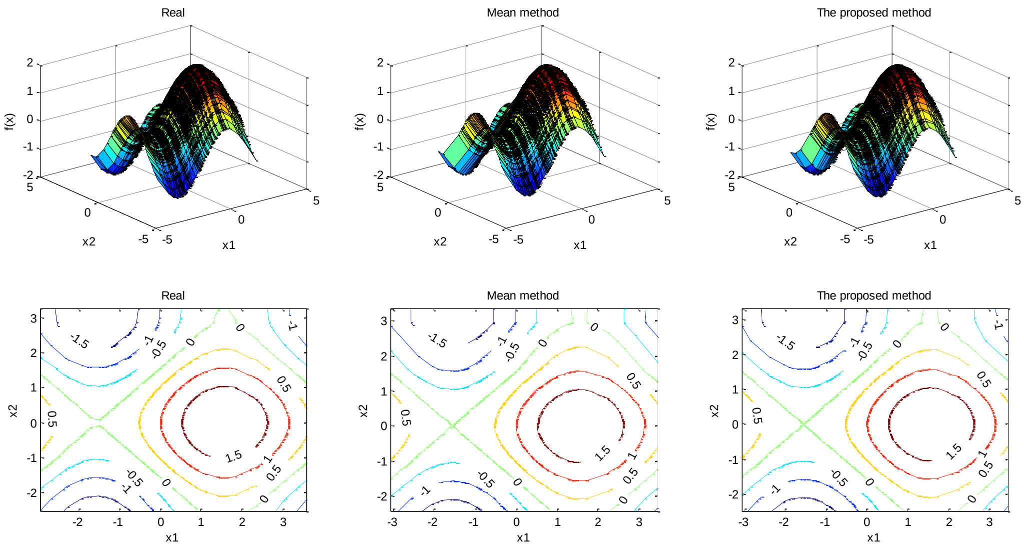
شكل رقم (٣)

يشير الى القيم الحقيقية والقيم التقديرية للقيم المفقودة لنموذج الانحدار المتعدد اللامعلمي وبحسب النمط الأول عند حجم عينة $n = 100$ وتباينات $\sigma_1^2 = 1\ ,\ \sigma_2^2 = 1$ ودرجة أرتباط $\rho = 0.50$ ونسبة فقدان **10%**



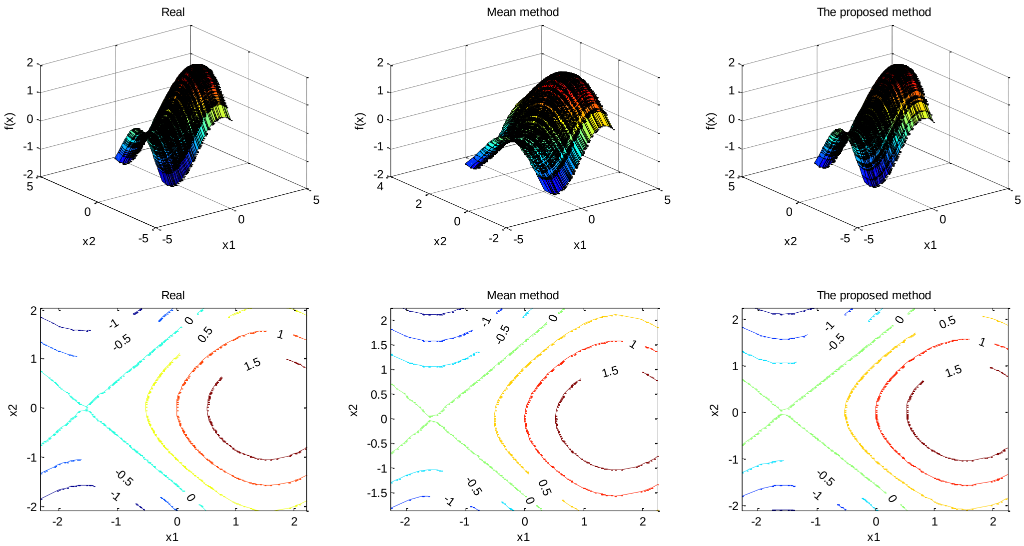
شكل رقم (٤)

يشير الى القيم الحقيقية والقيم التقديرية للقيم المفقودة لنموذج الانحدار المتعدد اللامعلمي وبحسب النمط الأول عند حجم عينة $n = 100$ وتباينات $\sigma_2^2 = 1$, $\sigma_1^2 = 1$ ودرجة أرتباط $\rho = 0.50$ ونسبة فقدان 20%



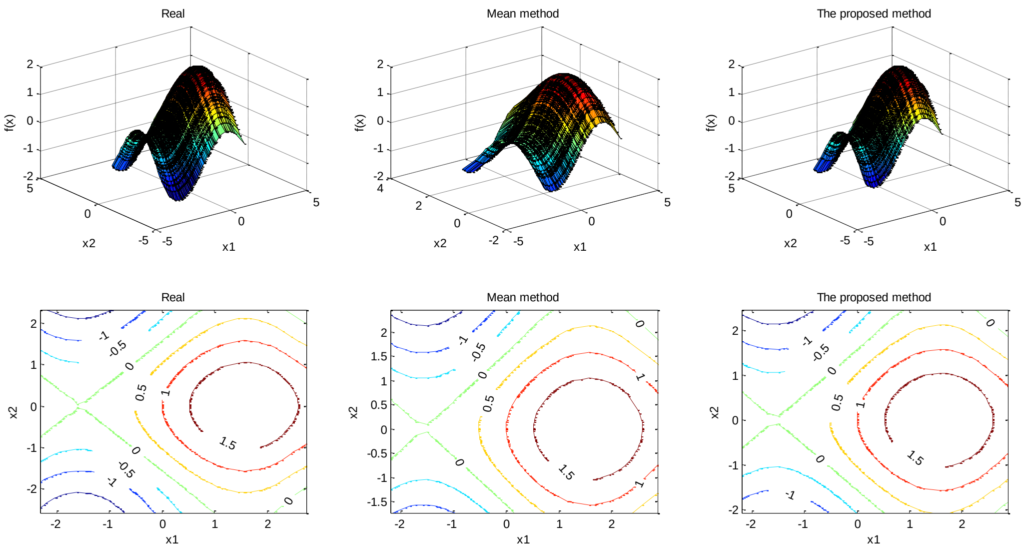
شكل رقم (5)

يشير الى القيم الحقيقية والقيم التقديرية للقيم المفقودة لنموذج الانحدار المتعدد اللامعلمي وبحسب النمط الأول عند حجم عينة $n = 100$ وتباينات $\sigma_2^2 = 1$, $\sigma_1^2 = 1$ ودرجة أرتباط $\rho = 0.50$ ونسبة فقدان 30%



شكل رقم (6)

يشير الى القيم الحقيقية والقيم التقديرية للقيم المفقودة لنموذج الانحدار المتعدد اللامعلمي وبحسب النمط الأول عند حجم عينة $n = 100$ وتباينات $\sigma_1^2 = 1$, $\sigma_2^2 = 1$ ودرجة أرتباط $\rho = 0.50$ ونسبة فقدان 40%



تفسير النتائج:

من الجدول (١) نلاحظ النتائج الاتية:

- **عند تباين $\sigma_1^2 = 1$, $\sigma_2^2 = 1$ و أرتباط $\rho = 0.20$ ولجميع نسب الفقدان ولجميع حجوم العينات أظهرت النتائج ان التعويض بالوسط الحسابي افضل من الاسلوب المقترح، ومن الناحية العلمية يكون هذا الواقع الصحيح، حيث ان حساب الوسط الحسابي لبيانات متجانسة يكون ممثلاً لهذه البيانات بصورة دقيقة جداً ولاسيما في حالة ضعف العلاقة السببية بين هذه المتغيرات.**
-

- لجميع نسب الفقدان ولجميع حجوم العينات وعند قيم ارتباط $\rho = 0.20\ ,\ 0.30$ وكذلك عند تباينات $\sigma_1^2 = 2\ ,\sigma_2^2 = 3$ و $\sigma_1^2 = 3\ ,\sigma_2^2 = 5$ أظهرت النتائج ان الأسلوب المقترح في التعويض عن القيمة المفقودة أفضل من طريقة التعويض بالوسط الحسابي ولاسيما عند زيادة نسب الفقدان (30% , 40%).
- نلاحظ عند زيادة قيم الارتباطات وقيم التباينات استقرار الأسلوب المقترح ولكن نلاحظ تأثر وبشكل كبير لطريقة التعويض بالوسط الحسابي وهذا ما يعاب على هذه الطريقة لانه عند زيادة قيم الارتباطات وقيم التباينات تقل كفاءتها [5] .
- من الشكل ٣ و ٤ و ٥ و ٦ نلاحظ قلة التحيز للأسلوب المقترح في التعويض عن القيمة المفقودة، اي أن البيانات التي تم تقديرها تكون قريبة جداً من البيانات الحقيقية.

الأستنتاجات:

لقد تم التوصل في هذا البحث الى ان الأسلوب المقترح ذو كفاءة عالية في التقدير ويطابق الواقع العلمي ولاسيما فيما يخص حالة ارتفاع نسب الفقدان وكذلك في حالة زيادة تشتت البيانات، كذلك تم التوصل في هذا البحث الى أمكانية الاسلوب المقترح من استخلاص المعلومات المناسبة في تقدير القيمة المفقودة لاحد المتغيرات بالاعتماد على العلاقة السببية التي تربطه بمتغير اخر.

المصادر:

١. القزاز، قتيبة نبيل نايف، ٢٠٠٧، "مقارنة أساليب بيز الحصين مع طرائق أخرى لتقدير معالم أنموذج الانحدار الخطي المتعدد في حالة البيانات غير التامة"، اطروحة دكتوراه، قسم الإحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

٢. القزاز، قتيبة نبيل نايف والسامرائي، مناف يوسف حمود، ٢٠٠٨، " مقارنة طرائق التعويض الأحادي عن القيمة المفقودة لأنموذج الانحدار اللامعلمي"، مجلة العلوم الإدارية والاقتصادية (كلية الإدارة والاقتصاد – جامعة بغداد)، المجلد ١٥، العدد ٥٣.

3. Hardle; Wolfgang; 1994; “Applied Nonparametric Regression” Humboldt – University; Berlin.
4. JON S. HORNE & EDWARD O. GARTON; 2010 “Likelihood Cross-Validation versus Least Squares Cross-Validation for Choosing the Smoothing Parameter in Kernel Home-Range Analysis” The Journal of Wildlife Management Volume 70, Issue 3, pages 641–648
5. Little; Roderick J. A.; 1992; “Regression with Missing X’s: A Review”; JASA; Vol. 87; No. 420.

Proposed method to estimate missing values in Non - Parametric multiple regression model

Abstract:

In this paper, we will provide a proposed method to estimate missing values for the Explanatory variables for Non-Parametric Multiple Regression Model and compare it with the Imputation Arithmetic mean Method, The basis of the idea of this method was based on how to employ the causal relationship between the variables in finding an efficient estimate of the missing value, we rely on the use of the Kernel estimate by Nadaraya – Watson Estimator , and on Least Squared Cross Validation (LSCV) to estimate the Bandwidth, and we use the simulation study to compare between the two methods.

Keyword/ Non-Parametric Multiple Regression Model- missing observation, Missing Data Mechanisms- Patterns of Missing Data- Nadaraya – Watson Estimator- Least Squared Cross Validation