مقارنة بين طريقتي التعويض المتعدد و التعويض المتعدد ثم الحذف في معالجة القيم المفقودة لمتغيرات أنموذج الانحدار الخطي المتعدد

أ.م.د. أحمد شاكر المتولي* منال جبار سلمان**

ahmutwali@uomustansiriyah.edu.iq

المستخلص:

تواجه عملية جمع البيانات لمتغيرات البحث في بعض الحالات فقدان بعض قيم مشاهدات تلك المتغيرات ويعود ذلك إلى عدة أسباب منها ما هو خارج عن ارادة الباحث واخرى تكون مقصودة، وبهذا فان البيانات توصف بانها غير تامة وأن أعتمادها في عملية التحليل تؤدي إلى نتائج غير دقيقة الامر الذي يدعو إلى تقدير قيم المشاهدات المفقودة في متغيرات البحث ومن ثم اعتماد البيانات التامة في عملية التقدير. تناول البحث طريقتين من طرائق معالجة البيانات غير التامة لمتغيرات أنموذج الانحدار هما ، طريقة التعويض المتعدد (Multiple imputation (MI))، وطريقة التعويض المتعدد ثم الحذف (multiple imputation, then deletion) ، اذ يهدف البحث إلى آجراء مقارنة بين هاتين الطريقتين في محاولة للتوصل إلى أفضل طريقة بالاعتماد على تجارب المحاكاة أفرزت نتائج تجارب المحاكاة وباستعمال متوسط مربعات الاخطاء لانموذج الانحدار المقدر ، كمعيار للمفاضلة ، إلى أن طريقة التقدير المتعدد ثم الحذف أفضل من طريقة التقدير المتعدد في معالجة البيانات غير التامة.

Comparison between the multiple imputation and the multiple imputation then deletion methods in treating missing values for the variables of multiple linear regression model

Abstract

process of collecting data for search variables in some cases faces the loss of some values that variables, due to several reasons, some of them is outside the will of the , others reasons is intentional, Thus, the data are described as incomplete and their reliability in the analysis leads to inaccurate results, Which calls for estimating the value of missing observations in the search variables and then adopting the complete data in the estimation process. The study dealt with two methods of incomplete data processing for regression model (Multiple Imputation(MI) and multiple imputation, then deletion (MID)).

The aim of the research is to compare these two methods in an attempt to arrive at the best method based on simulations , The results of the simulation experiments and using the mean error squares of the estimated regression model, as a criterion for differentiation, showed that the multiple estimation method then deletion were better than the multiple estimation method in incomplete data processing.

* الجامعة المستنصرية / كلية الادارة والاقتصاد .

** باحثة .

مستل من رسالة ماجستير

مقبول للنشر بتأريخ 2018/1/4

1- المقدمة وهدف البحث ومشكلة البحث:

1-1 المقدمة:

ترافق عملية جمع البيانات عدة مشاكل أهمها عدم أمكانية تسجيل كافة البيانات ويعود ذلك لعدة أسباب منها ما هو غير مقصود والمتمثلة بحالات التلف عند آجراء التجارب او الفقدان بسبب الحروب أو الكوارث أو المخاطرة عند جمع البيانات . ومنها ما هو مقصود والتي تشمل حالات الفقدان بسبب التكاليف الباهظة أو الاهمال أو سرية البيانات. وبالتالي حدوث مشكلة الفقدان في قيم مشاهدات المتغيرات موضوع البحث مما يؤثر ذلك في تحليل البيانات وتقدير المعلمات بالنسبة لتحليل الانحدار الامر الذي يدعو إلى البحث عن طرائق لمعالجة مشكلة البيانات غير التامة . ومن الجدير بالملاحظة أن فقدان القيم المشاهدة للمتغيرات تحدث وفق أنماط مختلفة وآليات متنوعة. هناك عدد من طرائق تقدير البيانات غير التامة والتي تم توضيفها من قبل الباحثين والتي تهدف إلى معالجة مشكلة الفقد ومن ثم تحويل البيانات من بيانات غير تامة إلى أخرى تامة والتي يمكن أعتمادها في عملية التحليل وبناء النماذج وتقدير معلمات تلك النماذج . أهتم البحث في التطرق إلى طريقتين من طرائق معالجة البيانات غير التامة لمتغيرات أنموذج الانحدار المتعدد وهما ، طريقة التقدير المتعدد وطريقة التقدير المتعدد ثم الحذف ، ثم آجراء مقارنة بين هاتين الطريقتين بالاعتماد على تجارب المحاكاة وباستعمال متوسط مربعات الاخطاء لانموذج الانحدار المقدر كمعيار للمفاضلة .

لغرض تحقيق الهدف المنشود من البحث فقد تم تقسيمه إلى أربعة جوانب ، الاول تضمن على المقدمة ومشكلة وهدف البحث ، والثاني شمل على الجانب النظري، والثالث فتمثل بالجانب التجريبي ومناقشة نتائج تجارب المحاكاة المستعملة في التوصل إلى المفاضلة بين الطريقتين موضوع البحث أما الجانب الرابع فشمل على الاستنتاجات والتوصيات.

1- 2 مشكلة البحث Research problem

أن مشكلة البحث تتلخص في عدم أمكانية الاعتماد على البيانات غير التامة في تقدير معلمات أنموذج الأنحدار اذ تودي إلى نتائج غير دقيقة مما يتوجب معالجتها ومن ثم أعتمادها في عملية التقدير.

1- 3 هدف البحث: Objective of The Research

يهدف البحث الى أجراء مقارنة بين طريقة التعويض المتعدد (Multiple Imputation (MI)) و طريقة التعويض المتعددة ثم حذف (multiple imputation, then deletion (MID)) المستعملتان في معالجة القيم المفقودة لمتغيرات أنموذج الانحدار المتعدد (الاستجابة والتوضيحية).

2- الجانب النظري:

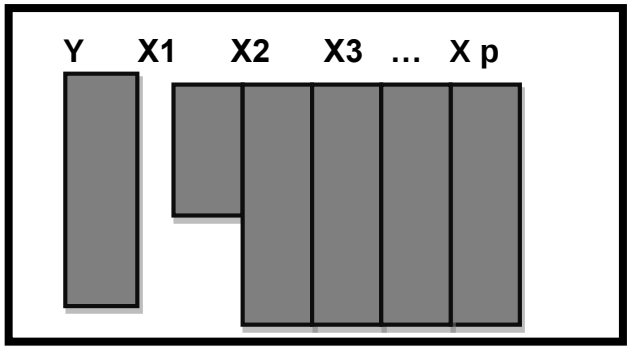
يتناول هذا الجانب اهم المفاهيم النظرية المتعلقة بالبيانات غير التامة لمتغيرات أنموذج الانحدار المتعدد وطريقتي التقدير (التعويض المتعدد و التعويض المتعدد ثم الحذف).

1-2 أنماط واليات البيانات المفقودة:

تحدث عملية فقدان البيانات للمتغيرات موضوع البحث وفق أنماط متعددة وفيما يلي عرض مختصر لتلك الآنماط :

1-2 -1 نمط البيانات المفقودة المنفرد (Univariate missing data)

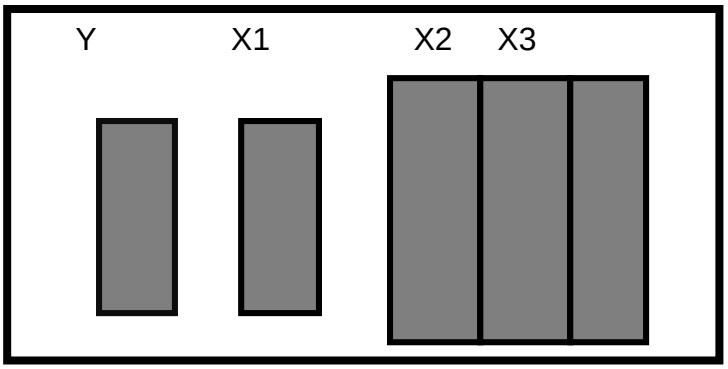
هذا النمط هو ابسط حالة ضمن حالات البيانات غير التامة والتي يكون فيها مشاهدات جميع المتغيرات تامة عدا متغيراً واحداً فقط يتضـــمن قيماً مفقودة في مشـــاهداته، وفيما يتعلق بأنموذج الانحدار فقد يكون المتغير ذو القيم المفقودة احد المتغيرات التوضيحية أو متغير الاستجابة. وهذا النمط يمكن تمثيله بالشكل (1) والذي يمثل الفقد في أحد المتغيرات التوضيحية،[3],[5],[1],[17],[16] .



الشكل (1)
يمثل فقد المشاهدات بشكل مفرد

2-2-2 نمط المتغيرين (Multivariate two pattern)

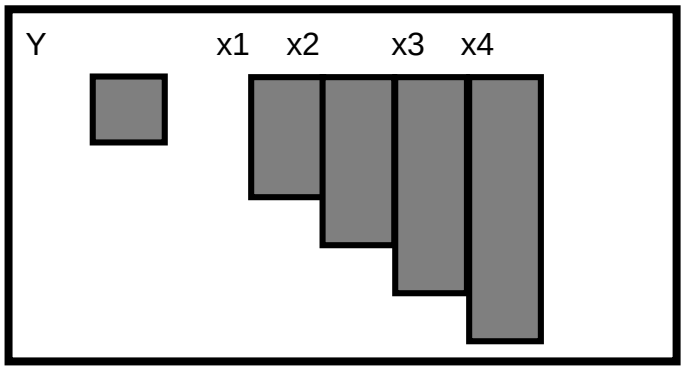
في هذا النمط يحدث فقدان قسم من البيانات في متغيرين فقط من المتغيرات الداخلة في عملية تحليل البيانات، وبخصوص أنموذج الانحدار فقد يكون هذين المتغيرين من بين المتغيرات التوضيحية او من بين متغير الاستجابة والمتغيرات التوضيحية معاً . كما موضح بالشكل (2)،[3],[1],[4],[17],[16].



الشكل (2)
يمثل فقد المشاهدات بشكل ثنائي

2-1-3:النمط الرتيب أو المتداخل (Monotone pattern)

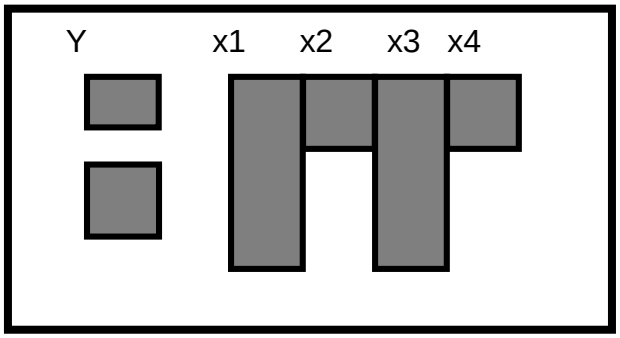
وفق هذا النمط ترتب المتغيرات حسب عدد المشاهدات المفقودة وبشكل تصاعدي أو تنازلي بان يكون المتغير ذو البيانات التامة بالترتيب الاول وبعد ذلك عدد أقل من المشاهدات المفقودة وهكذا الى اخر متغير والذي يحتوي على اكبر عدد من المشاهدات المفقودة (أو بالعكس) وكما ذكرنا في النمطين السابقين فان المتغيرات ذات البيانات غير التامة قد تكون ضمن المتغيرات التوضيحية او التوضيحية والمتغير المفسر معاً. وكما مبين في الشكل (3) ،[1],[4] ،[17].



الشكل (3)
يمثل فقد المشاهدات بشكل متداخل

2- 1-4 النمط العام (General pattern)

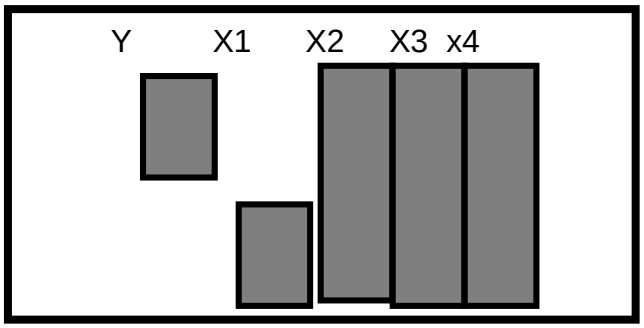
في هذا النمط تكون حالات البيانات غير التامة عشوائية ولا يوجد نمط محدد لها ، وقد يحدث فقدان المشاهدات في المتغيرات التوضيحية أو في المتغيرات التوضيحية ومتغير الاستجابة معاً وكما موضح بالشكل (4)،[3],[1],[4],[17],[16] .



الشكل (4)
يمثل فقد المشاهدات بشكل عمومي

5-1-2 النمط المتماثل (File matching pattern)

في هذا النمط يكون فقد البيانات في متغيرين فقط (بين المتغيرات التوضيحية او بين المتغيرات التوضيحية ومتغير الاستجابة معاً) بحيث لا توجد بيانات مفقودة مشتركة بين المتغيرين . وكما مبين في الشكل (5). [2],[6],[1],[11],[13] .



الشكل (5)
يمثل فقد المشاهدات بشكل متماثل

اما اليات فقد البيانات فتصنف إلى ثلاث أصناف هي الفقد العشوائي (Missing at Random (MAR) ، وغير العشوائي ((Missing not at Random (MNAR)، وتام العشوائية (Miss Completely at Random (MCAR))

3-2 طرائق معالجة القيم المفقودة Modalities for treating with missing values

هناك عدة طرائق لتقدير القيم المفقودة في متغيرات أنموذج الانحدار والتي تعتمد على نوع الفقد واليته. وقبل التطرق لتلك الطرائق نبين أدناه كيفية تمثيل العلاقة بين متغير الاستجابة وعدة متغيرات تفسيرية تتضمن قيماً مفقودة ،[14] .
ان العلاقة بين المتغيرات التوضيحية والمتغيرات المفسرة يمكن تمثيلها بنموذج الانحدار الخطي الأتي :

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \quad \text{.....(1)}$$

اذ ان Y هو متجه مشاهدات متغير الاستجابة ذو السعة (nx1) و X هو مصفوفة قيم المتغيرات التفسيرية ذات سعة (nxp) و$\underline{\varepsilon}$ هو متجه قيم البواقي ذو السعة (nx1) والذي يفترض ان تكون مستقلة وتتوزع وفق التوزيع الطبيعي $\varepsilon \sim N(\underline{\beta}, \sigma^2 I_n)$.

وبالاعتماد على طريقة المربعات الصغرى او طريقة الامكان الاعظم فان تقدير معلمات أنموذج الانحدار يكون وفق الصيغة الاتية

$$\beta = (X'X)^{-1}X'Y \quad \text{.....(2)}$$

وتحت تحقق فروض التحليل فأن تقديرات المربعات الصغرى تتصف بأنها أفضل تقدير خطي غير متحيز (Best Linear Unbiased Estimation)($BLUE$)

في حالة وجود بيانات غير التامة سواء كان في متغير الاستجابة أو المتغير التوضيحي أو في كلاهما فأن مصفوفة البيانات يمكن كتابتها بالشكل الاتي:

$$[Y\ X]=\begin{bmatrix} y_1 & x_{11} & x_{12} & . & . & x_{1p} \\ y_2 & . & . & . & * & . \\ * & . & . & . & . & * \\ . & * & . & . & . & . \\ y_n & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix}$$

اذ ان (*) تشير إلى القيم المفقودة في متغيرات عينة البحث . وبذلك يمكن تجزئة قيم متغيرات البحث بشكل عام إلى مجموعتين الأولى تضم القيم المشاهدات والمتمثلة بالرموز (X_{obs}, Y_{obs}) والثانية تضم قيم غير المشاهدة (المفقودة) والمتمثلة بالرموز (Y_{mis}، X_{mis})وبالاعتماد على تلك التجزئة يمكن كتابة معادلة انموذج الانحدار بالشكل الاتي :-

$$\begin{bmatrix} y_{obs} \\ y_{miss} \end{bmatrix} = \begin{bmatrix} \underline{1} & x_{obs} \\ \underline{1} & x_{miss} \end{bmatrix} \underline{\beta} + \begin{bmatrix} \varepsilon_{obs} \\ \varepsilon_{miss} \end{bmatrix} \quad \text{.....(3)}$$

اذ أن Y_{obs} متجه قيم متغير الاستجابة المشاهدة فعلاً ذو سعة (mx1)، Y_{mis} متجه قيم متغير الاستجابة المفقودة ذو سعة (n-m X 1)، اذ أن X_{obs} مصفوفة قيم المتغيرات المفسرة المشاهدة فعلاً ذو سعة (mxp)، X_{mis} مصفوفة قيم المتغير المفسرة المفقودة ذو سعة (n-m X p) ، $\underline{1}$ متجه احادي سعته تستمد من مصفوفة المعلومات ، ϵ_{obs} متجه قيم الاخطاء العشوائية ذو سعة (mx1) و ϵ_{mis} متجه قيم الاخطاء العشوائية ذو سعة (n-m X 1) ومنه نستطيع كتابة أنموذج الانحدار الجزئي الذي يضم المتغيرات ذات القيم المشاهدة فعلاً وبالشكل الاتي

$$\underline{Y}_{obs}= [\underline{1} \quad X_{obs}]\ \underline{\beta} + \underline{\varepsilon}_{obs} \quad \text{. .....(4)}$$

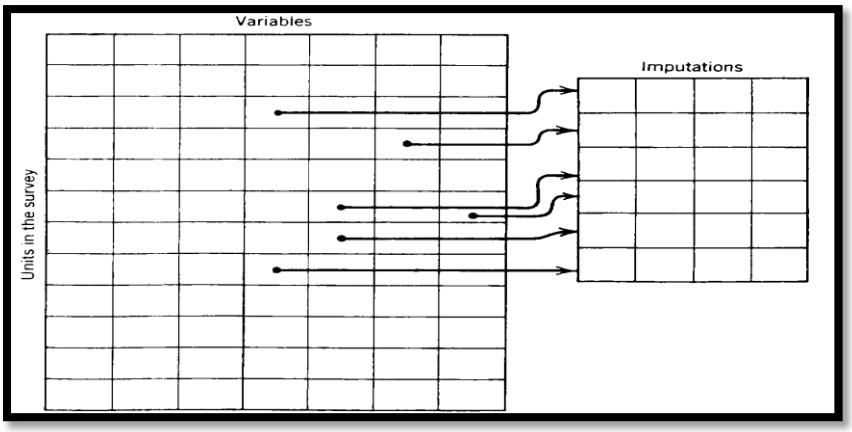
أما الجزء الثاني من معادلة الانحدار الخطي فيتضمن البيانات المفقودة وكذلك فأن أنموذج الانحدار الجزئي الذي يضم المتغيرات ذات القيم المفقودة يكون كالأتي ، [19]،[17]،[12]،[[14].

$$\underline{Y}_{miss}= \begin{bmatrix} 1 & X_{miss} \\ \underline{} & \end{bmatrix} \underline{\beta} + \varepsilon_{miss} \quad \text{.....(5)}$$

1-3-2 : طريقة التعويض المتعدد للقيم المفقوده Multiple Imputation Method

طريقة التعويض المتعدد للقيم المفقودة تم اقتراحها من قبل الباحث Rubin ، (1977) [15]، بموجب هذه الطريقة يتم التعويض عن كل قيمة مفقودة بعدد (M ≥2) من القيم التقديرية المحتملة، بمعنى ان كل قيمة مفقودة يتم تعويضها بمتجه من القيم التقديرية ذو سعة (M×1) ، تلك القيم التعويضية ترتب بحيث اول قيمة من كل متجه من متجهات القيم التعويضية والخاصة بكل قيمة مفقودة تستخدم لتكوين مجموعة من البيانات الكاملة ، والقيمة الثانية من كل متجه تستخدم لتكوين مجموعة ثانية من البيانات الكاملة وهكذا الى اخر قيمة من قيم متجهات القيم التعويضية، وبذلك تتكون لدينا (M) من مجموعات البيانات الكاملة وكل مجموعة منها يتم تحليلها بشكل مستقل. الشكل (6) يبين بيانات التعويض المتعدد للقيم المفقودة.

القيم التعويضية المحتملة (M) هي عبارة عن (M) من التكرارات التي يتم توليدها بالاعتماد على التوزيع التنبؤي اللاحق للقيم المفقودة للمتغير وكل تكرار يستند على اختبار مستقل لمعلمات توزيع القيم المفقودة ووفق نماذج بيزية مناسبة للبيانات،[5]،[7] ،[15]



الشكل (6)
توضيح عمل طريقة التعويض المتعدد للقيم المفقوده (MI)

التعويض المتعدد للقيم المفقودة يحتفظ بالميزتين الاساسيتين للتقدير المنفرد والمتمثلة بأمكانية أستعمال طرائق تحليل البيانات الكاملة وامكانية تضمين معرفة جامع البيانات وفي الحقيقة فان طريقة التعويض المتعدد لا تحتفظ فقط في الميزة الاساسية الثانية وانما تعمل على تحسينها اذ انها تسمح لجامعي البيانات لاستخدام معرفتهم لتعكس حالة عدم التأكد حول أي القيم التي سوف يتم اعتمادها. هناك ثلاثة ميزات مهمة تجعل طريقة التعويض المتعدد افضل من طريقة التعويض المنفرد الاولى انها تزيد من كفاءة التقدير وذلك لان التعويضات المتعددة يتم اختيارها بشكل عشوائي بالاعتماد على التوزيع الاحتمالي للبيانات .الميزة الثانية هو أمكانية الحصول على استدلالات فعاله والتي يتم الحصول عليها من خلال دمج الاستدلالات المستحصل عليها من مجاميع البيانات الكاملة اذ ان التعويضات المتعددة تمثل تكرارات عشوائية يتم اختيارها بالاعتماد على انموذج عدم الاستجابة .اما الميزة الثالثة فتتمثل بأمكانية اجراء دراسة دقيقة لحساسية الاستدلالات بالاعتماد على البيانات الكاملة المتكررة والتي تم توليدها بشكل متكرر وعشوائي بالاعتماد على اكثر من انموذج للبيانات.

أما عيوب هذه الطريقة فيمكن تلخيصها في أنها تؤدي إلى نتائج مختلفة في كل مرة تستخدم فيها، وذلك لأن القيم التعويضية هي عشوائية وليست ثابته. وكذلك وجود العديد من طرائق تقدير القيم المتعددة منها ، التقدير باستخدام الوسط الحسابى غير الشرطي أو الوسيط أو المنوال ، و التقدير باستخدام القيمة المتوقعة العظمى Expected Maximization Imputation والتقدير عن طريق تحليل الانحدار (Regression analysis) والتقدير باستخدام سلسلة ماركوف مونت كارلو ، مما قد يؤدي إلى التخبط والارتباك في اختيار طريقة التعويض من بين تلك الطرائق ،[5]،[15] ،[4].

1-1-3-2 : تحليل مجاميع بيانات طريقة التعويض المتعدد
analysis of aggregates data for Multiple Imputation Method

بينا سابقاً ان عدد (M) من التعويضات المتعددة للقيم المفقودة تولد (M) من مجاميع البيانات الكاملة كل منها يمكن تحليلها بطرائق التحليل المعتمده بعد ذلك يتم دمج نتائج تحليل (M) من مجاميع البيانات الكاملة لنحصل على استدلال واحد لطريقة التعويض المتعدد.

لنفرض ان β تمثل متجه المعلمات المراد تقديره ذو سعة (1×K) ، وعلى وجه الخصوص β تمثل متجه معلمات انموذج الانحدار الخطي البسيط أي ان

$\beta = (\beta_0 ، \beta_1)$ وعلى افتراض ان الاستدلال حول متجه المعلمات β يستند على التوزيع الطبيعي وكالاتي

$$(\beta - \hat{\beta}) \sim N(0,\sum) \quad \text{.....(6)}$$

اذ ان $\hat{\beta}$ هو تقدير متجه المعلمات β ، $\sum$ تمثل مصفوفة التباين والتباين المشترك لمتجه المعلمات المقدرة ، $N(0,\sum)$ هو التوزيع الطبيعي متعدد المتغيرات Multi- Variate normal (distribution) بمتوسط يساوي صفر وتباين $\sum$ تحت انموذج بيز محدد وباعتماد طريقة التعويض المتعدد للقيم المفقودة فأننا نحصل على (M) من التقديرات $\hat{\beta}_l$ اذ ان l =1,..,M وتباياناتها، Σ_l ، $l = 1,2,3,....M$ ، والتي نحصل عليها من كل مجموعات البيانات الكاملة بعد التعويض المتعدد للقيم المفقودة . ومن تلك التقديرات المتعددة نحصل على التقدير النهائي لمتجه المعلمات β وكالاتي

$\bar{\bar{\beta}} = \frac{\sum_{l=1}^{M} \hat{\beta}_l}{M}$ .....(7)

والتباين المرتبط بهذا التقدير هو معدل التباين لـ M في مجاميع البيانات الكاملة (والذي يمثل معدل التباين داخل التكرارات) يكون كالات

$\bar{\sum} = \frac{\sum_{l=1}^{M} \Sigma_l}{M}$ .....(8)

اما التباين بين تقديرات (M) من مجاميع البيانات الكاملة يكون كالاتي :-

$B_m = \frac{\sum_{l=1}^{M} (\hat{\beta}-\bar{\beta})'(\hat{\beta}-\bar{\beta})}{M-1}$..... (9)

وبذلك فان التباين الكلي الخاص بالتقدير النهائي $\bar{\bar{\beta}}$ يكون وفق الصيغة الاتية:

$T = \bar{\sum} + (1+M^{-1}) B$ .....(10)

التوزيع الاحتمالي التقاربي الذي يستخدم في تقدير فترة الثقة واختبارات المعنوية هو توزيع ستيودنت t-distribution اذ ان

$(\bar{\sum} - \bar{\bar{\beta}}) T^{-1/2} \sim t_v$ .....(11)

وان درجة الحرية V تحسب وفق الصيغة الاتية

$V = (M-1)(1+r^{-1})^2$.....(12)

اذ ان r تمثل الزيادة النسبية في التباين نتيجة لعدم الاستجابة المؤدية لوجود القيم المفقودة ، والتي تحسب كالاتي:

$r = (1+M^{-1}) tr(B \bar{\sum}^{-1})/K$ .....(13)

وعليه فان تقدير فترة الثقة لمتجه المعلمات β ولمستوى معنوية α تكون كالاتي:-

$\bar{\bar{\beta}} \mp t_{(V, \alpha/2)} T^{1/2}$ .....(14)

نسبة المعلومات المفقودة حول قيمة المعلمات (β) نتيجة عملية فقدان البيانات تكون وفق الصيغة الاتية

$\gamma = \frac{r+2/(V+3)}{r+1}$ ..... (15)

اما مستوى المعنوية بقيمة متجه المعلمات β تحت صحة فرضية العدم (β_0) عندما تكون تكرارات القيم المقدرة للقيم المفقودة (M) كبيرة نسبياً مقارنة مع عدد عناصر متجه المعلمات (K) يمكن حسابه كالاتي.

$D = (\beta_0 - \bar{\bar{\beta}}) T^{-1} (\beta_0 - \bar{\bar{\beta}})^{-} /K$ .....(16)

واعتبار مستوى المعنوية هذا يمثل احتمال ان المتغير العشوائي F (بدرجتي حرية K و V) اكبر من D أي ان :-

$Pr\{F_{(K,V)} > D\}$.....(17)

افضل اختبار لفرضية العدم وعندما تكون M ذو قيمة معتدلة قياساً بعدد عناصر متجه المعلمات (K) يكون وفق الصيغة الاتية.

$\tilde{D} = (1+r)^{-1} (\beta_0 - \bar{\bar{\beta}}) \bar{\Sigma}^{-1} (\beta_0 - \bar{\bar{\beta}})^{-} /K$ .....(18)

والتي تشير الى توزيع F بدرجتي حرية (K, (K+1)V/2)

احصاءة مكافئة تقريباً للاحصاء $\tilde{D}$ والمبينة بالصيغة (18) يمكن حسابها من (M) من مستويات المعنوية المتعلقة بقيمة متجه المعلمات تحت صحة فرضية العدم (β_0) وقيمة الزيادة النسبية في التباين (r) المبينة بالصيغة (13) والمستحصل عليهم من البيانات المتكررة. فعلى افتراض ان $d_1, d_2, \ldots, d_M$ عبارة عن M من القيم المكررة لاحصاءة مربع كاي (χ^2) لمجاميع البيانات الكاملة ، والمتعلقة بقيمة β_0 بمعنى ان مستوى المعنوية للمجموعة ذي التسلسل L من بين مجاميع البيانات الكاملة هو احتمال ان يكون المتغير العشوائي (مربع كاي) بدرجة حرية (K) اكبر من (d_L) ، $(L=1,2,\ldots,M)$.عندها

$\hat{D} = \frac{\frac{\bar{d}}{K} - \frac{M-1}{M+1} r}{1+r}$ .....(19)

اذ ان

$\bar{d} = \frac{\sum_{l=1}^{M} d_l}{M}$..... (20)

والذي يمثل معدل قيم احصاءات مربع كاي لـ M من التكرارات .الاحصاء $\widehat{D}$ تكون اكثر فائدة من D أو $\widetilde{D}$, لانها تعتمد على قيم مربع كاي المكررة ($d_1,d_2,\ldots,d_M$) وقيمة الزيادة النسبية r بدلاً من الاعتماد على مصفوفة من سعة (k × k) [15].

2-1-3-2: تقدير دالة الامكان وفق طريقة التعويض المتعدد للقيم المفقودة

لتكن Y_{obs} ، Y_{mis} تمثل القيم المشاهدة والقيم المفقودة لمتغير الاستجابة Y على التتابع ، X_{obs}، X_{mis} تمثل القيم المشاهدة والقيم المفقودة للمتغير التفسيري X، فان دالة الامكان الأعظم لمتجه المعلمات β بمعلومية القيم المشاهدة والقيم المفقودة بعد تعويضها بطريقة التعويض المتعدد و M من مجاميع البيانات الكاملة تقدير وفق الصيغة الاتيه

$$\hat{L}(\beta^{(m)}/ X_{mis}^{(m)}, X_{obs}, Y_{mis}^{(m)}, Y_{obs}) = f(Y_{mis}^{(m)}, Y_{obs} / X_{mis}^{(m)}, X_{obs}) \ldots\ldots\ldots (21)$$

اذ ان m=1,2,..,M

بعدها يتم حساب متوسط تلك التقديرات عبر M من جاميع البيانات الكاملة لنحصل على الوسط الحسابي لـ M من تقديرات الأمكان الأعظم وفق الصيغة الأتية

$$\hat{L}_{M,MI}(\beta/X_{obs}, Y_{obs}) = \frac{1}{M}\sum_{m=1}^{M}\hat{L}^{(m)}(\beta/ X_{mis}^{(m)}, X_{obs}, Y_{mis}^{(m)}, Y_{obs}) \ldots\ldots (22)$$

والتي تمثل تقدير طريقة MI لدالة الأمكان لمتجه المعلمات β، [20]، [10].

2-3-2 : طريقة التعويض المتعددة ثم الحذف multiple imputation, then deletion (MID)

تم اقتراح هذه الطريقة من قبل الباحث Vonn Hippel في عام 2007، فكرة هذه الطريقة مشابهة تقريباً لفكرة طريقة التعويض المتعدد للقيم المفقودة MI مع اختلاف بسيط هو حذف القيم التعويضية لمتغير الاستجابة قبل اجراء عملية التحليل .فعلى افتراض ان Y متغير استجابة العلاقة بعدة متغيرات توضيحية ($X_1,X_2,\ldots,X_P$) وان هناك قيم مفقودة بشكل عشوائي في متغير الاستجابة وكذلك المتغيرات التوضيحية ، فبالاعتماد على هذه الطريقة يتم اولاً التعويض عن القيم المفقودة باستعمال طريقة تقدير المتعدد للقيم المفقودة ، بالنظر لكون القيم التعويضية لمتغير الاستجابة Y لا تضيف معلومات حول انحدار متغير الاستجابة Y على المتغير التوضيحي X وذلك لان لوغاريتم دالة الامكان لحالات متغير الاستجابة الذي تم تعويض قيمها المفقودة يساوي وبالضبط الصفر اذا ان دالة الامكان لتلك الحالات تساوي =1 وكما مبين :

$$L_{Ymis}(\beta/ X_{obs|ymis}) = \int f(Y_{mis}/X_{obs|Ymis}, \beta)\, dY_{mis}=1 \ldots\ldots\ldots (23)$$

$$\log L_{ymis}(\beta/Y_{obs|ymis}) = 0 \ldots\ldots\ldots (24)$$

ولكن أضافتها في عملية التعويض المتعدد للقيم المفقودة يؤدي إلى تحسين القيم التعويضية للمتغيرات التوضيحية لذا يتم حذف الحالات المتضمنة قيماً تعويضية لمتغير الاستجابة بعد الأعتماد عليها في عملية التعويض المتعدد للقيم المفقودة بمعنى اخر يتم حذفها قبل اجراء عملية تحليل البيانات وتقدير معلمات أنموذج الانحدار. كما أن الأبقاء على تلك القيم التعويضية يؤدي إلى تقديرات مشوشة (noise estimates)، [20]، [18]، [10] .

2-3-2-1: تقدير دالة الامكان الأعظم بموجب طريقة MID .

تقدير دالة الامكان الاعظم بموجب طريقة MID يكون مشابه لتلك الدالة المقدرة بموجب طريقة MI ولكن بكفأة أكثر وذلك يعود إلى الاعتماد على قيم لمتغير الاستجابة المشاهدة فعلاً اذ كما بينا سابقاً أن القيم المفقودة لهذا المتغير بعد تعويضها بموجب هذه الطريقة لا تضيف أية معلومة لدالة الأمكان .ففي حالة كون القيم المفقودة في كل من متغيري الاستجابة ومتغيرات التوضيحية فان دالة الامكان تكون كالاتي

$$L(\beta|X_{obs}, Y_{obs}) = \int L(\beta/Y_{obs}X_{obs},Y_{mis},X_{mis},Y_{mis})P(X_{mis},Y_{mis}/X_{obs},Y_{obs})dX_{mis}dY_{mis} \ldots\ldots\ldots\ldots (25)$$

وان صيغة دالة الامكان المقدرة باستخدام القيم المشاهدة فقط لمتغير الاستجابة تكون بالشكل الاتي

$$\hat{L}_{Yobs}^{(m)}(\beta/ X_{mis}^{(m)}, X_{obs}, Y_{mis}^{(m)}, Y_{obs}) = f(Y_{obs}/X_{mis}^{(m)}, Y_{obs}, X_{obs}, \beta) \ldots\ldots\ldots (26)$$

ومتوسط تلك التقديرات عبر (M) من مجاميع البيانات بعد تعويض القيم المفقودة تمثل تقدير دالة الامكان بموجب طريقة MID وكما مبين في المعادلة الاتية:

$$\hat{L}_{M,MID}(\beta/X_{obs}, Y_{obs}) = \sum_{m=1}^{M} \hat{L}_{Y_{obs}}^{(m)}(\bar{\beta}/ X_{mis}^{(m)}, X_{obs}, Y_{obs}) \ldots (27)$$

وبمقارنة هذه الصيغة المقدرة مع الصيغة المقدرة بموجب طريقة MI والمبينة بالمعادلة (22) نجد ان هذه الصيغة تكون أكثر كفأة وذلك لان الصيغة المقدرة بموجب MI لها مصادر اضافية لخطأ التقدير وذلك لاعتمادها على قيم Y المفقودة بعد التعويض . بينما نجد ان طريقة MID تتجنب ذلك الخطأ اذا بفتراضها ان دالة الامكان لحالات قيم متغير الاستجابة المفقودة يكون مساوي الى الواحد الصحيح كما في المعادلة (23)، [20]،[18].

2-3-2-2: تقدير معلمات أنموذج الانحدار بموجب طريقة MID

التقدير بموجب MID يعتمد نفس اسلوب تقدير طريقة MI ولكن بموجب طريقة MID فان الحالات المعتمدة في التحليل تكون اقل (حجم العينة يكون اقل من حجم العينة المعتمد بطريقة MI) اذ كما بينا سابقاً ان بموجب هذه الطريقة يتم التعويض عن القيم المفقودة في متغيري الاستجابة والتفسيري وبعد ذلك أي قبل اجراء تحليل البيانات يتم حذف كل الحالات التي تضمنت قيم مفقودة لمتغير الاستجابة والتي تم تعويضها بموجب هذه الطريقة. مما تقدم فان صيغ التقدير تكون نفس صيغ التقدير لطريقة MI مع اختلاف في درجات الحرية فقط ، أي تقليل درجات الحرية. نتيجة التخفيض الحاصل في حجم العينة بمعنى ان الاختلاف بين الصيغ التقديرية لطريقتي التعويض MI وMID هو فقط في درجات الحرية و للتميز بين درجات الحرية للطريقتين نفرض لدرجة الحرية لمقدر MI بـ $V_{com,MI}$ ولدرجة الحرية لمقدرMID بـ $V_{com,MID}$، ($V_{com,MI}>V_{com,MID}$). ومن الجدير بالملاحظة قد يفهم خطأ ان حدود الثقة لمعلمات الأنموذج وفق طريقة MI لها درجات حرية اكبر من حدود الثقة بموجب طريقة MID ولكن بالمقارنة طريقة MI لها نسبة معلومات مفقودة حول قيمة المعلمات β اكبر من تلك النسبة الخاصة بطريقة MID

أي ان :

$$\gamma_{\infty,MID} \leq \gamma_{\infty,MI} \quad \text{.....(28)}$$

- اذ ان γ كما عرفنها سابقاً بانها تمثل نسبة المعلومات المفقودة المشاهدة للبيانات وبما ان درجة الحرية المشاهدة تساوي تقريباً :

$$V_{obs} = (1-\gamma)\, V_{com}$$

مما يؤدي إلى ان درجات الحرية المشاهدة تكون أقل بالنسبة لطريقة MI مقارنة مع درجات الحرية المشاهدة لطريقة MID ،[20]،[18]،[10]

تمتاز طريقة (MID) بكفأتها مقارنة مع طريقة التعويض المتعدد للقيم المفقودة MI اذا انها تعطي تقديرات اكثر دقة وبأقل خطا معياري مع فترات ثقة اضيق من تلك التي نحصل عليها بموجب طريقة (MI)، بالرغم من حذف Yi التي تتضمن فقد والتي غالبا ما تسبب الحذف زيادة في الانحراف المعياري داخل كل مجموعة البيانات محسوبة، ولكن عن طريق الحد من ثاثير القيم المحسوبة MID يقلل من قيمة التقديرات في كل مجموعة بيانات لذا فانه عندما يكون هناك تباين صغير داخل المجموعة افضل من ان يكون تباين صغير بين لمجموعة البيانات المحسوبة لان حساب التباين الداخلي للمجموعة يكون ادق من التباين بين مجموعاة البيانات المحسوبة لان بياناتها المحسوبة تكون اقل بكثير من عدد الحالات في كل منها أما فيما يتعلق باختبار الفرضيات فانها تؤدي إلى اختبارات ذو قوة أكبر والميزة الاخرى والتي تعد الاهم هو انها تعتبر حصينة لمشاكل نماذج التعويض وكذلك يمكن ادراج المتغيرات المساعدة لتحسين التنبؤ من القيم المفقودة في X .

اما سلبياتها فانها تكون غير فعالة في حالة كون البيانات قليلة و نسبة الفقد في البيانات كبيرة وكذلك عند الفقد يمكن ان تسبب خسارة معلومات مهمة ،[18]،[10]،[20] .

3-الجانب التجريبي

سيتم استخدام اسلوب المحاكاة لدراسة حالات مختلفة من البيانات من ناحية توزيعها والتي تجسد المشاكل التي تواجه الباحث في حالة تحليل أنموذج الانحدار الخطي المتعدد وبوجود المشكلة الرئيسة وهي كون بيانات كل من، المتغيرات التوضيحية X's والمتغير التفسيري Y ، تعاني من فقدان في بعض مشاهداتها ،وباستخدام طريقة مونت كارلو لتوليد البيانات

أ- تم حساب قيم لمتغير الاستجابة من العلاقة الخطية الآتية و بواقع ثلاث متغيرات توضيحية:

$$Y = X\beta + e \qquad \text{...(29)}$$

أذ ان

Y : تمثل متجه قيم المتغير الاستجابة من الدرجة nx1.
X : تمثل مصفوفة المتغيرات التوضيحية من الدرجة nx3.
β : تمثل متجه معاملات الانحدار للنموذج من الدرجة 4x1.
E : تمثل متجه الخطأ العشوائي من الدرجة nx1.

ب- تم فرض قيم أولية لمعلمات الأنموذج (β) وحسب الانموذج المطلوب دراسته وبما يتوافق مع طبيعة الظاهرة المدروسة اذا تم الاعتماد على القيم الافتراضية (1،0.2،0.5،0.3) لمعلمات الانموذج ($\beta0$،$\beta1$، $\beta2$،$\beta3$) على التتابع بهذا فان أنموذج الانحدار الخطي المتعدد في عملية المحاكاة يكون كالاتي.

$$y_i = 1 + 0.2\,x_{i1} + 0.5\,x_{i2} + 0.3\,x_{i3} + e_i \quad \text{...(30)}$$

i= 0,1,2,…,n ،

ج- تم توليد البيانات بقيم مفقودة لكل من متغير الاستجابة والمتغيرات التوضيحية وفق الية الفقد العشوائي MAR وبنسب فقد (5%،10%، 15%) ونمط الفقد العمومي.

د- تم استعمال ثلاث حجوم للعينات هي (110,75,40) وبقيمتين لتباين الخطأ العشوائي هي (1.5,1) .

هـ- تم تكرار التجربة 1000 مرة بغية التوصل إلى نتائج ذات دقة عالية يتم الأعتماد عليها للمقارنة بين طريقتي التقدير المعتمدة[1]،[3] .

لغرض المقارنة بين طريقتي التقدير المعتمدة تم الاعتماد على معيار للمفاضلة وهو مقياس متوسط مربعات الاخطاء لانموذج الانحدار المقدر للمحاولة الواحدة ولنفس النموذج وكالاتي:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p} \quad \text{...(31)}$$

وبالاعتماد على متوسط قيم متوسطات مربعات الاخطاء لجميع المحاولات يتم المقارنة بين طرائق التقدير المعتمدة ،[3] ، [4].

1-3 تحليل نتائج تجارب المحاكاة:

تم تلخيص نتائج تجارب المحاكاة والخاصة بمعيار المفاضلة MMSE لكل أنموذج في جداول خاصة حسب حجم العينة ونسبة الفقد وتباين الفقد وتباين الخطأ العشوائي لطريقتي التقدير المعتمدة وكما مبين في الجداول (1) و الجداول (2) . وفيما يلي مناقشة نتائج المفاضلة وحسب تباين الخطا العشوائي .

أولاً : المقارنة بين طريقتي التقدير بافتراض تباين الخطا العشوائي يساوي (1).

أفرزت قيم معيار المفاضلة MMSE والمبينة في الجدول (1) ان طريقة التقدير MID هي الافضل بالنسبة لحجم العينة n=40 وn=110 عند كافة نسبة الفقدان اذ تراوحت قيمة معيار المفاضلة المفاضلة MMSE بين (3.0208413e-08) و (7.05409073e-08) وكذلك نجد ان هذه الطريقة هي الافضل أيضاً عند حجم العينة n=75عند نسبة فقد 15% في حين كانت طريقة MI هي الافضل بالنسبة لحجم العينة n=75 ولنسبتي الفقد 5% و10% إذ كانت قيمتا معيار المفاضلة MMSE هي الاقل إذ بلغتا (3.8901132742e-08) و(8.208959e-08) .عند نسبتي الفقد على التتابع.

جدول (1)

تقدير معلمات أنموذج الانحدار ومتوسط متوسطات مربعات الاخطاء لتلك التقديرات ومتوسط متوسطات مربعات الاخطاء للأنموذج المقدر مصنفة حسب طرائق التقدير ونسب التقدير وحجوم العينات عندما ((e ~ N (0,1)

حجم العينة	طرائق التقدير	حجم العينة / نسبةالفقدان	مقدرات المعالم $MMSE(\beta)$				MMSE
			β_0	β_1	β_2	β_3	
40	MI	5%	-0.9561 (4.2673)	0.2102 (0.4410)	0.5040 (0.4409)	0.3353 (0.4421)	1.06935948836858 e-07
		10%	0.2759 (0.5385)	0.2082 (0.0142)	0.4873 0.0143)	0.3148 (0.0144)	9.01612610066285 e-08
		15%	5.0898 (22.4914)	0.1166 (5.7716)	0.4472 (5.7675)	0.3231 (5.7652)	1.24252812364298 e-07
	MID	5%	-0.8480 (3.7879)	0.2093 (0.3729)	0.5031 (0.3728)	0.3352 (0.3741)	6.99244034686689 e-08
		10%	0.0667 (0.9039)	0.2039 (0.0329)	0.4957 (0.0329)	0.3129 (0.0330)	7.05409072550811 e-08
		15%	2.2995 (2.6859)	0.1665 (0.9983)	0.4782 (0.9976)	0.3113 (0.9973)	6.97495325712469 e-08
75	MI	5%	1.5337 (0.6604)	0.2008 (0.3756)	0.4913 (0.3757)	0.3031 (0.3757)	3.89011327442182 e-08
		10%	2.6142 (3.9377)	0.1706 (1.3331)	0.4953 (1.3322)	0.2978 (1.3322)	8.20895945490962 e-08

		15%	4.3444 (15.2620)	0.1867 (4.0770)	0.4696 (4.0778)	0.2828 (4.0771)	6.53709561751009 e-08
	MID	5%	1.6021 (0.7807)	0.2000 (0.4182)	0.4912 (0.4183)	0.3032 (0.4182)	7.70190012587593 e-08
		10%	0.8209 (0.1082)	0.1870 (0.0763)	0.5045 (0.0761)	0.3131 (0.0763)	8.90734420906872 e-08
		15%	3.0752 (6.2144)	0.1938 (1.9080)	0.4828 (1.9083)	0.2925 (1.9080)	4.66188093060254 e-08
110	MI	5%	1.9121 (1.4773)	0.1890 (0.6454)	0.5006 (0.6453)	0.2901 0.6454)	5.73822240097616 e-08
		10%	1.3203 (0.3602)	0.1992 (0.2576)	0.4871 (0.2578)	0.3132 (0.2578)	3.39647508625872 e-08
		15%	4.9313 (20.8207)	0.1675 (5.3664)	0.4721 (5.3661)	0.2772 (5.3658)	9.14187722161169 e-08
	MID	5%	0.8736 (0.1045)	0.2010 (0.0886)	0.5027 (0.0886)	0.2985 (0.0886)	3.47551816394645 e-08
		10%	0.6853 (0.1443)	0.2034 (0.0453)	0.5002 (0.0453)	0.3098 (0.0454)	3.02763744802887 e-08
		15%	0.9524 (0.1158)	0.1906 (0.1136)	0.5039 (0.1135)	0.2978 (0.1135)	3.02084139580431 e-08

ثانياً: المقارنة بين طريقتي التقدير بافتراض تباين الخطا العشوائي يساوي (1.5).

بمراجعة قيم معيار المفاضلة MMSE لطريقتي التقدير موضوع البحث والمعروضة في الجدول (2) نجد ان طريقة MID هي الافضل لنسبة فقد 5% ولحجوم العينات بالنسبة n=40 و n=75 ولنسبة فقد 10% لحجمي العينتين n=40 و n=75 ، ولنسبة فقد 15% لحجم العينة n=110 إذ كان لها أقل قيمة لمعيار المفاضلة MMSE اذ تراوحت بين (2.41184090065853e-08) و (9.95705358e-08) في حين نجد ان طريقة MI هي الافضل لنسبة فقد 15% لحجمي العينة n=40 و n=75 ، و لنسبة فقد 10% لحجم العينة n=110 ، اذ كان لها أقل قيمة لمعيار المفاضلة MMSE التي تراوحت بين (1.284563e-08) و(9.3735292e-08) .

جدول (2)

تقدير معلمات أنموذج الانحدار ومتوسط متوسطات مربعات الاخطاء لتلك التقديرات ومتوسط متوسطات مربعات الاخطاء للأنموذج المقدر مصنفة حسب طرائق التقدير ونسب التقدير وحجوم العينات عندما ((e ~ N (0,1.5)

حجم العينة	طرائق التقدير	حجم العينة / نسبةالفقدان	مقدرات المعالم $MMSE(\beta)$ β_0	β_1	β_2	β_3	MMSE
40	MI	5%	1.8764 (1.3850)	0.1904 (0.6170)	0.4989 (0.6169)	0.2909 (0.6170)	1.232029594e-07
		10%	1.5968 (0.7705)	0.1991 (0.4143)	0.4866 (0.4145)	0.3087 (0.4144)	1.2845634e-08e-07
		15%	5.4786 (26.800)	0.1124 (6.7496)	0.4498 (6.7444)	0.3168 (6.7422)	9.3735292e-08
	MID	5%	0.8490 (0.1046)	0.2023 (0.0818)	0.5009 (0.0818)	0.2990 (0.0818)	9.95705358e-08
		10%	0.9492 (0.1116)	0.2036 (0.1090)	0.4994 (0.1090)	0.3057 (0.090)	3.3551618e-08
		15%	2.4535 (3.2701)	0.1651 (1.1585)	0.4829 (1.1576)	0.3053 (1.1573)	2.0739932e-07
75	MI	5%	1.3982 (0.4561)	0.2042 (0.2976)	0.4870 (0.2978)	0.3063 (0.2976)	6.977789e-08
		10%	2.3964 (3.0439)	0.1692 (1.0948)	0.4991 (1.0939)	0.2987 (1.0939)	8.80321121600341 e-08
		15%	4.2420 (14.3822)	0.1872 (3.8715)	0.4692 (3.8723)	0.2852 (3.8715)	1.74478677114647 e-08
	MID	5%	1.5042 (0.6114)	0.2029 (0.3572)	0.4868 (0.3574)	0.3064 (0.3573)	2.53068341065894 e-08
		10%	0.6021 (0.1936)	0.1857 (0.0355)	0.5081 (0.0353)	0.3142 (0.0355)	2.41184090065853 e-08
		15%	3.1465 (6.6145)	0.1940 (2.0072)	0.4815 (2.0075)	0.2936 (2.0072)	2.73922092749502 e-08
110	MI	5%	1.8764 (1.3850)	0.1904 (0.6170)	0.4989 (0.6169)	0.2909 (0.6170)	6.39200088937020 e-08
		10%	1.5968 (0.7705)	0.1991 (0.4143)	0.4866 (0.4145)	0.3087 (0.4144)	1.28456349710223 e-08
		15%	4.7767 (19.2753)	0.1688 (5.0126)	0.4738 (5.0124)	0.2768 (5.0122)	4.67289808931142 e-08
	MID	5%	0.8490 (0.1046)	0.2023 (0.0818)	0.5009 (0.0818)	0.2990 (0.0818)	4.66613864157871 e-08

		10%	0.9492 (0.1116)	0.2036 (0.1090)	0.4994 (0.1090)	0.3057 (0.1090)	3.35516183602002 e-08
		15%	0.8259 (0.1084)	0.1922 (0.0781)	0.5055 (0.0781)	0.2969 (0.0781)	3.02084139580431 e-08

ثالثاً: المفاضلة بين طريقتي التقدير لجميع حجوم العينات المعتمدة

لتحديد أي من طريقتي التقدير موضوع البحث أفضل تم حساب عدد مرات تفوق كل طريقة خلال حجوم العينات الثلاث لجميع نسب الفقد وحسب ما تم عرضه من نتائج التقدير في الجدولين (1) و (2) ، ثم تلخيص تكرارات الافضلية في جدول (3) بمراجعة تلك التكرارات نجد ان طريقة MID هي افضل من طريقة التقدير MI بالنسبة لتباين الخطأ المفترض (1) إذ حققت أعلى تكرار والبالغ (7) بينما نجد ان هناك تقارب بين الطريقتين بالنسبة لتباين الخطأ المفترض (1.5) ومع ذلك يمكن أعتبار طريقة MID هي الافضل .

مما تقدم يمكن القول ان طريقة التقدير المتعدد ثم الحذف هي أفضل من طريقة التقدير المتعدد على ضوء مافرزته نتائج تقدير تجارب المحاكاة لحجوم العينات ولقيمتي الخطأ العشوائي التي تم أفتراضها.

جدول (3)

ملخص لعد تكرار أفضلية طرائق التقدير بالنسبة لجميع حجوم العينات ونسبة الفقدان بالاعتماد على نسبة تباين الخطأ

تباين الخطأ	طرائق التقدير	
	MI	MID
1	2	7
1.5	4	5

4- الاستنتاجات والتوصيات The Conclusions and recommendations:

بناءً على ما تم التوصل اليه من نتائج في الجانب التجريبي يمكن ادراج الاستنتاجات الاتية:

1- عند تباين الخطا العشوائي (1) وجد ان طريقة التعويض المتعدد ثم الحذف (MID) افضل من طريقة التعويض المتعدد اذ حققت أكبر تكرار ويساوي (7) مرة لحجوم العينات ولنسب الفقد الثلاثة المعتمدة.

2- عند تباين الخطا العشوائي (1.5) وجد ان هناك تقارب بين طريقة التعويض المتعدد وطريقة التعويض المتعدد ثم الحذف اذ حققت الاولى تكرار في الافضلية بلغ (4) مرة أما الثانية فحققت تكرار في الافضلية يساوي (5) مرات .

3- بصورة عامة يمكن استنتاج ان افضل طريقة تقدير للقيم المفقودة لمتغيرات أنموذج الانحدار المتعدد هي طريقة التعويض المتعدد ثم الحذف .

4- أظهرت الزيادة في حجم العينة تاثير واضح على قيمة معيار المفاضلة MMSE اذ كلما زاد حجم العينة قلت قيمته بالنسبة لطريقتي التقدير موضوع البحث.

بناء على ماتم التوصل اليه من استنتاجات نقترح الاتي

1- في حالة التوزيع الطبيعي متعدد المتغيرات يمكن أعتماد على طريقة التعويض المتعدد ثم الحذف في تقدير معلمات أنموذج الانحدار الخطي في ظل وجود مشكلة البيانات المفقودة في كل من متغيري الاستجابة و المتغيرات التوضيحية عند آلية الفقد MAR ولنمط الفقد العمومي.

2- عندما تكون حجم العينة صغير نسبياً وكمية الفقد كبيرة يفضل استخدام طريقة التعويض المتعدد لتقدير معلمات أنموذج الانحدار الخطي في حالة التوزيع الطبيعي متعدد المتغيرات بوجود مشكلة فقد البيانات في كل من متغيري الاستجابة و المتغيرات التوضيحية .

المصادر العربية

1. القزاز، قتيبة نبيل نايف ، 2007 ،"مقارنة اساليب بيز الحصين مع طرائق اخرى لتقدير معالم انموذج الانحدار الخطي المتعدد في حالة البيانات غير التامة" اطروحة دكتوراه فلسفة في الإحصاء ، كلية الإدارة والأقتصاد ، جامعة بغداد.
2. النعيمي، اسوان محمد طيب، 2009 ، "معالجة البيانات غير التامة وتقديرها بطريقة انحدار المركبات الرئيسية"، المؤتمر العلمي الثاني للرياضيات – الاحصاء والمعلوماتية / كلية علوم الحاسبات والرياضيات - جامعة الموصل.
3. حسين، انعام عبود ، 2010، "تحليل البيانات غير التامة لنماذج الانحدار المتعدد بأستخدام الخوارزميات EM، ECM و ECME مع تطبيق عملي" رسالة ماجستير في الإحصاء , كلية الإدارة والأقتصاد , جامعة بغداد.
4. حسين، علي ناصر،2012 ، " تقدير القيم المفقودة لمتغيير الاستجابة في نموذج الانحدار الخطي المتعدد" العلوم الاقتصادية العدد (30) المجلد الثامن.

المصادر الاجنبية

5. Allison ,P.D., 2002, (Missing Data), ASAGE University PAPER , Sage publications, INS.
6. Allison ,P.D, 2012, " Handling Missing Data by Maximum Likelihood" , SAS Global Forum, Orlando, FL (2012), PA, USA., Paper 312-2012 .

7. Allison, Paul D. (2000) "Multiple imputation for missing data: A cautionary tale." Sociological Methods and Research ,vol.28,NO.3,pp. 301-309
8. C. ACOCK ,A., 2005, (Working With Missing Values) , Journal of Marriage and Family 67 (November),pp. 1012–1028 .
9. Dong ,Y.,& Peng,Ch.Y., 2013, "Principled missing data methods for researchers",. May 14;2(1):222. doi: 10.1186/2193-1801-2-222. Print Dec.
10. Enders ,C.K.,(2010)," APPLIED Missing Data", *Series Editor's Note by Todd D.* Little , A Division of Guilford Publications, Inc. ,72 Spring Street, New York, NY 10012.
11. ENDERS,C. K. , 2001The Performance of the Full Information Maximum Likelihood Estimator in Multiple Regression Models with Missing Data", Educational and Psychological Measurement, Vol. 61 No. 5, October ,713-740.
12. Fomby,TH.B., & Hill ,R.C., 2003 ,"MAXIMUM LIKELIHOOD ESTIMATION OF MISSPECIFIED MODELS: TWENTY YEARS LATER", USA, Publisher: Emerald Group Publishing Limited.
13. Hallgren, K. A. ,& Witkiewitz ,K., 2013, "Missing Data in Alcohol Clinical Trials: A Comparison of Methods", Published in final edited form as: Alcohol Clin Exp Res. ; 37(12): 2152–2160. doi:10.1111/acer.12205
14. Higgins,J.P. & Green,S. ,(2008), "Cochrane Handbook for Systematic Reviews of Interventions ", Published by A John Wiley & Sons, Ltd., Publication. Josse,J. & Husson ,F., 2016, "missMDA: A Package for Handling Missing Values in Multivariate Data Analysis" , Journal of Statistical Software - doi: 10.18637/jss.v070.i01.
15. Kennickell , A.B., (1998) ," Multiple Imputation in the Survey of Consumer Finances", Statistical Journal of the IAOS, vol. 33, no. 1, pp. 143-151, 2017.
16. Newgard,C.D., Haukoos, J.S., Lewis,R.J., 2006, "Missing Data: What Are You Missing?" , Society for Academic Emergency Medicine Annual Meeting San Francisco.
17. Peng ,C.Y., Harwell, M., Liou,Sh. M. ,& Ehman,L.H.,2006, "Advances in Missing Data Methods and Implications for Educational Research " Review of education research Ins.s sawilo wsky (Ed.),Real Data analysis (p.p.,31-78).
18. Sullivan ,TH. R., Salter, A. B., Ryan,P., & K. J. Lee , 2014, " Bias and Precision of the "Multiple Imputation, Then Deletion" Method for Dealing With Missing Outcome Data", American *Journal of Epidemiology*, Volume 182, Issue 6, 15 September 2015, Pages 528–534, https://doi.org/10.1093/aje/kwv100.
19. Toutenburg , H ., Heumann, C., Nittner, T., Scheid, S., 2002 ," Parametric and Nonparametric Regression with Missing X's - A Review",journal of the Irnian statistical society, URL: http://jirss.irstat.ir/article-1-87-en.html.
20. Von Hippel,P.T.," 2007", "REGRESSION WITH MISSING Y'S: AN IMPROVED STRATEGY FOR ANALYZING MULTIPLY IMPUTED DATA" , sociological Methodology journal,vol.37,No.1.

……
……..
……..