مقارنة بين طريقتي الحذف وطريقة الأمكان الأعظم للمعلومات الكاملة في معالجة البيانات غير التامة لمتغيرات أنموذج الانحدار المتعدد

أ.م.د. أحمد شاكر المتولي*
منال جبار سلمان**
ahmutwali@uomustansiriyah.edu.iq

المستخلص:

تعد مشكلة البيانات غير التامة لمتغيرات أنموذج الانحدار المتعدد واحدة من اهم المشاكل التي تواجه الباحث عند عملية جمع البيانات لمتغيرات البحث ويعود ذلك إلى عدة أسباب منها اسباب مسيطر عليه مثل الكلفة و المخاطرة و بسبب عدم توفير الإمكانيات للمعاينة واخرى تكون غير مسيطر عليها مثل حالات التلف عند آجراء التجارب او الفقدان بسبب الحروب او الكوارث ، أن أعتماد مثل هذه البيانات غير التامة في عملية التحليل تؤدي إلى نتائج غير دقيقة وعليه يتوجب معالجة فقدان البيانات موضوع البحث باستعمال بعض طرائق المعالجة التي تؤدي الى الحصول على نتائج دقيقة أو تقترب من الدقة. تناول البحث طريقتين من طرائق معالجة البيانات غير التامة لمتغيرات أنموذج الانحدار هما ، طريقة الحذف (list wise deletion)، وطريقة الأمكان الأعظم للمعلومات الكاملة (FULL information maximum likelihood) (FIML) ، يهدف إلى آجراء مقارنة بين هاتين الطريقتين للتوصل إلى أفضل طريقة بالاعتماد على تجارب المحاكاة أفرزت نتائج تجارب المحاكاة، إلى أن طريقة الحذف أفضل من طريقة الأمكان الأعظم للمعلومات الكاملة في معالجة البيانات غير التامة وذلك بالاعتماد على متوسط مربعات الاخطاء لانموذج الانحدار المقدر، كمعيار للمفاضلة.

Comparison between the list wise deletion method and the FULL information maximum likelihood in treating missing values for the variables of multiple regression model

Abstract

The problem of incomplete data for the variables of the multiple regression model is one of the most important problems faced by the researcher in the process of collecting data for the variables of research This is due to several reasons, including controlled reasons such as cost and risk and because of the lack of the possibilities for inspection and other are uncontrolled, such as cases of damage when the tests or loss due to wars or disasters , That the adoption of such incomplete data in the analysis leads to inaccurate results and therefore the loss of data should be addressed by using some methods of treatment that lead to accurate results or near precision. The study dealt with two methods of incomplete data processing for regression model variables .list wise

* الجامعة المستنصرية / كلية الادارة والاقتصاد .
** باحثة .

مقبول للنشر بتأريخ 2018/1/14
مستل من رسالة ماجستير

deletion methoed (LD) and FULL information maximum likelihood methoed (FIML).

, The aim of study is to compare these two methods in order to arrive at the best method based on simulations. The results of the simulation experiments revealed that the method of deletion is better than the method of maximizing the full information in the processing of incomplete data, based on the mean error squares of the estimated regression model.

1- المقدمة وهدف البحث ومشكلة البحث

1-1 المقدمة :

تعد مشكلة البيانات غير التامة لمتغيرات انموذج الانحدار المتعدد واحدة من اهم المشاكل التي تواجه الباحث والتي تؤدي إلى نتائج غير دقيقة في التحليل الاحصائي لذا فقد وجب ايجاد حلول مناسبة لهذه المشكلة من خلال استعمال معالجات أحصائية للبيانات غير التامة والتي يتم الاعتماد عليها لتقدير معلمات أنموذج الانحدار ويعود الفقد للعديد من الأسباب منها ما هو غير متعمد والمتمثلة بحالات التلف عند آجراء التجارب او الفقدان بسبب الحروب أو الكوارث أو المخاطرة عند جمع البيانات . ومنها ما هو متعمد والتي تشمل على حالات الفقدان بسبب التكاليف الباهظة أو الاهمال أو سرية البيانات.. 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 [3]، لتحقيق الهدف المنشود من البحث تم تقسيمه إلى أربعة جوانب ، الاول تضمن على المقدمة ومشكلة وهدف البحث ، أما الجانب الثاني فقد شمل على الجانب النظري، والثالث تمثل بالجانب التجريبي ومناقشة نتائج تجارب المحاكاة المستعملة في التوصل إلى المفاضلة بين الطريقتين موضوع البحث أما الجانب الرابع فشمل على الاستنتاجات والتوصيات.

2-1 مشكلة البحث Research problem

تتلخص مشكلة البحث في عدم امكانية الاعتماد على البيانات غير التامة لانها تؤدي إلى نتائج غير دقيقة اثناء عملية التحليل الاحصائي فيما يتعلق في تقدير معلمات أنموذج الأنحدار والاختبارات الاحصائية المتعلقة بمعلمات الانموذج ، الامر الذي يدعو الى معالجة تلك البيانات غير التامة ومن ثم أعتمادها في عملية التقدير.

3-1 هدف البحث Objective of The Study

يهدف البحث الى أجراء مقارنة بين طريقتي ، تقدير القيم المفقودة (طريقة الحذف List wise deletion (LD) وطريقة الأمكان الأعظم للمعلومات الكاملة FULL information maximum likelihood (FIML)) المستعملة في معالجة القيم المفقودة لمتغير الاستجابة و المتغيرات التوضيحية .

2- الجانب النظري:

في هذا الجانب نتطرق إلى اهم المفاهيم النظرية المتعلقة بالبيانات المفقودة وطريقتي معالجتها (الحذف والامكان الاعظم للمعلومات الكاملة) المستخدمتان في معالجة البيانات غير التامة لمتغيرات انموذج الانحدار.

1-2 أنماط واليات البيانات المفقودة

قبل البدأ بتحليل البيانات يعمد الباحث الى تحديد أنماط واليات فقد البيانات والتي تساعد في تحديد المعالجة المناسب للبيانات ويمكن تلخيص آنماط البيانات المفقودة بالنسبة لمتغيري الاستجابة والتوضيحي بالاتي,[1],[14], [2],[4]:

1. **نمط البيانات المفقودة المنفرد (Univariate missing data)**
2. **نمط المتغيرين (Multivariate two pattern)**
3. **نمط الرتيب أو المتداخل (Monotone pattern)**
4. **نمط العمومي (General pattern)**
5. **النمط المتماثل (File matching pattern)**

أما اليات فقد البيانات فقد صنفت من قبل الباحثين Little and Rubin في عام (1987) [19] ولتوضيح آليات الفقد نفترض ان المصفوفة D تمثل مصفوفة البيانات والتي تتضمن قيم كل من المتغيرات،الاستجابة Y والتوضيحيه X أي ان :

$$D = \{ X , Y\} \quad ...(1)$$

وعلى افتراض ان بعض من عناصر مصفوفة البيانات D تكون مفقودة . ولنفرض أيضا ان المصفوفة M تمثل مصفوفة المتغيرات المؤشرة والتي تضم المتغيرات الموشرة الدالة على فقدان كل قيمه في متغيرات مصفوفة البيانات D أي ان

$$M_{ij} = \begin{cases} 0 & if\ the\ ith\ observation\ of\ jth\ varible\ is\ not\ missing \\ 1 & if\ the\ ith\ observation\ of\ jth\ varible\ is\ missing \end{cases} \quad \dots(2)$$

مما تقدم فان مصفوفة M تكون كالاتي:

$$M= \{M_{obs} , M_{mis} \} \quad ...(3)$$

اذ ان M_{obs} تمثل جزء القيم المشاهدة للمتغيرات و M_{mis} تمثل جزء القيم المفقودة للمتغيرات ،[9].

1. آلية الفقد العشوائي : Missing at Random (MAR)

تفترض الية الفقد العشوائي ان احتمال القيم المفقودة لأحد المتغيرات يعتمد على بعض أو كل القيم المشاهدة لبقية المتغيرات ولكنها لا تعتمد على القيم غير المشاهدة لذلك المتغير، أي ان ، ,[15]، [7]،[10].

$$Pr (M|D)= Pr (M|D_{obs}) \quad ...(4)$$

2. آلية الفقد بشكل عشوائي تام: MissCompletely at Random (MCAR)

تفترض الية الفقد العشوائي التام ان احتمال فقدان قيم احد المتغيرات لا يعتمد على قيم ذلك المتغير أو على قيم بقية المتغيرات قيد البحث ، أي ان مصفوفة المتغيرات المؤشرة تكون مستقلة أحصائياً عن مصفوفة البيانات المشاهدة أي أن، [7]، [10].

$$Pr (M|D) = Pr (M) \quad ....(5)$$

كما وتفترض هذه الآلية إمكانية اعتماد فقدان قيم احد المتغيرات على فقدان قيم احد المتغيرات الاخرى قيد البحث.

3. آلية الفقد بشكل غير العشوائي: Missing not at Random (MNAR)

تفترض هذه الالية ان احتمال فقدان قيم المتغير يرتبط بالقيم غير المشاهدة لذلك المتغير ذلك يعني ان، [7]، [10].

$$Pr (M|D)= Pr (M|D_{obs},D_{mis}) \quad ....(6)$$

MAR أقل تقييدا من MCAR وبالتالي، فان MCAR تكون حالة خاصة من MAR، [16] ،[15].

3- طرائق معالجة القيم المفقودة dealing with missing values

يتناول طريقتان من طرائق معالجة البيانات غير التامة وهي ، طريقة الحذف وطريقة الامكان الاعظم للمعلومات الكاملة .

3-1طريقة الحذف list-wise deletion (LD)

تعد طريقة الحذف احدى الطرائق الإحصائية البسيطة والشائعة الاستخدام في معالجة البيانات المفقودة تقوم فكرة هذه الطريقة على حذف القيم المفقودة في المتغيرات الداخلة في التحليل مما يؤدي ذلك الى حذف القيم المناظرة لها (المشاهدة) في المتغيرات الاخرى ، بمعنى أخر اعتماد قيم المتغيرات المشاهدة فعلاً فقط في عملية التحليل لذا تسمى هذه الطريقة أيضاً بطريقة تحليل الحالة الكاملة (COMPLETE-CASE ANALYSIS) ،الامر الذي يؤدي إلى تقليص حجم العينة، [12] ، [13]. تمتاز هذه الطريقة فضلاً عن بساطتها أنها تختصر الوقت والجهد الخاص بعملية التحليل وتوفر الحزم الاحصائية التي يمكن اعتمادها كالحزمة الإحصائية SAS وSPSS والتي تقوم باجراء عملية التحليل بصورة تلقائية بموجب هذه الطريقة ، بالرغم من انه يمكن اجراء هذه الطريقة من دون اللجوء الى الحزم الاحصائية كاستخدام نظام Excel في تحليل البيانات كما في المعادلة الاتية:

$$Y_{obs}= [\underline{1} \quad X_{obs}] \underline{\beta} + \epsilon_{obs} \quad ....(7)$$

أما ما يعاب على هذه الطريقة هو استبعاد نسبة كبيرة من مشاهدات العينة الأصلية بسبب فقدان بعض القيم من المتغيرات الامر الذي يؤدي إلى فقدان بعض القيم المناظرة للقيم المفقودة والتي تعود لمتغيرات ذات اهمية في التحليل ، كما ان هذا الحذف يؤثر على عدد المشاهدات الفعالة في تحليل (يؤثر على درجة الحرية) مما يؤثر على القدرة الإحصائية (statistical power) والذي يؤدي الى استدلال احصائي ذو كفأءة أقل وبالاخص في حالة الية الفقد العشوائي التام (MCAR) اذ يؤدي الى أخطاء معيارية كبيرة وحدود ثقة

عريضة وخسارة في قوة اختبار الفرضية ومن جانب اخر في حالة العينات الكبيرة فان الاخطاء المعيارية المقدره بموجب هذه الطريقة تكون قريبة من الاخطاء المعيارية التقريبية لذا تعد طريقة جيدة لمعالجة البيانات المفقودة في هذه الحالة ولايمكن استعمالها عندما يكون حجم العينة صغير وتعاني من فقدان في قيم المتغيرات [18]،[19]، [17]، [8]. بموجب هذه الطريقة ولتقدير أنموذج الانحدار فانه يتم الاعتماد على قيم المتغيرات المشاهدة فعلاً إذ ان :

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

حيث 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 انموج الانحدار بالشكل الاتي :

$$\begin{bmatrix} y_{obs} \\ y_{miss} \end{bmatrix} = \begin{bmatrix} \underline{1} & x_{obs} \\ \underline{1} & x_{miss} \end{bmatrix} \underline{\beta} + \begin{bmatrix} \varepsilon_{obs} \\ \varepsilon_{miss} \end{bmatrix} \qquad \text{.....(8)}$$

اذ أن y_{obs} متجه قيم متغير الاستجابة المشاهدة فعلاً ذو السعة (mx1)، y_{miss} متجه قيم متغير الاستجابة المفقودة ذو السعة (n-m X 1)، اذ أن X_{obs} مصفوفة قيم المتغيرات المفسرة المشاهدة فعلاً ذو السعة (mxp)، X_{miss} مصفوفة قيم المتغير المفسرة المفقودة ذو السعة (n-m X p)، $\underline{1}$متجه احادي ذو سعته تستمد من مصفوفة المعلومات ،ϵ_{obs} متجه قيم الاخطاء العشوائية ذو السعة (mx1) و ϵ_{mis} متجه قيم الاخطاء العشوائية ذو السعة (n-m X 1) ومنه نستطيع كتابة أنموذج الانحدار الجزئي الذي يضم المتغيرات ذات القيم المشاهدة فعلاً كما في معادلة رقم (7) .

أما الجزء الثاني من معادلة الانحدار الخطي فيتضمن البيانات المفقودة وكذلك فأن أنموذج الانحدار الجزئي الذي يضم المتغيرات ذات القيم المفقودة يكون كالأتي ، [21]،[16]،[22]،[[14].

$$\underline{Y}_{miss}=\begin{bmatrix} 1 & X_{miss} \end{bmatrix}\underline{\beta}+\varepsilon_{miss} \qquad \text{.....(9)}$$

أي اعتماد نموج الانحدار الجزئي والمبين بالصيغة (7)، وباعتماد طريقة المربعات الصغرى (least square) فان تقدير معلمات أنموذج الانحدار تكون كالأتي:

$$\beta_{obs}=(X_{obs}{}^{\backslash}X_{obs})^{-1}X^{\backslash}_{obs}y_{obs} \qquad \text{.....(10)}$$

اما مصفوفة التباين والتباين المشترك لمتجه المعلمات المقدرة ($\hat{\beta}_{obs}$) فتكون وفق الصيغة الآتية

$$V\text{- cov}\ (\beta_{obs})=\sigma^2\ (X^{\backslash}_{obs}\ X_{obs})^{-1} \qquad \text{.....(11)}$$

أما فيما يخص بمقدار التحيز في تقديرات معلمات الأنموذج فذلك يعتمد على نوع إلية الفقد ، ففي حالة الية الفقد العشوائي التام (MCAR) فان المعلمات المقدرة تكون غير متحيزة تقريباً إذ تحت هذه الإلية فان العينة الجزئية ذات البيانات التامة يمكن اعتبارها كعينة عشوائية بسيطة مسحوبة من العينة البحث الأصلية وكما هو معروف بان العينة العشوائية البسيطة تعطي تقديرات غير متحيزة للمعلمات. أما إذا كانت البيانات وفق الية الفقد العشوائية(MAR) فان طريقة الحذف تؤدي إلى تحيز في عملية تقدير المعلمات،[18]،[13]، ،[5]،[6] .

2-3: طريقة الأمكان الأعظم للمعلومات الكاملة FULL information maximum likelihood (FIML)

يعد الباحثين (Hartley و Hocking) أول من استخدام طريقة الأمكان الأعظم وذلك في عام 1971 ، [19] وتسمى هذه الطريقة ايظاً بطريقة الامكان الاعظم المباشر (direct maximum likelihood) او طريقة الأمكان الأعظم الصفية (raw maximum likelihood) وفي بعض الاحيان يطلق عليه (ML) وذلك لانها تعد تطوير للامكان الاعظم بموجب هذه الطريقة يتم تقدير معلمات المجتمع بشكل مباشر بالاعتماد على كل المعلومات المتوفرة في مجموعة البيانات غير التامة اذ تتضمن هذه الطريقة خطوتين الأولى يتم فيها حساب دالة الامكان لمتجه المتغيرات Z ذو البعد $(p \times 1)$ ولكل حالة (i=1,2,.....n) مع

استبعاد المتغيرات ذات القيم المفقودة في كل حالة فعلى افتراض التوزيع الطبيعي المتعدد لمتجه المتغيرات Z فان لوغاريتم دالة الامكان للحالة i تكون كالاتي :

$$\text{Log } L_i = k_i - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(Z_i - \mu_i)' \Sigma_i^{-1}(Z_i - \mu_i) \qquad (12)$$

إذ ان μ_i هو متجه المتوسطات و Σ_i هو مصفوفة التباين والتباين المشترك لمتجه المتغيرات Z وZ_i تمثل متجه المتغيرات ذات القيم المشاهدة فعلاً للحالة i، K_i هو ثابت يحدد بالاعتماد على عدد p_i من المتغيرات ذات القيم المشاهدة فعلاً للحالة i اذ ان

($k_i = -\frac{p_i}{2}\log 2\Pi$)، [5],[9]. بعد الانتهاء من حساب لوغاريتم دوال الامكان لكل الحالات تأتي الخطوة الثانية والتي تتضمن تجميع لوغاريتم دوال الأمكان لكل الحالات والمبينة بالصيغة (12) لنحصل على دالة لوغاريتم دالة الامكان للعينة وكما مبين أدناه

$$\text{Log } L = k - \frac{1}{2}\sum_{i=1}^{n}\log|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{n}(Z_i - \mu_i)' \Sigma_i^{-1}(Z_i - \mu_i) \quad \ldots (13)$$

اذ ان $k=\sum_{i=1}^{n} k_i$ وبتعظيم لوغاريتم دالة الامكان للعينة نحصل على تقدير متجه معلمات المجتمع.

من الجدير بالملاحظة وعلى خلاف طريقة التعويض المتعدد للقيم المفقودة (MI) فان طريقة FIML لا تعوض أي قيمة مفقودة وانما تعتمد على المعلومات المتوفرة في البيانات الغير تامة والتي تعد ميزة تمتاز بها هذه الطريقة [16]،[11].

تمتاز هذه الطريقة أن الخوارزمية تستخدم جميع المعلومات فيها ولعدد غير محدود من أنماط البيانات المفقودة حيث انها تفترض متعددة المتغيرات الطبيعي ، لذا فهي تزيد من احتمالية الأنموذج، ويتم فيها تقدير متوسط وتباين الجزئي لقيم المفقودة والمشاهدة من المتغيرات الأخرى ولا توجد حدود لكمية البيانات المفقودة ، وكذلك تمتاز ايضاً بانها تعطي تقديرات غير متحيزة في حالة النمط العشوائي (MAR) والتام العشوائية (MCAR) ولكنها تعطي تقديرات متحيزة في حالة النمط غير العشوائي (MNAR) وتعطي تقديرات متسقة وفعالة عندما يكون الفقد تام العشوائية .

ويعاب عليها عدم قدرتها على التعامل مع بعض النماذج (مثل نماذج الاتجاهية duration models ، ونماذج سجل الأحداث event history models ،إلخ.) ،وكذلك عدم توفرها برامج أحصائية جاهزة مثل spss يمكن أستخدامها لتطبق هذه الطريقة في ما عدا توفرها في حزم برمجية تعتمد على المعادلات مثل LISREL و , MPLUS [9],[20].

4- الجانب التجريبي

في هذا الجانب سوف نستخدم اسلوب المحاكاة وذلك لدراسة حالات متنوعة من البيانات من ناحية توزيعها والتي من الممكن ان تجسد المشاكل التي تواجه الباحث عند تحليل الأنموذج الخطي المتعدد وبوجود المشكلة الرئيسة وهي وجود فقد في بعض من بيانات متغير الاستجابة Y والمتغيرات التوضيحية X^s، ،وباستخدام طريقة مونت كارلو لتوليد البيانات وفق ما يأتي:

1. تم توليد بيانات المتغيرات المستقلة (X's) لثلاثة متغيرات توضيحية) وفق التوزيع ثنائي الحدين وبحدين (a,b) و الخطأ العشوائي يتوزع توزيعاً طبيعياً بمتوسط مساوي للصفر وتباين مساوٍ لـ σ^2 أي أن $e \sim N(0\ ,\ \sigma^2)$
2. بعد توليد قيم المتغيرات التوضيحي يتم حساب قيم لمتغير الاستجابة من العلاقة الخطية التالية:

$$Y = X\beta + e \qquad \ldots(14)$$

أذ ان

Y تمثل متجه قيم المتغير الاستجابة من الدرجة nx1 وX تمثل مصفوفة المتغيرات التوضيحية من الدرجة nx3 و β تمثل متجه معاملات الانحدار للنموذج من الدرجة 3x1 اما e فتمثل متجه الخطأ العشوائي من الدرجة nx1.

تم فرض قيم أولية لمعلمات الأنموذج (β) وفق الانموذج المعتمد دراسته وبما يتوافق مع طبيعة الظاهرة المدروسة وهي (1،0.2،0.5،0.3) لمعلمات الانموذج (β0،β1، β2، β3) على التتابع بهذا فان أنموذج الانحدار الخطي المستعمل في عملية المحاكاة يكون كالاتي.

$$y_i = 1 + 0.2\,x_{i1} + 0.5\,x_{i2} + 0.3\,x_{i3} + e_i \qquad \ldots(15)\ ،i = 1,2,\ldots,n$$

3. توليد البيانات لكل من متغير الاستجابة والمتغيرات التوضيحية وبقيم مفقودة وفق الية الفقد العشوائي MAR وبنسب فقد (5%،10%، 15%) ونمط الفقد العمومي و لثلاث حجوم للعينات هي (110,75,40) وكذلك بقيمتين

لتباين الخطأ العشوائي هي (1,0.5) . وتكرار التجربة 1000 مرة بغية التوصل إلى نتائج ذات دقة عالية يتم الأعتماد عليها للمقارنة بين طرائق التقدير المعتمدة[2] .

4. تم الاعتماد في المقارنة بين طريقتي لمعالجة البيانات غير التامة على معيار للمفاضلة متوسط مربعات الاخطاء لانموذج الانحدار المقدر للمحاولة الواحدة ولنفس النموذج وفق الصيغة الاتية:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p} \quad \ldots(16)$$

ومن ثم حساب متوسط قيم متوسطات مربعات الاخطاء لجميع المحاولات لاعتماده لغرض المقارنة بين طريقتي المعالجة المعتمدة ،[3] ، [4].

1-4 تحليل نتائج تجارب المحاكاة:

أولاً : المقارنة بين طريقتي المعالجة بافتراض تباين الخطأ العشوائي يساوي (0.5).

أفرزت قيم معيار المفاضلة MMSE والمبينة في الجدول (1) ان طريقة التقدير LD هي الافضل بالنسبة لحجم العينة n=40 عند نسبة الفقد 5% و لحجم العينة n=75 لنسبتي الفقد 10% و 15% وكذلك لحجمي العينتين (n=40, n=110) لنسبة الفقد 15% اذ كان لها أقل القيم لمعيار المفاضلة MMSE فقد تراوحت قيمة معيار المفاضلة المفاضلة MMSE بين (6.323536487710e -08) و (1.3920428e-07) أما طريقة FIML هي افضل عند حجم العينة (n=40 وn=110) ونسبة فقد (10%) وكذلك لنسبة فقد (5%) لحجمي العينة (n=75 وn=110) اذ حققت أقل القيم لمعيار المفاضلة MMSE المعتمد والتي تتراوح بين (4.80794380e-08) (7.4147315e-08) هي الاقل وبناء على نتائج المفاضلة نجد ان هناك تقارب بين طريقتي المعالجة المعتمدة .

جدول (1)

تقدير معلمات أنموذج الانحدار ومتوسط متوسطات مربعات الاخطاء لتلك التقديرات ومتوسط متوسطات مربعات الاخطاء للأنموذج المقدر مصنفة حسب طريقتي التقدير ونسب التقدير وحجوم العينات عندما (e ~ N (0,0.5))

حجم العينة	طرائق التقدير	نسبةالفقدان	$MMSE(\beta)$ومقدرات المعالم				MMSE
			β_0	β_1	β_2	β_3	
40	LD	5%	0.7315 (0.1270)	0.2007 (0.055)	0.5020 (0.055)	0.3036 (0.055)	9.3794494e-08
		10%	1.3661 (0.4169)	0.1988 (0.2829)	0.4946 (0.2829)	0.3000 (0.2829)	1.35157721e-07
		15%	0.9769 (0.1203)	0.1995 (0.1198)	0.5051 (0.1198)	0.2955 (0.1198)	1.3920428e-07
	FIML	5%	0.7525 (0.1205)	0.2001 (0.0593)	0.5015 (0.0593)	0.3029 (0.0593)	2.41122706e-07
		10%	1.3278 (0.3708)	0.1982 (0.2634)	0.4965 (0.2634)	0.3003 (0.2634)	7.4147315e-08
		15%	1.0857 (0.1652)	0.1999 (0.1579)	0.5051 (0.1579)	0.2942 (0.1579)	2.2884671e-07
75	LD	5%	0.8159 (0.1066)	0.2037 (0.0727)	0.4961 (0.0727)	0.3035 (0.0727)	6.576109022e-08
		10%	0.7437 (0.1236)	0.1985 (0.0579)	0.5037 (0.0579)	0.3017 (0.0579)	6.97054454509351e-08
		15%	1.0516 (0.1469)	0.2001 (0.1443)	0.4986 (0.1443)	0.301 (0.1443)	6.558316877e-08
	FIML	5%	0.7988 (0.1092)	0.2037 (0.0687)	0.4958 (0.0687)	0.3031 (0.0687)	4.80794380e-08
		10%	0.7666 (0.1172)	0.1990 (0.0627)	0.5030 (0.0627)	0.3008 (0.0627)	1.35435370833619e-07
		15%	1.0085 (0.1293)	0.2008 (0.1292)	0.4990 (0.1292)	0.3009 (0.1292)	1.249418504e-07
110	LD	5%	0.9910 (0.1234)	0.2012 (0.1234)	0.4979 (0.1234)	0.3006 (0.1234)	5.99056026982418e-08
		10%	1.2892 (0.3285)	0.4994 (0.2449)	0.2952 (0.2449)	0.1999 (0.2449)	6.23539845000 3e-08
		15%	0.8052 (0.1090)	0.2016 (0.0710)	0.5021 (0.0711)	0.2998 (0.0710)	6.32353648776 16e-08
	FIML	5%	0.9233 (0.1081)	0.2015 (0.1022)	0.4985 (0.1022)	0.3006 (0.1022)	4.96031695771036e-08
		10%	1.3757 (0.4291)	0.2006 (0.2880)	0.4991 (0.2880)	0.2955 (0.2880)	5.04192300902 8e-08
		15%	0.8464 (0.1048)	0.2017 (0.0812)	0.5011 (0.0812)	0.2990 (0.0812)	1.60057162800 85e-07

ثانياً: المقارنة بين طريقتي المعالجة بافتراض تباين الخطا العشوائي يساوي(1).بمراجعة قيم معيار المفاضلة MMSE لطريقتي التقدير موضوع البحث والمعروضة في الجدول (2) نجد ان طريقة LD هي الافضل بالنسبة لحجمي العينة (n=40 و n=110) ولجميع نسب الفقدان وكذلك هي الافضل بالنسبة لنسبتي الفقد 10% و 15%عند حجم العينة n=75 و اذ حققت أقل قيمة لمعيار المفاضلة MMSE والذي تروح بين (5.812447e-08) و (1.4009469165e-07) أما معيار المفاضلة بالنسبة لطريقة التقدير FIML فقد بين ان هذه الطريقة هي الافضل فقط عند نسبة فقد 5% لحجم عينة n=75 اذ بلغت قيمة هذا المعيار المفاضلة (9.199648e-09) وهي الأقل مما تقدم نستنتج ان طريقة LD قد تفوقت على طريقة المعالجة (FIML) .

جدول (2)

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

حجم العينة	طرائق التقدير	نسبةالفقدان	مقدرات المعالم $MMSE(\beta)$				MMSE
			β_0	β_1	β_2	β_3	
40	LD	5%	0.4630 (0.3081)	0.2014 (0.0197)	0.5041 (0.0197)	0.3071 (0.0197)	8.85334657282409e-08
		10%	1.7323 (1.0432)	0.1977 (0.5069)	0.4893 (0.5070)	0.2999 (0.5069)	1.35903164928798e-07
		15%	0.9538 (0.1154)	0.1990 (0.1133)	0.5101 (0.1134)	0.2911 (0.1133)	1.40094691652909e-07
	FIML	5%	0.5051 (0.2678)	0.2002 (0.0229)	0.5029 (0.0229)	0.3058 (0.0229)	2.06412012230430e-07
		10%	1.6556 (0.8842)	0.1964 (0.4544)	0.4930 (0.4544)	0.3007 (0.4544)	2.40966105890218e-07
		15%	1.1713 (0.2222)	0.1998 (0.1928)	0.5102 (0.1929)	0.2884 (0.1930)	1.77805261027974e-07
75	LD	5%	0.6318 (0.1714)	0.2073 (0.0359)	0.4922 (0.0359)	0.3070 (0.0359)	6.67240802400982e-08
		10%	0.4874 (0.2850)	0.1971 (0.0223)	0.5073 (0.0224)	0.3033 (0.0223)	6.96691924538394e-08
		15%	1.1031 (0.1738)	0.2001 (0.1632)	0.4972 (0.1632)	0.3036 (0.1632)	6.65202552726796e-08
	FIML	5%	0.5974 (0.1931)	0.2073 (0.0312)	0.4917 (0.0312)	0.3063 (0.0311)	9.19964862091559e-09
		10%	0.5332 (0.2441)	0.1979 (0.0262)	0.5060 (0.0262)	0.3015 (0.0262)	9.21542553429016e-08
		15%	1.0169 (0.1321)	0.2015 0.1318)	0.4980 (0.1318)	0.3018 (0.1318)	1.66069755356532e-07
110	LD	5%	0.9819 (0.1204)	0.2025 (0.1201)	0.4957 (0.1201)	0.3012 (0.1201)	5.81244745431633e-08
		10%	1.5783 (0.7399)	0.1998 (0.4055)	0.4988 (0.4055)	0.2904 (0.4056)	6.34555322865894e-08
		15%	0.6105 (0.1863)	0.2032 (0.0346)	0.5042 (0.0346)	0.2997 (0.0346)	6.42670200112867e-08
	FIML	5%	0.8467 (0.1042)	0.2031 (0.0807)	0.4971 (0.0807)	0.301 (0.0807)	1.11222705896967e-07
		10%	1.7515 (1.0852)	0.2012 (0.5205)	0.4982 (0.5205)	0.2910 (0.5206)	1.87840425552656e-07
		15%	0.6928 (0.1420)	0.2034 (0.0476)	0.5022 (0.0476)	0.2980 (0.0476)	1.70487230301274e-07

ثالثاً: المفاضلة بين طريقتي التقدير لجميع حجوم العينات المعتمدة

للمفاضلة بين طريقتي معالجة البيانات غير التامة موضوع البحث تم الاعتماد على عدد مرات تكرار افضلية كل طريقة بالنسبة لحجوم العينات الثلاث ولنسب الفقدان الثلاثة التي تم افتراضها وحسب تباين الخطأ المفترض اذ تم تلخيص تلك التكرارات في الجدول (3) وبملاحظة تلك التكرارات نجد ان هناك تقارب بين طريقة الحذف وطريقة الامكان الاعظم للمعلومات الكاملة (0.5) اذا كان تكرار الاولى (5) وتكرار الثانية (4) . أما بالنسبة لتباين الخطأ المفترض (1) فنجد ان طريقة الحذف تفوقت على طريقة الامكان الاعظم للمعلومات الكاملة وبشكل واضح اذ كان لها اكبر تكرار ويساوي (8) في حين هناك حالة واحدة فقط تفوقت فيها الطريقة الثانية.

جدول (3)
ملخص لعد تكرار أفضلية طريقتي التقدير بالنسبة لجميع حجوم العينات ونسبة الفقدان بالاعتماد على نسبة تباين الخطأ

تباين الخطأ	طريقتي التقدير	
	LD	FIML
0.5	5	4
1	8	1

8- الاستنتاجات والتوصيات The Conclusions and recommendations:

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

1- **وجد ان هناك تقارب بين طريقتي المعالجة بالنسبة لتباين الخطأ العشوائي (0.5) إذ تفوقت طريقة الحذف (5) مرات بينما تفوقت طريقة الامكان الاعظم للمعلومات الكاملة (4) مرات .**
2- **بانسبة لتباين الخطا العشوائي المفترض (1) وجد ان طريقة الحذف LD قد تفوقت بشكل واضح على طريقة الامكان الاعظم للمعلومات الكامل اذ تفوقت (8) مرات في حين تفوقت الطريقة الثانية مرة واحدة فقط**
3- **وبصورة عامة يمكن التوصل إلى ان طريقة الحذف تعد افضل من طريقة الامكان الاعظم للمعلومات الكاملة وفق ما تم التوصل اليه من نتائج تجارب المحاكاة.**

بناء على ماتم التوصل اليه من استنتاجات نقترح الاتي

1- **في حالة التوزيع الطبيعي متعدد المتغيرات يمكن أعتماد على طريقة 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- **عند حجوم العينات الصغيرة وفي حالة نسب الفقد العالية لا يفضل الاعتماد على طريقة الحذف في معالجة البيانات ويمكن الاعتماد على طريقة FIML .**

المصادر العربية

1. **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.**
2. **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.**
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. **Bori ,M.S., 2013, "Dealing with missing data: Key assumptions and methods for applied analysis" , published in fulillment of the requirements for PM931 Directed Study in Health Policy and Management ,Technical Report No. 4.**
7. **C. ACOCK ,A., 2005, (Working With Missing Values) , Journal of Marriage and Family 67 (November),pp. 1012–1028.**
8. **Cool,A.L., (2000) , "A Review of Methods for Dealing with Missing Data" , ERIC Number: ED438311, Annual Meeting of the Southwest Educational Research Association.**
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., Bandalos D.L., 2001, " The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models" , structural equation modeling:A Multidisciplinary Journal , Pages 430-457 .**
12. **Little, R. J. A. (1992) "Regression with missing X'S", journal of the American statistical Association 87, 1227- 1237.**
13. **Myers,T.A.,(2011) , "Goodbye, Listwise Deletion:Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data",Journal Communication Methods and Measurs, Pages 297-310.**

14. 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.
15. Pantanowitz,A., & Marwala,M.,2008 "Evaluating the Impact of Missing Data Imputation through the use of the Random Forest Algorithm" ,School of Electrical and Information Engineering, University of the Witwatersrand, Private Bag x3, Wits, 2050, Republic of South Africa,Available http: //arxiv.org/arxiv/0812/0812.2412pdf.
16. 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). .
17. Raghunathan, T.E.,(2004), "WHAT DO WE DO WITH MISSING DATA ? SOME OPTIONS FOR ANALYSIS OF INCOMPLETE DATA", Journal Annual Reviews, Vol. 25:99-117.
18. Saunders,J.A., Morrow,N., Spitznagel,E., Dori,P., Proctor,E.K., and Pescarino,R. , 2006, "Imputing Missing Data: A Comparison of Methods for Social Work Researchers" National Association of Social Workers, Volume 30, Issue 1, 1 March ,06, Pages 19–31,https://doi.org/10.1093/swr/30.1.19.
19. Smiley,W.F., 2015 , "Exploring Listwise Deletion and Multilevel Multiple Imputation in Linear Two-Level Organizational Models", A*merican Journal of Epidemiology*, Volume 182, Issue 6, 15 September 2015, Pages 528–534, https://doi.org/10.1093/aje/kwv100.
20. TUFIŞ,C.D. ,(2008), "MULTIPLE IMPUTATION AS A SOLUTION TO THE MISSING DATA PROBLEM IN SOCIAL SCIENCES" , Journal: Calitatea vieții, nr. 1–2, P.R. 199–212.
21. 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
22. 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.

……..
……
……