بعض الطرق الإحصائية لمعالجة البيانات المفقودة (دراسة مقارنة مع التطبيق في مجال الأمراض السريرية)

د/حسن حسن عبد الملك

أستاذ العلوم المالية والمصرفية المساعد-بقسم العلوم المصرفية -كلية التجارة-جامعة إب-الجمهورية اليمنية

Email: hasmalik5@hotmail.com

الملخص:

هدفت الدراسة إلى تحليل وتقيم بعض الطرق الإحصائية لمعالجة البيانات المفقودة وذلك من خلال:

- دراسة تحليلية لبعض الطرق الإحصائية لمعالجة البيانات المفقودة التي تخضع للتوزيع الطبيعي.
- توليد البيانات باستخدام التوزيع الطبيعي الاحتمالي (Normal Distribution) كمعالجة إحصائية للبيانات قبل إدخالها لعملية التقدير.
- تحديد أفضل طريقة لمعالجة البيانات المفقودة وذلك من خلال عمل مقارنة لمصفوفة التباين-والتباين المشترك لمعلمات (β) لكل من طريقة تقدير المربعات الصغرى الاعتيادية (Ordinary Least Squares Method) وطريقة تقدير الإمكان الأعظم (Maximum Likelihood Method) وطريقة تقدير (The EM Algorithm Method).

وتوصلت الدراسة إلى مجموعة من الاستنتاجات التي تتفق مع الأهداف التي تم ذكرها آنفًا وخصوصًا في تحديد الطريقة المثلى لمعالجة البيانات المفقودة؛ حيث توصلت الدراسة إلى أن جميع الطرق (الأساليب) الإحصائية المذكورة آنفًا لها كفاءات متقاربة في معالجة وتقدير البيانات المفقودة، وايضًا توصلت الدراسة إلى أن طريقة المربعات الصغرى لها الأفضلية من بين الطرق الإحصائية المذكورة آنفًا في معالجة البيانات المفقودة.

الكلمات المفتاحية: البيانات المفقودة ـ طريقة تقدير المربعات الصغرى الاعتيادية ـ طريقة تقدير الإمكان الأعظم ـ طريقة تقدير (EM) الخوارزمية ـ التوزيع الطبيعي.

Abstract:

The study aimed at analyzing and evaluating some statistical methods for processing missing data through:

- *Analytical study of some statistical methods for processing missing data that are subject to normal distribution.*
- *Generating data using a normal distribution as a statistical treatment of data before entering it for the estimation process.*
- *Determining the best way to process the missing data by making a comparison of the variance matrix - and the covariance of the parameters (β) for each of those methods (Ordinary Least Squares Method, Maximum Likelihood Method, and The EM Algorithm Method).*

The study concluded that all the aforementioned statistical methods have the same efficiency in processing and estimating the missing data, and also the study concluded that the Ordinary Least Squares Method has the preference among the aforementioned statistical methods in processing the missing data.

***Keywords**: Missing Data, Ordinary Least Squares Method, Maximum Likelihood Method, the EM Algorithm Method, and Normal Distribution.*

أولًا-الإطار العام:

1-المقدمة:

يشهد العالم تطورًا متسارعًا في جميع مجالات الحياة ويعد علم الإحصاء من العلوم المهمة لما له من دور مهم وبارز في تحليل واستخراج النتائج لمختلف الدراسات العلمية في شتى المجالات[7]، ومن المعروف أن الدراسات العلمية تصنف تبعًا لطبيعة البيانات المستخدمة فيها[9].

تمر الدراسات العلمية بمجموعة من المراحل منها الحصول على بيانات الدراسات العلمية، وفي هذه المرحلة بالذات تنشأ مشكلة البيانات المفقودة، حيث يعد معرفة أسباب ظهور البيانات المفقودة وأنواعها (أنماطها) من أهم العوامل التي تساعد الباحثين للحصول على طريقة لمعالجة البيانات المفقودة، وخصوصًا إذا كان من بين البيانات المفقودة ما قد يكون ناجمًا عن أسباب فنية أو أسباب يمكن تلافيها كنقص التدريب أو عدم توفر المصادر المالية أو غيرها من الأسباب؛ لذلك ليس من السهل سرد كل ما يمكن عن أسباب ظهور البيانات المفقودة وأنماط البيانات المفقودة، لذلك حاول الباحث تلخيص أهم أسباب ظهور البيانات المفقودة وأهم أنماطها في الآتي:

1- أسباب ظهور البيانات المفقودة [6] [18]: التي تتمثل في الأسباب الآتية:

(أ)الاسقاط المتعمد أو غير المتعمد لبعض بيانات (مفردات) الدراسة على وجه التحديد في أثناء مرحلة جمع البيانات؛ (ب) عدم التمكن من توجيه الأسئلة بدقة ووضوح؛ (ج) حساسية موضوع الدارسة؛ (د) وقت وطول الزيارة لجمع بيانات الدراسة؛ (هـ) التكاليف المادية لجمع بيانات الدراسة؛ (و) في أثناء ترميز البيانات وتجهيزها لعملية تبويب البيانات وتحليلها؛ نتيجة لظهور البيانات المفقودة للأسباب المذكورة أنفًا وجب علينا ذكر أنماط (آلية) البيانات المفقودة.

2- أنماط (آلية) البيانات المفقودة [4] [8] [16] [19]: تتمثل في ثلاث نقاط أو عناصر رئيسة وهي: الفقدان العشوائي التام (MCAR)، الفقدان العشوائي (MAR)، وأخيراً الفقدان غير العشوائي (NMAR). ومما سبق وعلى وجه التحديد أنماط البيانات المفقودة مختلفة الأنواع فعلى سبيل المثال فهناك فقدان بمتغير واحد (X_i) فقط بينما جميع المتغيرات المستقلة الأخرى تكون تامة أي لا يوجد بها بيانات مفقودة، وأيضًا هناك بيانات مفقودة المرتبة والمتداخلة؛ حيث يتم ترتيب المتغيرات المستقلة ترتيبًا تصاعديًا أو تنازليًا، ومن ثم يتم اختيار عينات عشوائية جزئية للمتغيرات التابعة بحيث كل عينة لا تمثل العينة الأخرى، وأخيرًا هناك نمط فقدان للبيانات ناتج عن دمج المتغيرات؛ حيث يظهر لنا متغير جديد للبيانات المفقودة وبشكل مختلف عن سابقتها. من وثم تمثل البيانات المفقودة مشكلة كبيرة وضخمة، وذلك بسبب تشويه طبيعة البيانات الأصلية، ومن ثم قد يلجأ بعض الباحثين إلى إهمال أو حذف هذه البيانات لمعالجة هذا التشوه ونتيجة لهذه المعالجة (الإهمال أو الحذف) يكون الباحثون قد فقدوا معلومات أو بيانات مهمة قد تفيد الدراسة، وهذا الأمر

غير مقبول في الدراسات العلمية، ومن ناحية أخرى عند التحليل الإحصائي في حالة وجود بيانات مفقودة بالبيانات ستكون نتائج هذا التحليل الإحصائي نتائج متحيرة وغير دقيقة[3].

ولمعالجة أوحل مشكلة البيانات المفقودة ظهرت العديد من الكتب المنهجية والدراسات العلمية الإحصائية سواءً كانت عربية أو أجنبية، ففي عام (2003) قام الباحثان (Singh & Deo) [17] باقتراح أربع طرق لمعالجة البيانات المفقودة وهي كالآتي: طريقة متوسط الفقد؛ وطريقة نسبة الفقد؛ طريقة حاصل ضرب الفقد، وأخيرًا الفقد بتحويل القوة. وفي عام (2005) قام الباحثون (Wang & et al) [19] بمعالجة البيانات المفقودة باستخدام إشارات التحليل الطيفي. وفي عام (2010) قامت (National Research Council) [16] بإصدار كتاب عن البيانات المفقودة وتم استخدام النماذج المختلطة وتحليل الحساسية لمعالجة البيانات المفقودة.

وفي عام (2012) قام الباحث (المنجي)[3] بمعالجة البيانات المفقودة استنادًا إلى ما قام به الباحثان (Singh & Deo) في عام (2003). وفي عام (2014) قام الباحثان (الرحيل والدرابسة) [1] بمعالجة البيانات المفقودة باستخدام طريقة (EM) الخوارزمية وطريقة (MI) المقصودة بها التعويض المتعدد وطريقة الإمكان الأعظم، وأخيرًا طريقة بيز. وفي العام ونفسه؛ حيث قام الباحثون (KNOPP & et al) [12] بمعالجة البيانات باستخدام التوزيعات الاحتمالية وعلى التحديد توزيع برنولي. وفي عام (2014) قام الباحثون (Taylor & et al) [11] بمعالجة البيانات شبه المفقودة باستخدام بمزج طريقة الفازي وطريقة بيز.

وفي عام (2016) قام الباحثان (رشيد وحمزة) [8] بمعالجة البيانات المفقودة باستخدام طريقة تعويض المتوسط – الوسيط، طريقة التمهيد اللامعلمي، وطريقة الجار الأقرب. وفي عام (2016) قام الكاتب (Raghunthan) [18] بوضع كتاب عن البيانات المفقودة وتم معالجة هذه البيانات باستخدام الانحدار وتحليل المتعامد والخوارزميات. وفي عام (2017) قام الباحثان (المشهداني وأحمد) [2] بمعالجة البيانات المفقودة باستخدام تحليل التغاير.

وفي عام (2018) قام الباحثون (Zhang & et al) [20] بمعالجة البيانات المفقودة باستخدام طريقة (CV) الخوارزمية. وفي عام (2019) قام الباحثان (Wu & Ma) [10] بتقدير البيانات باستخدام (EM) الخوارزمية وذلك بتوليد البيانات باستخدام العمليات العشوائية للنماذج المختلطة وطريقة مونت كارلو ودراسة كفاءة التقدير لهذه البيانات. وفي عام (2020) قام الباحثون (Zhao & et al) [13] بمعالجة البيانات المفقودة بمزج طريقة كل من نموذج هيكمان وطريقة (EM algorithms).

مما سبق لاحظ الباحث أن الطرق الإحصائية لمعالجة البيانات المفقودة في الكتب المنهجية والدراسات السابقة اختلفت من ناحية المعالجة لهذه البيانات، حيث تضمنت طرقًا لمعالجة البيانات المفقودة على مستوى المتغير الواحد وعلى وجه التحديد على المستوى العمودي للمتغير مثل طريقة

(طريقة متوسط، وطريقة نسبة، طريقة حاصل الضرب، الفقد بتحويل القوة، وطريقة الجار الأقرب، وغير ذلك)، وأيضًا هناك معالجة للبيانات على مستوى المتغيرات بشكلٍ عام وعلى وجه التحديد على مستوى الحالة الواحدة (المستوى الأفقي) مثل (النماذج المختلطة، الانحدار، الانحدار اللامعلمي، (EM) الخوارزمية، والتعويض المتعدد، تحليل التغاير، وبيز وغير ذلك)، وأخيرًا هناك معالجات للبيانات المفقودة وذلك بالمزج بين طريقتين أو نموذجين إحصائيين في آن واحد مثل (استخدام إشارات التحليل الطيفي، التقدير المويجي لدالة الانحدار اللامعلمي وغير ذلك).

ومهما يكن من أمر ذلك نجد أن لكل من هذه الطرق مزايا وعيوبًا من ناحية شروط الأساليب الإحصائية وأيضًا من ناحية البرامج الإحصائية المستخدمة لتطبيق هذه الطرق.

لذا ما يميز هذه الدراسة عن الدارسات السابقة بمعالجة البيانات المفقودة على مرحلتين: المرحلة الأولى وهي إخضاع البيانات وتوليد قيم احتمالية على مستوى المتغير الواحد وهذا الجانب لم يتم التطرق إليه في الدراسات السابقة.

ومن ثم تأتي المرحلة الثانية وهي تقدير البيانات المفقودة باستخدام نماذج إحصائية تخضع لشرط التوزيع الاحتمالي المختار نفسها. وأخيرًا دراسة مقارنة بين نتائج النماذج المقترحة، وأيضًا ما يميز هذه الدراسة عن الدراسات السابقة أنه تم التطبيق على بيانات طبية وعلى وجه التحديد المرض السكري وهذا الجانب لم يتم التطرق إليه في الدراسات السابقة وخصوصًا التطبيق على البيانات المفقودة في هذا المجال، أخيرًا الهدف الرئيس للدراسة هو الكشف عن أفضل طريقة لمعالجة البيانات المفقودة وليس دراسة أي علاقات أو تأثيرات طبية؛ لذا تم التطبيق بهذا المجال لتوفر البيانات المفقودة فقط.

2-مشكلة الدراسة:

في العادة تشترط الأساليب الإحصائية في عملية التقدير عدم وجود بيانات مفقودة، ومن ثم قد يتم التعامل مع البيانات المفقودة بالإهمال والتجاهل، ونتيجة لذلك (الإهمال أو التجاهل) تظهر لنا مجموعة من المشاكل ومنها حجم البيانات المفقودة قياسًا بحجم البيانات الأصلية، وأيضًا شكل أو توزيع (نمط) البيانات المفقودة؛ إذ إن لكل نمط منها تعاملا خاصًا للبيانات المفقودة، ومن ثم عند استخدام الأساليب (النماذج) الإحصائية لتقدير البيانات المفقودة ستكون نتائجها متحيزة وغير دقيقة، ونتيجة لذلك ستكون الأساليب (النماذج) الإحصائية المختارة في عملية التقدير ليست الأداة المناسبة؛ وهو ما يصيب الباحثين بارباك في اختيار الأساليب الإحصائية التي تلائم طبيعة بيانات دراستهم، على وجه التحديد هل هذه البيانات تخضع لتوزيع معلمي أو لامعلمي.

ومن هنا نشأت مشكلة الدراسة الرئيسة والمتمثلة بالتساؤل الآتي" ما هي أفضل طريقة إحصائية لمعالجة البيانات المفقودة؟" ويتفرع من هذا السؤال أسئلة فرعية:

1- ما مدى صلاحية البيانات الطبية في دراسة البيانات المفقودة؟ وما مدى الفقدان العشوائي فيها؟

2- كيف يتم التحليل الإحصائي للبيانات المفقودة وتحديد (آلية) نمطها؟

3- ما هي أفضل طريقة إحصائية لمعالجة أو (تقدير) البيانات المفقودة عند إخضاع البيانات للتوزيع الطبيعي (Normal Distribution) الاحتمالي كمعالجة إحصائية للبيانات قبل إدخالها لعملية التقدير؟

3-أهداف الدراسة:

تهدف الدراسة إلى تحليل وتقييم بعض الطرق الإحصائية لمعالجة البيانات المفقودة وذلك من خلال:

1- دراسة تحليلية لبعض الطرق الإحصائية لمعالجة البيانات المفقودة التي تخضع للتوزيع الطبيعي.

2- توليد البيانات باستخدام التوزيع الطبيعي (Normal Distribution) الاحتمالي كمعالجة إحصائية للبيانات قبل إدخالها لعملية التقدير.

3- تحديد أفضل طريقة لمعالجة البيانات المفقودة، وذلك من خلال عمل مقارنة لمصفوفة التباين-والتباين المشترك لـ معلمات (β) لكل من طريقة تقدير المربعات الصغرى الاعتيادية (Ordinary Least Squares Method)، طريقة تقدير الإمكان الأعظم (Maximum Likelihood Method)، وأخيرًا طريقة تقدير (The EM Algorithm Method).

4-أهمية الدراسة:

تكمن أهمية الدراسة في العديد من الجوانب منها:

أ-الأهمية العلمية:

تتلخص أهمية الدراسة في توضيح أسباب ظهور هذه البيانات، وأيضًا في توضيحها لبعض الأساليب الإحصائية المستخدمة لمعالجة البيانات المفقودة التي تخضع للتوزيع الطبيعي، وأيضًا اقتراح معالجة إحصائية لتوليد البيانات باستخدام التوزيعات الاحتمالية وعلى وجه التحديد التوزيع الطبيعي (Normal Distribution) بدل اللجوء إلى المحاكاة (طريقة البوتستراب أو طريقة مونت كارلو) أو استخدام أساليب إحصائية متقدمة، وأخيرًا تحديد أفضل طريقة لمعالجة البيانات المفقودة.

ب-الأهمية العملية:

ستضيف معرفة علمية تكاد تكون غير متوفرة في المكتبات اليمنية، وأيضًا إعطاء القيادات ومتخذي القرار والمخططين والباحثين في شتى المجالات والتخصصات معلومات عن كيفية التعامل مع البيانات المفقودة، وأيضًا تزويدهم بالمعلومات العملية التي تساعدهم في كيفية معالجة وتقدير البيانات المفقودة باستخدام بعض الأساليب الإحصائية التي تخضع للتوزيع الطبيعي.

5-حدود الدراسة:

تناولت الدراسة تحليلا وتقويمًا لبعض الطرق الإحصائية لمعالجة البيانات المفقودة، وعلى وجه الخصوص في تحديد أفضل الطرق الإحصائية الممكن استخدامها لمعالجة البيانات المفقودة وذلك من خلال عمل مقارنة لكل من طريقة تقدير (OLS)، طريقة تقدير (LME)، طريقة تقدير (EM Algorithm)؛ إذ تم الحصول على بيانات الدراسة من المرضى المترددين على العيادات الخارجية في مستشفى الثورة –اب لـ (434) مريضًا، خلال الفترة الزمنية (2018/5/3 – 2018/8/25)؛ حيث تم تشخيص هؤلاء المرضى بناء على الصورة السريرية والتحاليل الطبية التي أجريت لهم وفق الأصول الطبية المعتمدة، وبموافقتهم الكاملة تم أخذ البيانات المطلوبة من ملفاتهم الطبية.

6-منهجية الدراسة وإجراءاتها:

اعتمدت الدراسة على منهجيين هما: المنهج الوصفي التحليلي المتمثل في قياس حجم البيانات المفقودة في بيانات الدراسة وتحديد آلية (نمط) البيانات المفقودة، وأخيرا وصف درجة كفاءة المعالجة للبيانات المفقودة بين النماذج الإحصائية المذكورة آنفًا باستخدام معيار الكفاءة النسبية؛ حيث الأخير (الكفاءة النسبية) تعتمد على تباينات المعالم المقدرة باستخدام النماذج الإحصائية المذكورة أنفًا وهنا تم الاعتماد على المنهج الاستدلالي بشكل ضمني. أما بالنسبة لإجراءات الدراسة تمت الدراسة على مرحلتين:

أ–مرحلة الدراسة النظرية:

بالاعتماد على المراجع المتخصصة والدراسات السابقة، تم توضيح أسباب ظهور هذه البيانات وأنماطها، وأيضًا تم شرح وتوضيح بعض الأساليب الإحصائية لمعالجة البيانات المفقودة وعلى وجه التحديد تم شرح وتوضيح طريقة تقدير (OLS)، طريقة تقدير (LME)، وأخيرًا طريقة تقدير (EM Algorithm).

ب–مرحلة الدراسة التطبيقية:

تم أخذ عينة من المرضى المترددين على العيادات الخارجية في مستشفى الثورة –اب لـ (434) مريضًا، خلال الفترة الزمنية (2018/5/3 – 2018/8/25). وبالتطبيق على بيانات العينة المختارة لتوليد قيم احتمالية باستخدام التوزيع الطبيعي (Normal Distribution) الاحتمالي كمعالجة إحصائية، وعلى ضوء البيانات المستخرجة من التوزيع الطبيعي (Normal Distribution) نعالج (نقدر) البيانات المفقودة باستخدام طريقة تقدير (OLS)، طريقة تقدير (LME)، طريقة تقدير (EM Algorithm). وأخيرًا مقارنة بين هذه الطرق باستخدام معيار الكفاءة النسبية، وذلك باستخدام برنامج (SPSS).

ثانيًا-الجانب النظري

يعرض هذا الجانب كيفية معالجة البيانات المفقودة باستخدام الأساليب الإحصائية. كما أوضحنا في المقدمة وأسباب ظهور البيانات المفقودة وأنواعها (أنماطها)، ونتيجة لذلك وبشكل عام، فإن للبيانات المفقودة أنماطًا وأشكالًا مختلفة في الدراسات العلمية بحسب أسباب ظهور هذه البيانات المفقودة، ومن ثم أي إهمال أو تجاهل لهذه البيانات يؤثر بدوره على عملية التحليل الإحصائي لهذه البيانات.

ومن ثم هناك عدة أساليب إحصائية لمعالجة البيانات المفقودة، وهذه الأساليب تعد وسائل لتقدير البيانات بشكل عام وبشكل خاص تعد وسيلة للتقدير للبيانات المفقودة؛ لذا تم التركيز في هذه الدراسة على نمط الفقدان العشوائي (MAR) وذلك بسبب اعتمادها على العشوائية في فقدان البيانات، ونتيجة لذلك (العشوائية) صار بالإمكان استخدام النماذج (الأساليب) الإحصائية لتقدير البيانات المبني على الافتراض نفسه (عشوائية ظهور المتغيرات).

وقبل التطرق للنماذج (الأساليب) الإحصائية لمعالجة البيانات المفقودة، من الضروري التعرف على المفاهيم والرموز الأساسية للبيانات المفقودة؛ حيث تمثل (M) مؤشر القيم المفقودة الناتج من تقدير المتغير العشوائي التابع (Y) الذي يحتوي على القيم المشاهدة والقيم المفقودة (Y_{obs} , Y_{Miss}) بحيث ؛ن كلا من القيم المشاهدة والقيم المفقودة (Y_{obs} , Y_{Miss}) لهما توزيع احتمالي معين، والهدف الرئيس من معرفة التوزيع الاحتمالي لكل من القيم المشاهدة والمفقودة يتمثل في وضع متغير وهمي (M) يحتوي على قيمتين الأولى عندما (m=0) ويقصد بها القيمة المفقودة (Y) عندما يكون المتغير المستقل (X) يحتوي على مشاهدات غير مفقودة، أما القيمة الأخرى ($m= 1$) ويقصد بها قيمة (Y) محددة عندما يكون المتغير المستقل ذا قيمة مفقودة لمعرفة التوزيع الاحتمالي، بمعنى آخر أن القيم التقديرية للبيانات المفقودة (Y_{Miss}) تقدر في حالة إذا ما استطعنا معرفة الدالة الاحتمالية المشتركة لكل من (M, Y_{obs}, X) (2) (16).

ومن أجل تطبيق نمط (آلية) الفقدان العشوائي (MAR) من الضروري معرفة الافتراضات الرئيسة لهذه الآلية وتتلخص في الآتي: عملية الفقدان للبيانات تتم بشكل عشوائي وليس لأحد الأسباب المذكورة آنفًا، وأيضًا تفترض أن يتم توزيع البيانات المفقودة وذلك بإنشاء عينات أخرى جزئية مستخرجة من بيانات الدراسة، حيث يتم توليد هذه العينات باستخدام طريقة المحاكاة (طريقة البوستراب أو طريقة مونت كارلو)، ونتيجة لتكوين هذه العينات تتكون لدينا مجاميع أو مجموعات عشوائية احتمالية أخرى تحتوي على مشاهدات وقيم مفقودة، ويمكن تمثيلها رياضيا بالشكل الآتي: $Pr(Y=0 \mid Y_{obs}, M)$ ، و أخيرًا تفترض على وجود ارتباطات بين العينات الجزئية التي تم توليدها(18).

بعد التعرف على المفاهيم والرموز الأساسية للبيانات المفقودة وتحديد نمط (MAR) للبيانات المفقودة، عند استخدام أي نموذج إحصائي لتقدير البيانات وعلى وجه التحديد البيانات المفقودة من الضروري التعرف على شروط وضوابط وخطوات التقدير لهذا النموذج (الأسلوب) الإحصائي.

لذا تم التركيز في هذه الدراسة على النماذج (الأساليب) التي تخضع التوزيع الطبيعي الاحتمالي، ذلك بسبب أن حد الخطأ –الذي يقصد به الفرق بين القيمة الحقيقية والقيمة المتوقعة (المتنبأ بها) –له عدة افتراضات منها أن يتبع التوزيع الطبيعي الاحتمالي، ومن النماذج (الأساليب) الإحصائية الي تعتمد على حد الخطأ طريقة تقدير (OLS) وطريقة تقدير (LME) وأخيرًا طريقة (EM) الخوارزمية، من ثم وجب علينا التعرف على شروط وضوابط وخطوات التقدير لهذه الطرق المقترحة.

1-طريقة تقدير المربعات الصغرى الاعتيادية (OLS):

تهدف طريقة تقدير (OLS) في إيجاد الخط المستقيم الذي يمر بنقاط الشكل الانتشاري بشكل يجعل مجموع مربعات الأخطاء أقل ما يمكن، وبمعنى آخر تحديد قيمة (β) المقدرة التي تجعل هذا المجموع أقل ما يمكن[7]؛ حيث يمكن تمثيل مجموع مربعات الأخطاء بالأشكال الآتية:

$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_i X_i)^2 = 0$ ، ويتم تقدير معلمات (β_0, β_i) وذلك بأخذ المشتقة الجزئية بالنسبة (β_0) مرة وبالنسبة (β_i) مرة أخرى وجعلها مساوية للصفر بذلك نحصل على معادلات يتم حلها لإيجاد معلمات نموذج الانحدار بواسطة العلاقة الآتية $\beta = (X/X)^{-1} (X/Y)$ وفي ضوء المعطيات المستخرجة من العلاقة السابقة نحدد مصفوفة التباين والتباين المشترك لمعلمات (β) التي تحدد بالعلاقة الآتية: $(X/X)^{-1}\sigma^2$ [14].

2-طريقة تقدير الإمكان الأعظم (LME):

تهدف طريقة تقدير (LME) في إيجاد المعلمات التي تجعل الإمكان في نهايتها العظمى؛ حيث تستخدم هذه الطريقة لتقدير معلمات نموذج الانحدار في ضوء المشتقات الجزئية، وبمعنى آخر يمكن تقدير معلمات نموذج الانحدار عندما يمتلك متغير حد الخطأ توزيعًا احتماليًا معروفًا[15]، يمكن إيجاد مقدرات (معلمات) الإمكان الأعظم لدالة حد الخطأ التي تكتب رياضيًا بالشكل الآتي[7]:

$$\Pr(Y_i, \beta_0, \beta_1, \sigma^2) = (2\prod_{i=1}^{n} \sigma_u^2)^{-\frac{n}{2}} \exp\left(\frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_i X_i)^2}{2\sigma_u^2}\right)$$

ويتم تقدير معلمات (β_0, β_i) بالخطوات الآتية: 1) بأخذ اللوغاريتم الطبيعية لدالة الإمكان الأعظم، ومن ثم إيجاد المشتقات الجزئية بالنسبة لـ (β) ومساواتها بالصفر؛ 2) وبحل نتيجة الخطوة السابقة نحصل على مقدرات الإمكان الأعظم التي تعبر بالعلاقة الآتية: [8]

$\beta = (X'X)^{-1}(X'Y)$ وفي ضوء المعطيات المستخرجة من العلاقة السابقة نحدد مصفوفة التباين والتباين المشترك لمعلمات (β) التي تكتب بالشكل الآتي: $(X'X)^{-1}\sigma^2$ وهي مقدرات طريقة (OLS) نفسه[14].

3-طريقة تقدير (EM Algorithm):

كما هو متعارف في الدراسات الإحصائية فإن طريقة تقدير (EM Algorithm) تعد طريقة بديلة للإمكان الأعظم؛ حيث تعمد طريقة (EM Algorithm) في التقدير إلى المنهج التكراري للحصول على معلمات تتمتع بالكفاية الإحصائية[15]؛ لذا طريقة (EM) الخوارزمية للتقدير المعلمات لها أهمية قصوى وذلك للأسباب الآتية: 1) كما ذكرنا سابقًا أن هذه الطريقة تعد طريقة بديلة للإمكان الأعظم، وذلك بسبب أن طريقة تقدير (LME) في أغلب الأحيان لا يمكنها تقدير معلمات الدوال الاحتمالية أو الحصول على حلول بشكل نهائي لمعلمات الدوال الاحتمالية، وبمعنى آخر طريقة تقدير (LME) لا يمكنها إيجاد قيمة المشتقات الجزئية لمعلمات (β) بالذات في حالة إذا كان التوزيع الاحتمالي لحد الخطأ غير معروف، ومن ثم لا نستطيع إيجاد حلول نهائية لمعلمات الدوال الاحتمالية؛[18] 2) تعد طريقة من طرق التقدير للبيانات في حالة وجود البيانات المفقودة؛[15] 3) احتواء طريقة كل من (LME) أو(OLS) ضمن خطوات الحل لهذه الطريقة [18].

ومن ثم يتم تقدير معلمات هذه الطريقة بالخطوات الآتية: 1) يتم إيجاد الدالة الاحتمالية $Pr(Y_i \mid \theta) = g(Y_i)$ حيث $g(Y_i)$ دالة لا تتضمن بها (θ) والمؤشر الأخير عبارة عن معلمة تتمتع بالكفاية الإحصائية.2) يتم التقدير لكل من المتوسط والتباين باستخدام طريقة (OLS) أو(LME)؛ 3) يتم تعديل مؤشر (θ) بشكل متكرر إما باستخدام طريقة(The E-step) وهي تقوم على أساس وضع قيمة أولية لمعلمة البيانات المفقودة المستخرجة في الخطوة رقم (2) ويتم تعديل هذه القيم حتى الحصول على معلمة تتمتع بالكفاية الإحصائية، أو باستخدام طريقة (The M-step) وهي تقوم على أساس استخدام المشاهدات والمعلمات التي تتمتع بالكفاية الإحصائية المستخرجة من نموذج (OLS) أو (LME) ووضعها كقيمة أولية، ويتم تكرار وتعديل قيم (θ) للحصول على مؤشر يتمتع بالكفاية الإحصائية [18].

خلاصة الجانب النظري هناك أسباب عدة لظهور البيانات المفقودة وخصوصًا ما قد يكون ناجمًا عن أسباب فنية أو أسباب يمكن تلافيها، لكن إذا استطعنا تلافي هذه الأسباب وظهرت لدينا بيانات مفقودة في اللحظة يتوجب علينا تحديد نمط (آلية) البيانات المفقودة، وفي ضوء نوع آلية البيانات المفقودة يتحدد لدينا نوع النماذج (الأساليب) الإحصائية؛ حيث إن لكل من هذه الأساليب أو النماذج ضوابط وشروط معينة لاستخدام تلك النماذج؛ ونتيجة لذلك (الشروط والضوابط) تم التركيز في هذه الدراسة على النماذج (الأساليب) الإحصائية التي تخضع للتوزيع الطبيعي وعلى وجه التحديد تم التركيز على ثلاثة نماذج وهي نموذج (OLS) ونموذج (LME) وأخيرًا نموذج (EM Algorithm).

بعد التعرف على الشروط والضوابط للأساليب (النماذج) الإحصائية المذكورة في هذا الجانب (الجانب النظري)، فإن بإمكاننا الآن التأكد من سلامة هذا الجانب وذلك من خلال التطبيق العملي (الجانب التطبيقي).

ثالثًا-الجانب التطبيقي:

يعرض هذا الجانب من الدراسة الإجابة عن مشكلة الدراسة الرئيسة المتمثلة بالسؤال الآتي (ما هي أفضل طريقة إحصائية لمعالجة البيانات المفقودة؟) ويتفرع من هذا السؤال أسئلة فرعية أخرى تتمثل في تحديد طبيعة البيانات والبيانات المفقودة ونوعها، وكيفية تحليلها إحصائيًا وتحديد نمط (آلية) البيانات المفقودة، وأيضًا إيجاد القيم الاحتمالية للبيانات المفقودة باستخدام التوزيع الطبيعي (Normal Distribution) الاحتمالي كمعالجة إحصائية للبيانات قبل إدخالها لعملية التقدير، وأخيرًا دراسة مقارنة أو المفاضلة بين طرق التقدير المذكورة آنفًا في متن الدراسة، وذلك باستخدام قانون الكفاءة النسبية. للإجابة عن الأسئلة تم جمع بيانات الدراسة من مستشفى الثورة في محافظة إب لـ (434) مريض؛ إذ تم اعتماد (13) متغيرًا أحدها متغير تابع (Y) والباقي متغيرات مستقلة عن الأمراض السريرية وعلى وجه التحديد عن مرض السكري (Blood Sugar)، والجدول الآتي يوضح قائمة المصطلحات الطبية المستخدمة في هذه الدراسة.

جدول(1)
جدول يبين قائمة المصطلحات الطبية المستخدمة في هذه الدراسة

التوضيح	المصطلح
ويقصد به طب المختبرات وهو تخصص طبي يهتم بتشخيص المرضى بالاعتماد على التحليل المختبري لسوائل الجسم مثل الدم والبول، الأنسجة أو مقتطفات من الجسم باستخدام أدوات الكيمياء وعلم الأحياء المجهرية وعلم الأمراض الجزيئي وأخيرًا علم أمراض الدم.	الامراض السريرية Clinical Pathology
يعد مرض السكري أحد أمراض الدم ويقصد به متلازمة تتصف باضطراب الأيض وارتفاع شاذ في تراكيز سكر الدم الناجم عن عوز (احتياج) هرمون الانسولين أو كلا الأمرين معًا، ويصاحب مرض السكري العديد من الحالات المرضية التي تتطلب الفحص لمعرفة الأمراض المصاحب لمرض السكري أم لا مثل مستوى الكوليسترول، فحص الدم العام، وتراكم الدهون على الكبد، والفيروسات، ويرمز له بالرمز (Blood Sugar).	المرض السكري Diabetes Mellitus
يعد الكوليسترول أحد أمراض الدم المصاحبة لمرض السكري وهي مادة دهنية شمعية أساسية في تكوين أغشية الخلايا في جميع أنسجة الكائنات الحية، ويرمز له بالرمز (Chol).	الكوليسترول Cholesterol
يعد البروتين الدهني مرتفع الكثافة أحد أنواع الكوليسترول، وهذا النوع مفيدٌ لصحة الإنسان، ويرمز له بالرمز (HDL).	البروتين الدهني مرتفع الكثافة High-density lipoprotein
يعد البروتين الدهني منخفض الكثافة أحد أنواع الكوليسترول، وهذا النوع ضارٌ لصحة الإنسان، ويرمز له بالرمز (LDL).	البروتين الدهني منخفض الكثافة Low-density lipoprotein
تعد أحد خلايا الدم، ووظيفتها الدفاع عن الجسم من الأمراض المعدية، ويرمز له بالرمز (WBC).	كريات الدم البيضاء White Blood Cell
وهو بروتين محمول داخل خلايا الدم الحمراء ويحتوي على ذرات الحديد، التي تساعد في التقاط الأوكسجين من الرئتين ويسلمه إلى الأنسجة بواسطة كريات الدم الحمراء، ويرمز له بالرمز (Hb).	الهيموغلوبين Hemoglobin
ويقصد به متوسط قياس حجم كرية الدم الحمراء ويستخدم للتفريق بين أنواع فقر الدم (Hb)، ويرمز له بالرمز (MCV).	تحليل الحجم الكروي الوسطي Mean Corpuscular Volume

المصطلح	التوضيح
تحليل الهيموغلوبين الكروي الوسطي Mean Corpuscular Hemoglobin	ويقصد به متوسط كتلة الهيموغلوبين في كريات الدم الحمراء الواحدة في عينة الدم، ويرمز له بالرمز (MCH).
تحليل الهيموغلوبين الكروي الوسطي-C Mean Corpuscular Hemoglobin-C	وهو هيموغلوبين غير طبيعي حيث إذا حدث أي إحلال أو استبدال بقايا حمض الجلوتاميك، ووضع بقايا الحمض الأميني لايسين على الموقع السادس من سلسلة (B-globin)، وهذا يعد مرضا يصيب به الإنسان، الدم، ويرمز له بالرمز (MCHC).
حجم الخلايا المكدسة في الدم Hematocrit	ويقصد به الراسب الدموي أو النسبة المئوية لحجم خلايا الدم الحمراء من إجمالي الدم، ويرمز له بالرمز (PCV).
البروتين المتفاعل-C C-reactive protein	وهو بروتين متواجد في الدم بمستويات مختلفة، ترتفع في حالة تواجد أمراض فيروسية والالتهابات في الدم، ويرمز له بالرمز (CRP).
ناقلة أمين الألانين Alanine Transaminase	وهو أنزيم ناقل يتواجد بكميات مرتفعة في الكبد، ويستخدم هذا المؤسر لتشخيص مرض الكبد المزمن، ويرمز له بالرمز (ALT).
الفيروسات Virus	وهي عامل ممرض صغير لا يمكنه التكاثر إلا داخل خلايا كائن حي آخر؛ حيث ترتبط فحوصات الفيروسات بفحوصات أخرى مثل البروتين المتفاعل-C، نقص المناعة وغيرها من الفحوصات، ويرمز له بالرمز (VIRUS).
ملاحظة: 1. تم جمع هذه المعلومات من موقع (الويكيبيديا).	

وبالاعتماد على النموذج الانحدار الآتي: $Y_i = \beta_0 + \beta_i X_i + U_i$ ، لتقدير قيم معلمات (β) وأيضًا مصفوفة التباين– والتباين المشترك لـ (β) باستخدام الطرق المذكور آنفًا في متن الدراسة، كانت الإجابات بشكل الآتي:

1-تحديد طبيعة البيانات والبيانات المفقودة ونوعها في الدراسة:

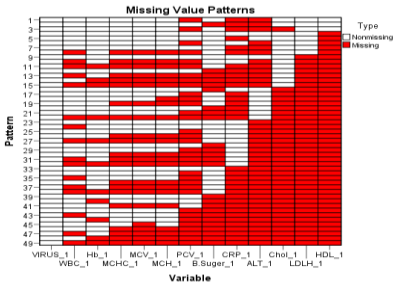
ذكرنا سابقًا الرموز الأساسية للبيانات المفقودة بشكل عام وبحسب بيانات الدراسة التي تم سحبها من المجتمع، تم تحديد المتغيرات المستقلة (X_i) إلى نوعين: متغيرات كمية وتتمثل بـ (Hb , PCV, MCH, MCHC, WBC, ALT, CHOL , HDL, LDL)، ومتغيرات نوعية (أسمي و رتبي) وتتمثل بـ (CRP, Virus) على التوالي، وتم تحديد المتغير التابع المتمثل بـ (Blood Sugar) وهو متغير كمي، نضيف متغيرات جديدة لضمن متغيرات الدراسة؛ حيث كل متغير من المتغيرات عبارة عن متغير (M) يحتوي على قيمتين، الأولى عندما (m=0) ويقصد بها قيمة مفقودة عندما يكون المتغير المستقل (X) يحتوي على مشاهدات غير مفقودة، أما القيمة الأخرى (m= 1) ويقصد بها قيمة محددة عندما يكون المتغير المستقل (X) ذا قيمة مفقودة.

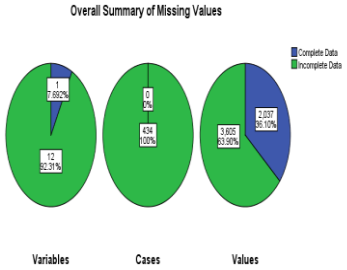
2-التحليل الوصفي للبيانات المفقودة:

هذا الجانب يعرض التحليل الوصفي لبيانات الدراسة ودراسة أنماط (آلية) البيانات المفقودة، وعلى وجه التحديد دراسة آلية أو نمط الفقدان العشوائي (MAR) للتأكد من صلاحية البيانات وآلية الفقدان العشوائي، وعند التطبيق باستخدام (SPSS) كانت النتائج بالشكل الآتي:

جدول(2)

جدول يبين أنماط والتحليل الوصفي للبيانات المفقودة التابعة لمتغيرات الدراسة





متغيرات الدراسة	الرمز العلمي	البيانات المفقودة		عدد المشاهدات	المتوسط	الانحراف المعياري
		العدد	النسبة			
البروتين الدهني مرتفع الكثافة	HDL	430	99.1%	4	2.50	1.291
البروتين الدهني منخفض الكثافة	LDL	424	97.7%	10	6.50	3.028
الكوليسترول	Chol	416	95.9%	18	9.50	5.339
ناقلة أمين الألانين	ALT	415	95.6%	19	10.00	5.627
البروتين المتفاعل-C	CRP	403	92.9%	31	3.23	1.087
المرض السكري	B.Suger	363	83.6%	71	30.62	15.852
حجم الخلايا المكدسة في الدم	PCV	305	70.3%	129	50.23	27.401
تحليل الهيموغلوبين الكروي الوسطي	MCH	192	44.2%	242	49.29	23.811
تحليل الحجم الكروي الوسطي	MCV	192	44.2%	242	83.89	41.881
تحليل الهيموغلوبين الكروي الوسطي-C	MCHC	191	44.0%	243	47.42	19.527
الهيموغلوبين	Hb	143	32.9%	291	33.75	21.693
كريات الدم البيضاء	WBC	131	30.2%	303	98.70	52.774

يلاحظ من الجدول (2) الآتي:

أولًا–من ناحية العرض البياني: يلاحظ أن عدد متغيرات الدراسة بلغ (13) متغيرًا أحدها المتغير التابع كان نسبة الفقدان بين المتغيرات بشكل عام (92.31 %) بمعنى أن هناك (12) متغيرًا من متغيرات الدراسة وجد بها بيانات مفقودة، ومتغير واحد لا يوجد فيه بيانات مفقودة وهو متغير (virus)؛ لذا تم حذف هذا المتغير من التحليل، وأيضا يلاحظ أن كل الحالات المرضية كانت بها بيانات مفقودة؛ إذ بلغت نسبة البيانات المفقودة بين الحالات المرضية بشكل عام (100%)؛ حيث كانت الحالة المرضية رقم (49) يوجد بها أعلى نسبة في البيانات المفقودة، وأقل حالة مرضية للبيانات المفقودة هي الحالة المرضية رقم (1)؛ إذ إن نسبة البيانات المرصودة (المشاهدة) بين الحالات المرضية تتراوح ما بين (29 % – 3 %)، وأخيرًا يلاحظ أن نسبة البيانات المفقودة بين قيم المتغيرات بشكل عام كانت (63.90 %).

ثانيًا–من ناحية الجدول: يلاحظ أن أعلى البيانات المفقودة بين المتغيرات الدراسة هي (HDL, LDL, Chol, ALT, CRP)؛ حيث بلغت نسبة البيانات المفقودة بنسبة (99.1%, 97.7% ، 95.9% , 95.6% , 92.9%) على التوالي. وتأتي المجموعة الثانية المتغيرات (Blood

(Sugar , PCV)؛ حيث بلغ البيانات المفقودة (70.3% ,83.6%) على التوالي. و أخيرًا المجموعة الثالثة (MCH, MCV, MCHC, Hb, WBC)؛ حيث بلغت البيانات المفقودة (30.2% ,32.9% ,44.0% ,44.2% ,44.2%) على التوالي. ويلاحظ أن المتوسطات والانحرافات المعيارية لمتغيرات الدراسة اختلفت بحسب القيم المفقودة في متغيرات الدراسة كما هو موضح في الجدول السابق.

مما سبق نستنتج أن بيانات الدراسة احتوت على بيانات مفقودة ويرجع ذلك إلى أن التحاليل الطبية التي طلبت من المرضى كانت بحسب تشخيص الطبيب لحالة المريض في تلك اللحظة، ونتيجة لاختيار عينة وفحوصاتهم العشوائية وظهور البيانات المفقودة، وبحسب الرسم البياني الموضح في الجدول رقم (2) للبيانات المفقودة التي تبين توزيع ونمط البيانات المفقودة، من ثم أخذت البيانات المفقودة نمط (آلية) الفقدان العشوائي (MAR). وبعد التأكد من آلية (نمط) البيانات المفقودة في الدراسة، نقوم بمعالجة بيانات الدراسة وذلك بإنشاء وتوليد القيم الاحتمالية للبيانات على نمط الفقدان العشوائي.

3-إنشاء القيم الاحتمالية على نمط الفقدان العشوائي (MAR):

يتم إضافة متغيرات جديدة (Z_i) بحسب عدد متغيرات الدراسة وعلى وجه التحديد المتغيرات التي تم ذكرها في الجدول رقم (2)؛ حيث إن هذه المتغيرات (Z_i) تحتوي على قيم متغيرات الدراسة في أي ربع تقع هذه القيم، بشرط أن يكون لهذه القيم فقدان بنسبة $Z_i = (0.20), (0.40), (0.60), (0.80)$ على التوالي. والهدف الرئيس من إضافة هذه المتغيرات (Z_i) هو أن تكوين مجاميع أو عينات أخرى من متغيرات الدراسة يكون البيانات فيها عبارة عن قيم احتمالية لظهور أول قيمة مفقودة في بيانات الدراسة، وهذه القيم الاحتمالية تم استخراجها باستخدام برنامج (SPSS)، بالاعتماد على التوزيع الطبيعي (Normal Distribution)، ولتحديد معالم هذا التوزيع (المتوسط والانحراف المعياري) تم استخدام المتوسط والانحراف المعياري للبيانات المفقودة المذكور في الجدول رقم (2) لتوليد قيم احتمالية لكل متغير من متغيرات الدراسة. وكانت نتائج هذه العملية بالشكل الآتي:

جدول(3)

جدول يبين القيم الاحتمالية لظهور أول قيمة مفقودة باستخدام توزيع Normal Distribution لمتغيرات الدراسة

متغيرات الدراسة	الرمز العلمي	عدد المشاهدات	متوسط القيم الاحتمالية	الانحراف المعياري	البيانات المفقودة	
					العدد	النسبة
الهيموغلوبين	Hb	291	.0129013791	.0052705843465	143	32.9
حجم الخلايا المكدسة في الدم	PCV	129	.0091444233	.0041845381237	305	70.3
تحليل الحجم الكروي الوسطي	MCV	242	.0071908368	.0022874903480	192	44.2
تحليل الهيموغلوبين الكروي الوسطي	MCH	242	.0095341226	.0053662058488	192	44.2
تحليل الهيموغلوبين الكروي الوسطي-C	MCHC	243	.0088975214	.0067695318708	191	44.0
كريات الدم البيضاء	WBC	303	.0056098087	.0018432577690	131	30.2
المرض السكري	B.SUGER	71	.0086428638	.0095148371202	363	83.6
ناقلة أمين الألانين	ALT	19	.0004274436	.0008504230984	415	95.6
الكوليسترول	CHOL	18	.0002424304	.0004666948724	416	95.9
البروتين الدهني منخفض الكثافة	LDL	10	.0000000002	.0000000004356	424	97.7
البروتين الدهني مرتفع الكثافة	HDL	4	.0000000000	.0000000000000	430	99.1
البروتين المتفاعل-C	CRP	31	.0000000000	.0000000000000	403	92.9

يلاحظ من الجدول (3) الآتي:

- أن نسب البيانات المفقودة لمتغيرات الدراسة في هذا الجدول مطابقة لما جاء في الجدول رقم (2) حيث بلغ أكبر متغير في فقدان البيانات هو متغير (HDL) بنسبة (99.1 %) وأقل متغير في فقدان البيانات هو متغير (WBC) بنسبة (30.2 %).
- أن كل من المتغيرات (LDL,HDL,CRP) كانت المتوسطات الاحتمالية والانحرافات المعيارية قريبة جدا من الصفر، ومن ثم يمكننا حذف هذه المتغيرات من عملية التقدير؛ لأن قيمتها في عملية التقدير ستكون مساوية أو قريبة جدًا من الصفر، لكن الباحث رأى أن تبقى هذه المتغيرات للتأكد من قيمتها الفعلية بعد عملية التقدير.

وبشكل عام فإن هدف هذه الخطوة هو تكوين القيم الاحتمالية لمتغيرات الدراسة بحيث يكون لهذه القيم فقدان بنسبة $Z_i = (0.20), (0.40), (0.60), (0.80)$، وأيضا معرفة أي من المتغيرات التي يمكن أن تستبعد أو تلغى بشكل أولي (مبدئي)، وأخيرًا إخضاع متغيرات الدراسة لتوزيع الطبيعي كمعالجة إحصائية قبل إدخال البيانات في عملية التقدير لكل من طرق التقدير (OLS) و (LME) و (EM) الخوارزمية.

4-المفاضلة بين تقدير (مقدرات) معلمات طريقة (OLS) و(LME) و (EM) الخوارزمية:

وبعد اختبار البيانات لفروض التحليل (البيانات المفقودة)، تم إيجاد معلمات (β) لنماذج الانحدار وعلى وجه التحديد معلمات طريقة (OLS) وطريقة (LME) وأخيرًا طريقة (EM) الخوارزمية وأيضًا

تحديد مصفوفة التباين–والتباين المشترك لـ معلمات (β)، وباستخدام برنامج (SPSS) تم عمل جدول مقارنة لمصفوفة التباين–والتباين المشترك لـ معلمات (β)، وكانت النتائج كما في الجدول رقم (4):

جدول(4)

جدول يبين تباين معلمات طريقة (OLS) (LME) و (EM) الخوارزمية لمتغيرات الدراسة

متغيرات الدراسة	الرمز العلمي	تباين طريقة(OLS)	تباين طريقة(LME)	تباين طريقة(EM)
الهيموغلوبين	_Hb	.0000284964977	.0000284964977	.0000278893506
حجم الخلايا المكدسة في الدم	PCV	.0000152443669	.0000152443669	.0000420223693
تحليل الحجم الكروي الوسطي	MCV	.0000050778229	.0000050778229	.0000052115391
تحليل الهيموغلوبين الكروي الوسطي	MCH	.0000289589916	.0000289589916	.0000286941966
تحليل الهيموغلوبين الكروي الوسطي-C	MCHC	.0000434189293	.0000434189293	.0000457600931
كريات الدم البيضاء	WBC	.0000032888941	.0000032888941	.0000034086309
المرض السكري	B.SUGER	.0000884036527	.0000884036527	.0000953129583
ناقلة أمين الألانين	ALT	.0000006884124	.0000006884124	.0000012673830
الكوليسترول	CHOL	.0000001994313	.0000001994313	.0000006105298
البروتين الدهني منخفض الكثافة	LDL	.0000000000000	.0000000000000	0.000000000000
البروتين الدهني مرتفع الكثافة	HDL	.0000000000000	.0000000000000	0.000000000000
البروتين المتفاعل-C	CRP	.0000000000000	.0000000000000	0.000000000000

يلاحظ من الجدول (4) الآتي:

- أن معلمات كل من طريقة (OLS) و (LME) تساوت في قيم التقدير، وهذا مما يؤكد الجانب النظري في صفحة (10–11).
- أن كل من المتغيرات (LDL,HDL,CRP) كانت نتيجة تقدير التباين مساوية للصفر، وهذا مما يؤكد ما جاء في الجدول رقم (3)؛ لذا تم حذف هذه المتغيرات من الدراسة.
- هناك فروق طفيفة في تباين مقدرات بين طرق (OLS , LME) و (EM) الخوارزمية.
- وعند المفاضلة لمصفوفة التباين–والتباين المشترك لمعلمات (β) للطرق المذكورة آنفًا، وباستخدام صيغة الكفاءة النسبية المتمثلة في العلاقة الآتية: $\mathrm{eff}(\hat{\beta}_i) = \frac{\mathrm{var}(\hat{\beta}_i)\ \mathrm{OLS}}{\mathrm{var}(\hat{\beta}_i)\ \mathrm{LME}}$، فإذا كانت النتيجة أقل من الواحد يعني التقدير بموجب (OLS) أكثر كفاءة من التقدير بطريقة (LME) والعكس صحيح، وبحسب الجدول رقم (4) وعلى وجه التحديد قيم التقدير كل من (OLS) و (LME) كانت متساوية أي تساوي الواحد، ومن ثم فإن كفاءة التقدير لكل من (OLS) و (LME) متساوية، وأيضًا عند المفاضلة بين طريقتين (OLS) و (EM) الخوارزمية كانت كفاءة التقدير بين الطريقتين متذبذبة؛ حيث كانت كفاءة التقدير طريقة (EM) الخوارزمية أفضل من طريقة (OLS) بالنسبة للمتغير (Hb)، وجاء العكس للمتغير (PCV)، أما بالنسبة لبقية المتغيرات كانت كفاءة طريقة التقدير (OLS) هي الأفضل. للفصل بين هذه الطريقتين تم احتساب المتوسط العام للتباينات المقدرة، وكانت النتائج كما هو في الجدول رقم (5):

جدول(5)
جدول يبين المتوسطات العام لتباينات المقدرة

متغيرات الدراسة	تباين طريقة(OLS)	تباين طريقة(LME)	تباين طريقة(EM)
المتوسط العام لمتغيرات الدراسة	0.0000178	0.0000178	0.0000205

من الجدول رقم (5) نستطيع الحكم بين الطريقتين (OLS) و (EM) الخوارزمية؛ حيث كانت أفضل كفاءة لتقدير متغيرات الدراسة هي طريقة (OLS).

رابعًا-الاستنتاجاتوالتوصيات:

مما سبق يمكن عرض أهم الاستنتاجات التي أسفرت عنها الدراسة بما يتفق مع أهداف الدراسة، وخاصة في تحديد الطريقة المثلى لمعالجة البيانات المفقودة، ووضع بعض التوصيات.

1-الاستنتاجات:

خلصت الدراسة إلى مجموعة من الاستنتاجات وذلك من خلال النتائج التي تم التوصل إليها من تطبيق الطرق الإحصائية المقترحة لمعالجة البيانات المفقودة في الدراسة، والتي تتلخص في عدة نقاط هي:

- جميع الطرق الإحصائية المذكورة آنفًا لها كفاءة متقاربة في معالجة وتقدير البيانات المفقودة، وذلك بسبب إخضاع متغيرات الدراسة للتوزيع الاحتمالي الطبيعي (Normal Distribution) وتحويلها إلى قيم احتمالية متجانسة؛ وهو ما ساعد في تحسين جودة تقدير البيانات المفقودة باستخدام الطرق الإحصائية المذكورة في متن الدراسة.
- على الرغم من أن الطرق الإحصائية المذكورة آنفًا لها الكفاءة نفسها في التقدير، إن طريقة المربعات الصغرى لها الأفضلية من بين هذه الطرق، وهذه النتيجة تؤكد النتيجة السابقة.
- على الرغم من أن طريقة (EM) الخوارزمية أكثر تطورًا من طريقة (OLS) من الناحية الإحصائية، إن نتيجة التقدير للبيانات المفقودة جاءت لصالح طريقة (OLS)، وذلك بسبب أن طريقة (OLS) تعد جزءًا لا يتجزأ من عملية التقدير باستخدام طريقة (EM) الخوارزمية، ونتيجة هذا التداخل بين الطريقتين وقع تحيز بسيط في عملية تقدير البيانات المفقودة.
- تتساوى معلمات (مقدرات) طريقة المربعات الصغرى(OLS) وطريقة الإمكان الأعظم (LME)، ومن ثم تتساوى كفاءة التقدير لكلا الطريقتين.

2-التوصيات:

- الأخذ بعين الاعتبار البيانات المفقودة عند إجراء أي دراسة علمية والتعامل معها بشكل عملي (إحصائي) وعدم اللجوء لحذف هذه البيانات من الدراسة.
- استخدام التوزيعات الاحتمالية في معالجة البيانات المفقودة لما تتمتع بها من خصائص ومميزات بحسب طبيعية المتغيرات المدروسة.
- الاستمرار في استخدام النماذج الإحصائية والاعتماد عليها في معالجة البيانات المفقودة.

خامسًا-قائمة المراجع:

1-المراجع العربية:

1. الرحيل، راتب صايل الخضر والدرابسة، رياض أحمد صالح (2014)" أثر طريقتي التعامل مع القيم المفقودة، وطريقة تقدير القدرة على دقة تقدير معالم الفقرات والافراد" المجلة الدولية التربوية المتخصصة-الجمعية الأردنية لعلم النفس، الأردن، العدد (6)، المجلد (3).
2. المشهداني، كمال علوان خلف وأحمد، أحمد شهاب (2017)" تقدير القيمة المفقودة باستخدام تحليل التغاير لتصميم القطاعات المنشقة" مجلة العلوم الاقتصادية والإدارية-جامعة بغداد، العراق، العدد (100)، المجلد(23)، ص 455-472.
3. المنجي، هشام محمد (2012)" تطوير بعض أساليب التحليل الإحصائي لرفع كفاءة المسح البعدي لسكان" رسالة دكتوراه غير منشورة، قسم الإحصاء التطبيقي والتامين، كلية التجارة، جامعة المنصورة، مصر.
4. كاظم، أموري هادي ونايف، قتيبة نبيل (2008) " أسلوب بيز في تحليل البيانات غير التامة" مجلة العلوم الاقتصادية والإدارية-جامعة بغداد، العراق، العدد (49)، المجلد (14)، ص 251-263.
5. عبد الملك، حسن حسن علي (2020) "توفيق نماذج توزيعات (Tweedie) الاحتمالية مع التركيب العمري والنوعي للسكان في اليمن -دراسة تطبيقية" مجلة جامعة الناصر-اليمن، (مقبول وقيد النشر)، يناير-يونيو العدد (15).
6. عبد الملك، حسن حسن علي (2020) "دراسة مقارنة بعض طرق تقويم بيانات التركيب العمري والنوعي لسكان اليمن" مجلة جامعة الجزيرة – محافظة أب-اليمن، يوليو، العدد (6)، المجلد(3)، ص 237-256.
7. طه، حازم عمار، صفاء يونس الصفاوي (2005)" بعض طرائق المقدرات التقليدية ومقدر بيز لمعلمات نموذج الانحدار الخطي" مجلة تنمية الرافدين-كلية الإدارة والاقتصاد – الموصل، العراق، العدد (27)، المجلد(8)، ص 91-103.
8. رشيد، ظافر حسين وحمزة، سعد كاظم (2016) " مقارنة بعض الطرائق التقدير المويجي لدالة الانحدار اللامعلمي عند فقدان متغير الاستجابة عشوائيًا" مجلة العلوم الاقتصادية والإدارية-جامعة بغداد، العراق، العدد (90)، المجلد(22)، ص382-406.
9. ديلو، فضيل (2014)" معايير الصدق والثبات في البحوث الكمية والكيفية" مجلة العلوم الاجتماعية-فلسطين، العدد(83).

2-المراجع الاجنبية:

10. Wu,Di & Ma ,Jinwen (2019)" An effective EM algorithm for mixtures of Gaussian processes via the MCMC sampling and approximation" Neurocomputing, ELSEVIER, Volume 331, 28 February 2019, Pages 366-374
11. Taylor, Jennifer & Lacovara ,Alicia & Smith Gordon & RaviPandian and MarkLehto (2014) " Near-miss narratives from the fire service: A Bayesian analysis" Accident Analysis & Prevention, ELSEVIER , Volume 62, January 2014, Pages 119-129
12. KNOPP, Jeremy& GRANDHI, Ramana & Li ZENG and ALDRIN, John (2012)"Considerations for Statistical Analysis of Nondestructive Evaluation Data: Hit/Miss Analysis" Japan Society of Maintenology, E-Journal of Advanced Maintenance, Volume 4, No.3, (2012), Pages 105-115.
13. Zhao, Jun & Kim , Hea-Jung and Kim, Hyoung-Moon (2020)" New EM-type algorithms for the Heckman selection model" Computational Statistics & Data Analysis, ELSEVIER ,Volume 146, June 2020, 106930

14. Marques de Sá, Joaquim (2007) "Applied Statistics Using SPSS, STATISTICA, MATLAB and R" Springer Berlin Heidelberg, New York, ISBN 978-3-540-71971-7.
15. Chihara, Laura & Hesterberg, Tim (2019)" Mathematical Statistics with Resampling and R "second edition, John Wiley & Sons, Inc, United States of America, and ISBN 9781119416531.
16. National Research Council (2010)" The Prevention and Treatment of Missing Data In clinical Trial" Panel and Handling Missing Data In clinical Trial, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Washington , DC: The National Academies Press.
17. Singh, s. & Dew, B (2003) "Imputation by Power Transformation" Statistical Papers, 44.
18. Raghunthan ,Trivellore (2016)" Missing Data Analysis in Practice" CRC Press is an imprint of The Taylor and Francis Group, an Informal Business, Intentional Standard. Book Number 13:978-1-4822-1193-1.
19. Wang, Yanwei & LI, Jian & and Stoical, Peter (2005)" Spectral Analysis of signals, the Missing Data Case" Library of Congress Cataloging-in-Publiction Data, USA, INBN: 15 -98290002.
20. Zhang ,Zhipeng & Trivedi ,Chintan and Liu ,Xiang (2018) " Automated detection of grade-crossing-trespassing near misses based on computer vision analysis of surveillance video data" Safety Science, ELSEVIER ,Volume 110, Part B, December 2018, Pages 276-285.