التعامل مع مشاكل القيم المفقودة في البيانات الطولية

Dealing with Missing Values Problems in Longitudinal Data

الأستاذ الدكتور/ إبراهيم محمد مهدي

أستاذ الرياضيات والإحصاء الإكتواري

الأستاذ الدكتور/ البيومي عوض طاقية

أستاذ الإحصاء التطبيقي ووكيل

كلية التجارة لشؤون الطلاب

حنين ناجي صبري أبو صالح

مدرس مساعد بقسم الإحصاء التطبيقي والتأمين

كلية التجارة- جامعة المنصورة

الملخص

تكمن مشكلة الدراسة في وجود فقد متقطع (فقد غير متكرر على وتيرة واحدة) في التغايرات المستمرة الطولية في حين أن الإستجابات ثنائية كاملة. وتعتبر هذه المشكلة من المشاكل التي لم يتم التطرق إليها كثيرا في الدراسات التي إهتمت بفقد البيانات لصعوبة معالجتها وخاصة أن الفقد غير متكرر على وتيرة واحدة.وضع البحث مجموعة من الأسس التي تمكن الباحثين من التعامل مع الفقد في البيانات بشكل أفضل بسبب عدم التحديد عند حل مشكلات الفقد مما يشتت الباحث ويبعده عن موضوع البحث، ومعالجة الفقد في التغايرات المستمرة الطولية والإستجابة الثنائية الكاملة في حالة الفقد القابل للتجاهل وغير القابل للتجاهل وغير المتكرر على وتيرة واحدة في النماذج الخطية المحتلطة المعممة بإستخدام التعويض المتعدد. النماذج الخطية المختلطة المعممة لتوفيق البيانات الطولية المستمرة ومقارنتها بطريقة تحليل الحالة الكاملة.

المقدمة

تلعب الدراسات الطولية دورا بارزا في العديد من التخصصات مثل الطب والصحة العامة والعلوم الإجتماعية. وتستخدم البيانات الطولية بشكل كبير سواء في الدراسات التي تعتمد على بيانات ملاحظة أو في الدراسات التجريبية. ويتم فيها تتبع الأفراد خلال فترة من الزمن، ويتم تجميع البيانات في نقاط زمنية متعددة. وبالتالي فإن الصفة المميزة للبيانات الطولية هي أنه يتم تجميع قياسات متكررة ومتعددة لنفس المتغيرات لكل فرد في الدراسة خلال فترة من الزمن (Wu, 2009). وتختلف الدراسات الطولية عن الدراسات المستعرضة التي تقيس المتغير مرة واحدة فقط عند نقطة زمنية واحدة (Gad and Darwish, 2013).

ويمكن تصنيف الفقد حسب نوعه كما يلي: الإنقطاع Dropout (متكرر على وتيرة واحدة) وفيه تنسحب بعض المفردات من الدراسة قبل إنتهائها أي أن القيمة المفقودة لا تتبعها أي قيمة ملاحظة، والنمط المتقطع في الفقد intermittent missing data (غير المتكرر على وتيرة واحدة) ويحدث عندما توجد قيمة أو قيم ملاحظة بعد حدوث القيمة المفقودة وهو ما يسمى بالفقد غير المتكرر على وتيرة واحدة non monotone وعندما يكون الفقد في البيانات متقطع فهناك تحد كبير يواجه الإحصائي في عملية النمذجة، وقد يحدث كلا النوعين معا.

ومن الملاحظ أن أسباب فقدان البيانات متعددة، فقد يكون السبب في الفقد وجود عطل في المعدات أو أن الظروف المناخية غير مواتية أو بسبب حدوث أخطاء في إدخال البيانات

(Fitzmaurice, 2008). وقد تعاني الإستبيانات مثلا من فقد البيانات عندما يرفض أحد أفراد العينة الإجابة، أو عندما لا يعرف الإجابة أو عندما يتخطى عنصر من الإستبيان بطريق الخطأ. وتعقد هذه المشكلة التحليل الإحصائي للبيانات (Tshering, et.al., 2013).

عادة يقوم الباحث بإستخدام حل بسيط لمعالجة الفقد هو أن يقوم الباحث بإستبعاد أي حالة فقد في البيانات لأي من المتغيرات في التحليل. ويترك هذا الإجراء مجموعة البيانات دون أي فقد وبالتالي يصبح من الممكن تحليلها بأي من الأساليب التقليدية. وتعرف هذه الإستراتيجية بإسم تحليل الحالة الكاملة (Myers, 2000). وتمتلك هذه الإستراتيجية العديد من الخصائص الجذابة ولكن العيب الرئيسي فيها أنها تستبعد نسبة كبيرة من العينة الأصلية. أي أن إعتماد الباحثين لعدة عقود على تقنيات Ad-hoc يعتبر غير مجدي حيث أنها تتعامل مع البيانات بإهمال الحالات غير الكاملة أو إضافة معلومات عن القيم المفقودة مثل الشطب بطريقة القائمة أو الشطب بالطريقة المزدوجة أو غيرها ويعتمد بعض منها على منع حالات الفقد بتقليل وحدات عدم الإستجابة أو تقليل فقد المتابعة للمريض في الدراسات الطولية ونجد أن الكثير من هذه التقنيات تحتاج إلى فروض حازمة نسبيا حول سبب فقد البيانات وتتعرض للتحيز الحقيقي.

ومن المهم أن نفرق بين أنماط فقد البيانات وآليات فقد البيانات. ويشير مصطلح أنماط فقد البيانات إلى شكل القيم المفقودة والملاحظة داخل فئة البيانات فهى تصف ببساطة موقع الفجوات في البيانات ولا تشرح أسباب فقدها، بينما يصف مصطلح آليات فقد البيانات

العلاقات الممكنة بين المتغيرات المقاسة وإحتمال فقد البيانات. وعلى الرغم من أن آليات فقد البيانات لا تقدم تفسير لسبب فقد البيانات إلا أنها تمثل العلاقات الرياضية بشكل عام بين البيانات والفقد (Enders, 2011). وتلعب آليات الفقد دورا كبيرا في نظرية Rubin لفقد البيانات. وسيتم تناول كل من أنماط فقد البيانات وآليات فقد البيانات حيث يوجد العديد من أنماط فقد البيانات ومنها النمط وحيد المتغير والنمط المتكررعلى وتيرة واحدة وغيرها من الأنماط.

وتلعب آليات الفقد دورا كبيرا في نظرية Rubin(1976) لفقد البيانات حيث وضعت الإطار النظري لمشاكل البيانات المفقودة والذي بقي إستخدامه واسع الإنتشار حتى اليوم فقد قام بتقسيم آليات الفقد كما يلي:

- الفقد كامل العشوائية
- الفقد العشوائي
- الفقد غير العشوائي

وتستند المصطلحات التالية على الإطار النظري الذي وضعه كل من Rubin (1976) and Little and Rubin (2002) حيث تسمح تلك المصطلحات بوضع الشروط الشكلية لآليات فقد البيانات والتي تحدد كيفية تأثير تلك الآليات على الإستدلالات اللاحقة.

ومن الملاحظ أن أداء نماذج تحليل البيانات الطولية يعتمد بشكل كبير على آلية فقد البيانات، لذا فلابد من الإهتمام بآلية فقد البيانات في إختيار التحليل المناسب. وعلى الرغم من أن هذه المصطلحات مستخدمة على نطاق واسع إلا أنها قد تكون غير واضحة. وربما كان جزء من السبب في ذلك هو أن هناك بعض الجوانب الخفية نسبيا والمهمة التي تميز

الآليات المختلفة بعضها عن بعض (Hedeker and Gibbons, 2006).

فإذا كان الفقد مستقل عن كل من البيانات الملاحظة وغير الملاحظة تسمى آلية الفقد أنه كامل العشوائية. وإذا كان الفقد مستقل عن القياسات غير الملاحظة بمعلومية البيانات الملاحظة تسمى آلية الفقد في هذه الحالة بالفقد العشوائي. وإذا كان الفقد يعتمد على القيم المفقودة والملاحظة يطلق على آلية الفقد بأنها غير عشوائية (Rubin, 1976).

تعتمد معالجة الفقد في البيانات على دراسة طبيعة البيانات بشكل جيد من حيث نوعية المتغيرات سواء أكانت مستمرة أو متقطعة، ودراسة نوعية البيانات من حيث كونها طولية أو تصنيفية أو غير ذلك، ودراسة نوعية الفقد من حيث كونه فقد متقطع أو إنقطاع ومن حيث وجود الفقد في التغايرات فقط أم في المخرجات أم في كل منهما، ودراسة آلية الفقد في البيانات من حيث كون الفقد كامل العشوائية أم عشوائي أم غير عشوائي، ومعرفة النموذج الذي يناسب هذه البيانات في حالة عدم وجود فقد، ومعرفة توزيعات المتغيرات وتوزيع الفقد والتوزيع المشترك بينهما.

ويظهر من ذلك التعقيدات التي تواجه الباحث من حيث إلمامه بكل هذه الأمور ولابد أن يتمتع بخلفية إحصائية قوية تمكنه من إختيار النماذج الأفضل في توفيق هذه الأنواع من البيانات في ضوء كل هذه المعطيات، فكلما إختلف أحد هذه المعطيات كلما نتج توفيق جديد للبيانات بنماذج تناسبها ومن الصعب أن نجد نفس التوفيق لنفس نوع البيانات في بحثين مختلفين.

إهتمت معظم الدراسات السابقة بالفقد في الإستجابات عندما تكون التغايرات كاملة

وإهتمت بعض الدراسات بالفقد في التغايرات والإستجابات معا ولم تتوافر الكثير من الدراسات التي تهتم بفقد التغايرات المستمرة الطولية في وجود إستجابات ثنائية كاملة. كما إهتمت معظم الدراسات بالفقد الذي يكون على شكل إنقطاع في حين أن الفقد المتقطع لم يحظى بنفس الإهتمام.

ولم تتلقى البيانات الطولية التي تتضمن الفقد المتقطع دراسات مستفيضة ولذا إهتم هذا البحث بالفقد المتقطع في التغايرات المستمرة الطولية في حالة النماذج الخطية المعممة المختلطة وهو الأمر الذي لم تتم تغطيته في دراسة واحدة في نفس الوقت. وقد إعتمدت الدراسة على إستخدام طريقة تحليل الحالة الكاملة وطريقة التعويض المتعدد والمقارنة بينهما.

وتساهم نتائج الأبحاث التي تتضمن بيانات طولية في التجارب الطبية في علاج المرضى أو تطوير الأجهزة الطبية أو إختيار الإجراءات المناسبة للعلاج. وإذا إحتوت البيانات الخاصة بهذه الدراسات على قيم مفقودة فإن جودة توفيق نماذج مناسبة لها تقل مما ينتج عنه نتائج مضللة خاصة أن عملية الفقد في مثل هذه البيانات متكررة، فمن الممكن أن يمتنع المريض عن الإستجابة أو أن يتوقف عن المشاركة في الدراسة لأي سبب ومن هذه الاسباب الوفاة إلى غير ذلك.

وقد تم إختيار مجال تفتيت حصوات الكلى بالموجات الصدمية Extracorporeal Shock Wave Lithotripsy (ESWL) ليكون مجالا تطبيقيا في هذه الدراسة. تترسب أحيانا بلورات من أملاح مختلفة على السطح الداخلي للكلية أو الحالب أو المثانة، وتكبر هذه البلورات مكونة حصوات الكلية أو الحالب أو المثانة.

وفي حالات عديدة نجد أن حصوات الكلى تنمو بدون أي أعراض، ولكن عندما تستقر في الحالب فإن الأعراض يمكن أن تكون شديدة جدا وتختلف إعتمادا على موقع الحصوة ونموها.

ولا علاقة تربط شدة الألم بحجم الحصوة فقد تكون الحصوة كبيرة ولكنها مستديرة فلا تسبب نفس الألم الذي تسببه حصوة صغيرة ذات حواف حادة.

وستتعرض الدراسة للموجات التصادمية ESWL في تفتيت حصوات الكلى لمجموعة من المرضى الذين تم إختيارهم عشوائيا وتتوافر فيهم الشروط التالية:

أن تكون الحصوة معتمة إشعاعيا. وأن تكون الحصوة ذات حجم $\leq$ ٢٥ مم. لم تتم معالجتها بأي وسيلة أخرى. وأن تكون كل القيم المختبرية الخاصة بالمريض لكل من الكرياتنين ووظائف الكبد والدم ومعاملات التجلط طبيعية.

ولا يتم التعامل مع الحصوات بالموجات التصادمية إذا وجدت عدوى بالمسالك البولية لم تتم معالجتها أو وجد إختناق بالقناة البولية التي سيمر منها فتات الحصوة أو إذا وجد أي عيب خلقي بالمسالك البولية (Sheir et al ,2007).

الطرق المستخدمة

بعد الإلمام بأنواع الفقد وآليات الفقد والطرق المتاحة لمعالجة الفقد أصبح من الضروري للباحث أن يجمع المعلومات الخاصة ببياناته ليستطيع تحديد الطريقة المثلى لمعالجة الفقد

تعتمد الدراسة على تغايرات مستمرة وإستجابة ثنائية والبيانات طولية ولذا فكان النموذج الملائم لهذه البيانات هو النموذج الخطي

المختلط المعمم. وتتضمن التغايرات فقط على الفقد حيث أن الفقد متقطع أي غير متكرر على وتيرة واحدة. ولم يتسنى للباحثه عمل دراسة إستطلاعية لمعرفة آلية الفقد ولذا تستخدم الدراسة الفقد كامل العشوائية والعشوائي وغير العشوائي. ولذا سيتم التعامل مع الفقد في حالتين حالة الفقد القابل للتجاهل وحالة الفقد غير القابل للتجاهل. ولذا فإن مشكلة هذه الدراسة هي أن المتغيرات هي عبارة عن تغايرات مستمرة وإستجابات ثنائية والبيانات إجمالا طولية والفقد في التغايرات فقط وهو غير متكرر على وتيرة واحدة.

ولأن الدراسات الطولية تتعرض للفقد أكثر من غيرها من أنواع الدراسات الأخرى ويكون تقدير المعالم في حالة البيانات المفقودة غير القابلة للتجاهل معقدا وتتطلب الطرق التي تعتمد على الإمكان تحديد التوزيع المشترك وآلية الفقد، وهذا التحديد يمكن تصنيفه إلى ثلاثة أنواع من النماذج هي: النماذج المختارة ونماذج النمط الخليط ونماذج المعلمة المشتركة كما ذكرنا في الفصول السابقة (Little, 1995) فإن المشكلة تزداد تعقيدا.

وقد تظهر العديد من المشكلات عند تقدير المعالم عندما تكون دالة الإمكان معقدة مما ينتج عنه مشاكل في حساب التعظيم وهذه الصعوبات قد تكون تحليلية أو حسابية أو كلا منهما. وقد يحدث ذلك في البيانات المقطوعة والتي تحتوي على بعض البيانات المفقودة أو البيانات من توزيعات مختلطة. في العديد من هذه المشاكل يمكن وضع صيغة مصاحبة للمشكلة الإحصائية بنفس المعالم للبيانات الكاملة والتي يمكن العمل عليها وإيجاد تقديرات الإمكان الأعظم منها في سلوك تحليلي أو

حسابي ثم إستغلال صيغ البيانات الكاملة لحساب الإمكان الأعظم للبيانات الناقصة (Ramadan, 2005).

يتم عادة إجراء التجارب الطولية للتحقيق في التغيرات الفردية للأفراد مع مرور الوقت لإستكشاف تأثير مجموعة من العوامل التي يحتمل أن تؤثر على التغيير. ويحدث إرتباط بين القياسات المتكررة للأفراد بشكل طبيعي. ولعمل نموذج للإرتباط بين الملاحظات وأيضا لدراسة تأثيرات فرد (أو مجموعة) على الإستجابات غالبا ما نستخدم النماذج الخطية المختلطة أو النماذج الخطية المختلطة المعممة. وعندما يكون الفقد في البيانات غير قابل للتجاهل فلابد من عمل نموذج لآلية الفقد للحصول على إستدلالات إحصائية صحيحة (Sinha, et al, 2014).

ستقدم هذه الدراسة طرق لمعالجة الفقد في البيانات بإستخدام التعويض المتعدد في حالة النماذج الخطية المعممة المختلطة وذلك في حالة نماذج الإختيار لتقدير معالم النماذج الخطية المختلطة المعممة في وجود تغايرات مفقودة وآلية فقد قابلة للتجاهل وغير قابلة للتجاهل في حالة نمط الفقد غير المتكرر على وتيرة واحدة أي أن فقد البيانات هو فقد متقطع. وذلك للتغايرات المستمرة والطولية في الحالة التي يكون فيها متغير الإستجابة ملاحظ بالكامل.

طريقتي الإمكان الأعظم وبايز تكونان الأصعب عند تقدير المعالم في الأنماط غير المتكررة على وتيرة واحدة من الفقد لأنه عمليا لا يمكن في جميع الأحوال تحليلها إلى عوامل بسيطة. ففي حالة الفقد العشوائي عندما تطبق القابلية للتجاهل فإنه يمكن إستخدام أدوات البرمجة التقليدية في حالة البيانات غير المتوازنة

لتعطي تقديرات كافية بإستخدام الأمكان الأعظم في حين أن المشكلة تزداد صعوبة في حالة الفقد غير العشوائي.

الفقد في التغايرات النماذج الخطية المختلطة المعممة

تسمى النماذج الخطية المعممة التي لها تأثيرات عشوائية بالنماذج الخطية المعممة المختلطة وهي تعميم للنماذج الخطية المعممة لنموذج التأثيرات العشوائية الطبيعي (Laird and Ware, 1982) المذكور في وتعرف عادة على النحو التالي. فلفرد معين i بـ $j = 1, \ldots\ldots, n_i$ من القياسات المكررة، يتم عمل نموذج للمخرجات y_{ij} على النحو التالي:

$$f(\boldsymbol{y}_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \tau) = exp\left[\tau\left\{\boldsymbol{y}_{ij}\boldsymbol{\theta}(\eta_{ij}) - g\left(\boldsymbol{\theta}(\eta_{ij})\right) + c(\boldsymbol{y}_{ij}, \tau)\right\}\right]$$

حيث

$\boldsymbol{y}_{ij}$:هو متجه من الدرجة $n_i \times 1$.

τ :هي معلمة إنتشار الوزن.

$\theta(.)$: هي دالة الربط.

$\eta_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i$: هي تنبؤ خطي.

$\boldsymbol{\beta}$: متجه معالم الإنحدار غير المعلومة من الدرجة $p \times 1$.

$\mathbf{x}_{ij}^{'}$: هي الصف رقم j في مصفوفة التغايرات الثابتة $\boldsymbol{X}_i$ من الدرجة $n_i \times p$.

$\boldsymbol{z}_{ij}^{'}$: هي الصف رقم j في مصفوفة التغايرات الثابتة $\boldsymbol{Z}_i$ من الدرجة $n_i \times q$ لمتجه التأثيرات العشوائية $\boldsymbol{b_i}$ من الدرجة $q \times 1$.

ويقال عن الربط أنه ربط قانوني عندما تكون $\theta\left(\eta_{ij}\right) = \eta_{ij}$.

وبدون خسارة العمومية سنفترض أن $\tau = \tau_0$ حيث أن τ_0 معلومة كما أن $\tau_0 = 1$ في إنحدار بواسون أو الإنحدار اللوجستي. ولهذا ستكتب $c(y, \tau_0) = c(y)$ و $f\left(\boldsymbol{y_{ij}}|\boldsymbol{\beta}, \boldsymbol{b_i}, \tau\right) = f\left(\boldsymbol{y_{ij}}|\boldsymbol{\beta}, \boldsymbol{b_i}\right)$ في المعادلة السابقة. بالإضافة إلى ذلك، يفترض أن $b_i \sim N_q(0, \boldsymbol{D})$ حيث $\boldsymbol{D}$ هي مصفوفة تغاير مجهولة من الدرجة $q \times q$.

وفيما يلي عرض لطرق التعويض المتعدد والإمكان الأعظم بإستخدام طريقتي مونتكارلو لتعظيم التوقع والإمكان الأعظم المتدرجة بإستخدام تقريب لابلاس وبابيز الكاملة:

التعويض المتعدد

لقد برزت طريقة التعويض المتعدد كأسلوب شائع للتعامل مع مشاكل القيم المفقودة. ويتضمن هذا الأسلوب إنشاء عدة مجموعات من البيانات الكاملة عن طريق ملء قيم البيانات المفقودة. ثم يتم تحليل كل مجموعة من مجموعات البيانات التي تم إستكمالها على أنها مجموعة من البيانات الكاملة. ثم يتم الجمع بين الإستدلالات لمجموعات البيانات في نتيجة واحدة بواسطة المتوسط لفئات البيانات التي تم إكمالها. يوجد العديد من الإبحاث التي وصفت التعويض المتعدد وبعض أشكاله للفقد في التغايرات (Ibrahim, et al., 2012, Little and Rubin, 2002).

وهناك نوعين من تقنيات التعويض المتعدد والتي تسمى التعويض الصحيح أو المناسب والتعويض غير الصحيح او غير المناسب (Nielsen,

(2003. ويستخدم التعويض غير المناسب نموذج تعويض يختلف عن نموذج التحليل في حين أن التعويض المناسب يعتمد فيه نموذج التعويض على نموذج التحليل. فعلى سبيل المثال، التعويض غير المناسب يكون فيه نموذج التعويض هو النموذج الطبيعي الخطي وتحليل البيانات الكاملة كان بإستخدام الإنحدار اللوجستي Little, and Rubin, (2002). ويعطي التعويض غير المناسب تقديرات متحيزة، ولكن التعويض المناسب على الرغم من كونه يتطلب حسابات أكثر إلا أنه يعطي حسابات غير متحيزة في العينات الكبيرة (Rubin, 1987).

والفكرة الأساسية للتعويض هي:

- إنشاء M مجموعة بيانات كاملة.
- الحصول على $\hat{\gamma}^{(m)}$ لمجموعة التعويض رقم m، $m = 1, \ldots\ldots, M$.
- تقدير المعلمة هو $\hat{\gamma} = \sum_{m=1}^{M} \frac{\gamma^{(m)}}{M}$.

ولحساب تقدير التباين نفترض أن تقدير التباين من مجموعة التعويض رقم m والذي يرمز له بالرمز $\widehat{\boldsymbol{V}}^{(m)}$ يتم الحصول عليه من إفتراض أن فئات التعويض هي فئات البيانات الكاملة وحساب تقدير التباين بالطريقة العادية (مثل مقلوب مصفوفة التباين). وبتعريف $\overline{\mathbf{V}}$ لتكون متوسط تقدير التباين حيث $\overline{\mathbf{V}} = \sum_{m=1}^{M} \frac{\widehat{\boldsymbol{V}}^{(m)}}{M}$ والتي هي متوسط (داخل) تباين التعويض. ويعرف تباين التعويض (بين) على أنه:

$$\hat{\beta} = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\gamma}^{(m)} - \hat{\gamma}\right) \times \left(\hat{\gamma}^{(m)} - \hat{\gamma}\right)'$$

وتعمل طريقة التعويض المتعدد كطريقة تعظيم التوقع ولكن طريقة التعويض المتعدد تستبدل تعويض المتوسط الشرطي في خطوة التوقع بسحب مفرد من توزيع التعويض كما يلي:

$$f\left(\mathbf{x}_{mis,i}|\boldsymbol{y}_i, \mathbf{x}_{obs,i}, \boldsymbol{\gamma}, \boldsymbol{b}\right) \propto f(\boldsymbol{y}_i|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)$$
$$f\left(\mathbf{x}_{miss,i}|\boldsymbol{\alpha}, \mathbf{x}_{obs,i}\right)$$

الدراسة التطبيقية

تتكون عينة الدراسة من المرضى الذين تم علاجهم في مركز الكلى والمسالك البولية بجامعة المنصورة في الفترة من مارس ٢٠٠٣ إلى ديسمبر ٢٠٠٦ وقد تم تجميع البيانات لـ ١٠٦ مريض لابد أن يتم تفتيت الحصوات لديهم عن طريق الموجات الصدمية. وقد تم قياس مجموعة من المتغيرات لكل مريض تشمل هذه المتغيرات مجموعة من التحاليل التي تم عملها للمريض قبل التفتيت، وبعد التفتيت مباشرة، وبعد التفتيت بشهر واحد، وذلك لمعرفة نتيجة إجراء عملية تفتيت الحصوة.

وتم تجميع البيانات قبل التفتيت، وبعد التفتيت مباشرة، وبعد التفتيت بشهر واحد حيث أن:

الإستجابة هي $\mathbf{y}$ وهو متغير ثنائي حيث أن القيمة صفر تعني نجاح التفتيت والقيمة ١ تعني فشل التفتيت ويحتاج المريض لإعادة التفتيت مرة أخرى.

والتغايرات هي

$\boldsymbol{z_1}$: هو مؤشر كتلة الجسم وهو متغير مستمر ذو بيانات كاملة.

$\boldsymbol{z_2}$: هو نوع جهاز التفتيت وهو متغير ثنائي ذو بيانات كاملة.

x_1: هو درجة حمضية أو قلوية البول وهو متغير مستمر قيس في ثلاثة مناسبات ويحتوي على فقد.

x_2 : هو الكثافة النوعية للبول وهو متغير مستمر قيس في ثلاثة مناسبات ويحتوي على فقد.

x_3: هو لتحديد نسبة الضررمن التفتيت على الكلية وهو متغير مستمر قيس في ثلاثة مناسبات ويحتوي على فقد.

x_4: هو الكرياتينين في مصل الدم وهو متغير مستمر قيس في ثلاثة مناسبات ويحتوي على فقد.

وبملاحظة التغاير x_1 وجد به ١٦ قيمة مفقودة من (١٠٦×٣=٣١٨ قيمة) أي حوالي ٥% وذلك في ١٣ حالة (١٢.١٥%). وفي التغاير x_2 يوجد به ١٦ قيمة مفقودة من (١٠٦×٣=٣١٨ قيمة) أي حوالي ٥% وذلك في ١٤ حالة (١٣.٢%). وفي التغاير x_3 يوجد به ١٦ قيمة مفقودة من (١٠٦×٣=٣١٨ قيمة) أي حوالي ٥% وذلك في ١٦ حالة (١٥.٠٩%). وفي التغايرات x_4 يوجد به ١٢ قيمة مفقودة من (١٠٦×٣=٣١٨ قيمة) أي حوالي ٣.٨% وذلك في ١٢ حالة (١١.٣%).

النموذج المناسب لهذه البيانات هو النموذج الخطي المختلط المعمم حيث أن $\mathbf{y}$ هو متغير الإستجابة ويتبع توزيع ذو الحدين والتغايرات x_1 و x_2 وx_3 وx_4 وz_1 وz_2 هي التغايرات، ويأخذ النموذج الشكل التالي:

$$\boldsymbol{Logit}(\mathbf{Y}) = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \beta_3\mathbf{X_3} + \beta_4\mathbf{X_4} + \beta_5\mathbf{Z_1} + \beta_6\mathbf{Z_2} + \mathrm{U_1}\sigma_1^2 + \mathrm{U_2}\sigma_2^2$$

وهذا النموذج يستخدم في توفيق البيانات الكاملة التي لا تحتوي على فقد وسيتم توفيق البيانات

بهذا النموذج في عدة حالات هي تحليل الحالة الكاملة والتي تحذف فيها جميع الحالات التي تحتوي على فقد في البيانات أي بإعتبار أن آلية الفقد هي الفقد كامل العشوائية ثم في حالة آلية الفقد العشوائي وذلك بإستخدام طريقتي التعويض وباييز ثم في حالة آلية الفقد غير العشوائي بإستخدام طريقة باييز.

وفي حالة آلية الفقد كامل العشوائية بإستخدام طريقة تحليل الحالة الكاملة وفي هذه الحالة سيتم حذف الحالات التي بها فقد وبالتالي سيقل عدد المرضى إلى ٨٢ مريض بدلا من ١٠٦ مريض وأصبح إجمالي المشاهدات ٢٢١٤ مشاهدة بدلا من ٢٨٦٢ مشاهدة. وتكون تقديرات المعالم كما يلي:

	Estimate
β_0	-12.16214
β_1	-0.38320
β_2	11.46527
β_3	-0.02844
β_4	0.27142
β_5	-0.02280
β_6	0.43905

وكانت المعايير AIC = 251.2 و BIC = 286.2.

ثانيا : حالة آلية الفقد العشوائي بإستخدام طريقة التعويض

وفي هذه الحالة سيتم إستخدام طريقة التعويض للحصول على إستكمال للبيانات المفقودة وبعد أن تصبح البيانات كاملة يتم توفيقها بإستخدام النموذج الخطي المختلط المعمم وفي هذه الحالة يكون عدد المرضى ١٠٦ و إجمالي المشاهدات ٢٨٦٢ مشاهدة. و

نتائج توفيق النموذج بعد إستكمال البيانات المفقودة بإستخدام طريقة التعويض والتي كانت كما يلي:

	Estimate
$\boldsymbol{\beta_0}$	-9.042418
$\boldsymbol{\beta_1}$	-0.393956
$\boldsymbol{\beta_2}$	13.611531
$\boldsymbol{\beta_3}$	0.002407
$\boldsymbol{\beta_4}$	-0.614846
$\boldsymbol{\beta_5}$	-0.047428
$\boldsymbol{\beta_6}$	0.075984

وكانت المعايير AIC = 321.2 و BIC = 358.8.

النتائج

بقسمة قيمة DIC على حجم العينة في كلا الحالتين في حالة تحليل الكاملة هو 4.1 وفي حالة التعويض المتعدد هو 3.1 وهذا يعني أن إستخدام حالة التعويض في معالجة الفقد كانت أفضل من إستخدام تحليل الحالة الكاملة.

المراجع

Enders, C.K. (2011). Missing Not at Random Models for Latent Growth Curve Analyses. Psychological Methods, Vol. 16, No. 1, 1–16.

Fitzmaurice, G. (2008). Missing data: implications for analysis. *Nutrition*, 24, 200-202.

Gad, A.M. and Darwish, N.M. (2013). A Shared Parameter Model for Longitudinal Data with Missing Values. *American Journal of*

Association, Vol. 100, No. 469, 332-346

Little R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Am Stat Assoc*; **90**:1113–1121.

Little, R. J. A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). New York: Wiley.

Ramadan, M. (2005). Extentions of the expectation-maximization (EM) algorithm using a baysian approach. Unpublished M.Sc Dissertation, Benha University, Dept. of Statistic.

Applied Mathematics and Statistics, Vol. 1, No. **2**, 30-35.

Hair, J., Anderson, R., Tatham, R. and Black, W. (1998). *Multivariate Data Analysis*. Upper Saddle River, N.J.: Prentice Hall.

Hedeker, D. and Gibbons R. (2006). *Longitudinal Data Analysis*. John Wily& Sons, Inc., Hoboken, New Jersey.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring A. H. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. Journal of the American Statistical

Tshering, S., Okazaki, T., and Endo, S. (2013). A Method to Identify Missing Data Mechanism in Incomplete Dataset. *IJCSNS International Journal of Computer Science and Network Security*, **13**, No.3.

Wu, L. (2009). *Mixed effects models for complex data*. London: Chapman & Hall

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**,581-592.

Rubin D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, NY, Wiley.

Sheir, K., Elhalwagy, S., Abo-Elghar, M. (2008) Evaluation of a synchronous twin-pulse technique for shock wave lithotripsy: a prospective randomized study of effectiveness and safety in comparison to standard single-pulse technique. B J U I N T E R N A T I O N A L | 101 , 1 4 2 0 – 1 4 2 5