التعويض الجزئي للقيم المفقودة

م. م. هيثم عبد الأمير الوردي*

المستخلص

إذا كانت مشكلة فقدان القيم، في مرحلة جمع البيانات، قد تمت بشكل عشوائي تام، فيمكننا معالجتها بالحذف بسهولة، أما إذا تم الفقدان بشكل عشوائي، فنستطيع غالباً جعله عشوائي تام بالتعويض الجزئي، الذي يستند إلى إحلال قيم مقدرة محل مجموعة من القيم المفقودة وليس جميعها، وبإحدى طرق التعويض الأحادي للسهولة، المتوفرة في معظم الحزم الإحصائية، ثم استخدام الحذف للمعالجة. إن بساطة هذه الخطوات تمكن الباحث غير المتخصص، وهو غالباً ما يكون كذلك، من تطبيقها بمفرده، دون طلب مساعدة، يصعب غالباً الحصول عليها.

Abstract

If the problem of missing values, during the data collection stage, happened completely at random (MCAR), we can simply use the deletion method to solve it. But if the missing happened at random (MAR), we can often make it as MCAR by partial imputation. That is based on replacing estimated values with some of the missing values instead of all of them, using one of the single imputation methods for simplicity, which are available in most of the statistical packages. Then we can use the deletion method as a solution. The simplicity of these steps enables unspecialized researcher to imply them alone, without asking help, which is usually hard to obtain.

1- المقدمة

غالباً ما تتضمن البيانات المجمعة على قيم مفقودة، تحتاج إلى معالجة دقيقة للحصول على أفضل النتائج. لقد تعدّدت طرق المعالجة وفقاً لآليات الفقدان، وأبسطها هي حذف حالات القيم المفقودة، ولكن الحذف يتطلب فرضية الفقدان العشوائي التام لها، فعندما لا تكون البيانات كذلك، يلجأ الباحث إلى طرق التعويض الكلي، ومعظمها تحتاج إلى خبرة لا يستطيع غير المتخصص تطبيقها، مثل طريقتي التعويض المتعدد والإمكان الأعظم، وهذا البحث هو من أجل تقديم طريقة جديدة، بسيطة وفعالة، تمكن الباحثين في مختلف التخصصات من تطبيقها، وتتلخص بتعويض جزئي عن القيم المفقودة بحيث نحصل على حالة فقدان عشوائي تام، وبالتالي المعالجة بالحذف حيث نحصل على تقديرات غير متحيزة للمعلمات. لقد قسّمت البحث على أربعة أقسام، تناولت في مقدمته ما يتعلق بالمشكلة والهدف والفرضية وعينة البحث والدراسات السابقة، وشرحت في الإطار النظري منه آليات فقدان القيم وطرق المعالجة وصيغ الاختبارات، أما القسم الثالث فهو الجانب التطبيقي، إذ استخدمت فيه بيانات ملفين جاهزين ضمن حزمة SPSS v18، واستخدمتُ هذه الحزمة في استخراج النتائج، وأخيراً أدرجتُ الاستنتاجات وما أخطط له من عمل مستقبلي في القسم الرابع .

* معهد الادارة / الرصافة
مقبول للنشر بتأريخ 2015/2/18

1-1 مشكلة البحث وأهميته

غالباً ما تحتوي البيانات المجمعة على قيم مفقودة، ومعالجة الشكل العشوائي التام (Missing Completely at Random) لها يمكن أن يتم ببساطة بإحدى طريقتي الحذف (Deletion)، إذ تتوافر طريقتيه، القائمة (Listwise) والمزدوج (Pairwise)، في معظم الحزم الإحصائية، وستكون المعلمات المقدرة غير متحيزة (Unbiased) [7]. إن تطبيق الحذف مرتبط بفرضية MCAR للبيانات، ولكن عند عدم توافر هذا الفرض، وعندما تكون البيانات قد فقدت بشكل عشوائي (Missing at Random)، وليس عشوائياً تاماً (MCAR) ، يلجأ الباحث إلى استخدام إحدى طرق التعويض (Imputation)، التي تستند إلى إحلال قيم تقديرية محل جميع القيم المفقودة، ومشكلة التقدير تتطلب استخدام الأفضل من بين طرق التعويض للحصول على تقديرات غير متحيزة (Unbiased) وفعالة (Efficient).

وبالفعل نحصل باستخدام طريقة التعويض المتعدد (Multiple imputation)، أو الإمكان الأعظم (Maximum likelihood) [1] على مثل هذه التقديرات، ولكن تطبيق أي منهما يحتاج إلى خبرة عالية في هذا المجال، نادراً ما تتوافر لدى الباحث، فضلاً عن عدم توافر طريقة الإمكان الأعظم في معظم الحزم الإحصائية. من هنا تأتي أهمية هذا البحث في تقديم طريقة التعويض الجزئي (Partial imputation) للقيم المفقودة، التي تستند إلى خطوات بسيطة، تمكن الباحث غير المتخصص من تطبيقها.

1-2 هدف البحث

يهدف البحث إلى تقديم طريقة جديدة في معالجة القيم المفقودة، تستند إلى تحديد المتغير المسبب للمشكلة أولاً، ثم التعويض الجزئي بإحدى طرق التعويض الأحادي (Single imputation) البسيطة لقيمه المفقودة ثانياً، بحيث نحصل على حالة MCAR للفقدان، وأخيراً استخدام الحذف لمعالجة المشكلة.

1-3 فرضية البحث

يمكن الحصول على حالة MCAR لفقدان القيم، الضرورية لتطبيق الحذف، بالتعويض الجزئي للقيم المفقودة من المتغير المسبب للمشكلة.

1-4 عينة البحث

عند تثبيت حزمة SPSS (مثل إصدار 18 الذي أستخدمه) على القرص الصلب للحاسوب، تنسخ معها تلقائياً مجموعة ملفات بيانات جاهزة، منها الملفان telco.sav و telco_missing.sav، اللذان سأستخدمهما في هذا البحث. يحتوي الأول على بيانات كاملة (Complete data) عن أشخاص، ولعينة عشوائية حجمها 1000، أما الثاني فهو نسخة ناقصة من الأول؛ لاحتوائه على قيم مفقودة (Missing values) في عدة متغيرات. البيانات خاصة بشركة اتصالات هاتفية، هدفها تحسين خدمات الشركة. يمثل كل سطر بيانات عن شخص، أي حالة (Case)، أما بيانات كل عمود فهي قيم متغير (Variable values).

1-5 الدراسات السابقة

تستند معظم الدراسات، على حد ما هو متوفر على شبكة المعلومات العالمية (الإنترنت)، وفيما يخص آلية MAR لفقدان القيم، إلى استخدام طرق التعويض الكلي عن القيم المفقودة، وهي تقارن بين تلك الطرق، أما بأسلوب المحاكاة (Simulation) أو بالصيغ الرياضية (Mathematical formulas)، وبالرغم من أن تلك الدراسات لا تستخدم تعبير "التعويض الكلي" (Complete imputation)، إلا أنه الأسلوب المفترض، أما الدراسات عن التعويض الجزئي، فقد اقتصرت على الأسلوب المسمى من قبل Yang and Shoptaw بالتعويض الجزئي المتعدد (Multiple Partial Imputation) [12]، وبموجبه يتم فصل القيم المفقودة على مجموعتين، الأولى خاصة بالقيم المفقودة بشكل متقطع (Intermittent Missing Values)، والثانية خاصة بالقيم المتروكة (Dropouts)، ويتم التعويض عن الأولى فقط عدة مرات، فنحصل على عدة مجموعات من قيم التعويض، وقد طبق أسلوب التعويض الجزئي المتعدد بدراسة طبية عام2011على تجارب سريرية [11] ، استخدمت فيها الحزمة الإحصائية MPI 2.0 الخاصة بذلك .

2- الإطار النظري لفقدان القيم

2-1 آليات فقدان القيم

هناك ثلاث آليات لفقدان القيم، هي:

2-1-1 فقدان عشوائي تام Missing Completely at Random (MCAR)

لنفترض أن Y متغيراً يحتوي على قيم مفقودة، وأن X هو متجه (Vector) من متغيرات لا تحتوي على قيم مفقودة . هنا ، تعد القيم (أو البيانات) مفقودة بشكل عشوائي تام (MCAR) إذا كان احتمال الفقدان في Y لا يعتمد على X أو على Y نفسها [5]. وقد صاغ Allison [1] ذلك بالشكل التالي:

$$\Pr(R=1|X,Y)=\Pr(R=1) \quad (1)$$

حيث R مؤشر استجابة (Response indicator)، يأخذ القيمة 1 في حالة الفقدان في Y، ويأخذ القيمة 0 في حالة المشاهدة في Y؛ فالقيم المفقودة في Y لا علاقة لها بالمشاهدات في X ولا بالقيم المفقودة في Y نفسها. وكمثال، إذا كان احتمال فقدان قيم في الدخل لا يرتبط بالعمر ولا بالدخل نفسه، فلدينا حالة MCAR.

يمكن تعميم الحالة السابقة عند وجود قيم مفقودة في عدة متغيرات في التحليل، فاحتمال الفقدان $\Pr(R=1)$ لأي متغير لا يعتمد على أي متغير آخر في التحليل ولا على القيم المفقودة نفسها [8].

بقي لنا أن نذكر خاصية مهمة لحالة الفقدان العشوائي التام للقيم، وهي أن المعلمات المقدرة غير متحيزة (Unbiased)، أي أن عدم التحيز في التحليل الإحصائي يبقى كذلك بالرغم من فقدان في القيم، كما تشير إلى ذلك مصادر عديدة، منها [5] و [9].

2-1-2 فقدان عشوائي Missing at Random (MAR)

تعد القيم المفقودة في Y مفقودة عشوائياً إذا كان احتمال الفقدان لا يعتمد على Y نفسها، ولكنه يمكن أن يعتمد على متغيرات أخرى في التحليل. وكمثال، إذا كان احتمال فقدان قيم في الدخل لا يعتمد على الدخل نفسه، ولكن يمكن أن يعتمد على العمر. وباستخدام نفس الرموز السابقة، صاغ Allison [1] فرضية MAR ، كالأتي :

$$\Pr(R=1|X,Y)=\Pr(R=1|X) \quad (2)$$

إن MAR يمكن أن تصبح MCAR عندما لا يعتمد احتمال الفقدان في Y على أي من المتغيرات الأخرى في التحليل فضلاً عن عدم اعتماده على Y نفسها. وبتعبير آخر، إذا كانت البيانات MCAR فهي أيضاً MAR.

2-1-3 فقدان غير عشوائي Not Missing at Random (NMAR)

ببساطة، إذا لم تكن القيم المفقودة في Y مفقودة عشوائياً (MAR)، فإنها مفقودة بشكل غير عشوائي (NMAR). هنا، يرتبط احتمال الفقدان في Y على Y نفسها، كما هو موضّح في الصيغة الآتية (لم ترد في المصدر السابق):

$$\Pr(R=1|X,Y)=\Pr(R=1|Y) \quad (3)$$

وكمثال، الأشخاص ذوو الدخل العالي لا يميلون إلى كتابة دخلهم في استمارة جمع البيانات؛ ففقدان القيم في متغير الدخل هنا مرتبط بالدخل نفسه، وبالتالي يكون تقدير متوسط الدخل مثلاً، من بيانات تحتوي على قيم مفقودة، متحيزاً (Biased) بالتأكيد؛ فهناك بيانات مهمة مفقودة بشكل غير عشوائي. وللأسف، لا تتوافر لغاية الآن طريقة فعالة متفق عليها في معالجة هذه المشكلة، ولكن توجد طرق مقترحة، معقدة جداً، منها تلك المقدمة من قبل Heckman [1] و Dunning and Freedman [2]. ولهذا السبب، تستند معظم الحزم الإحصائية، التي تعالج مشكلة فقدان القيم، على فرضية الفقدان العشوائي .

2-2 الوقاية والمعالجة للقيم المفقودة

الوقاية خير من العلاج، وهذه حقيقة وليست مجرد مثل، فالباحث يريد بيانات دقيقة وكاملة من المشمولين بالاستبيان، وهناك عوامل يجب أن تأخذ بالاعتبار قبل البدء بعملية جمع البيانات، منها يتعلق بالاستبانة نفسها من حيث التصميم وعدد الأسئلة وصياغتها، ومنها يتعلق بزمان ومكان جمع البيانات، ومنها يتعلق بآليات جمع البيانات، وغير ذلك من الأمور المهمة، التي يعتبر الالتزام بها في غاية الأهمية.

بعد جمع البيانات، نلاحظ غالباً، رغم إتباعنا طرق الوقاية، نقصاً في البيانات المجمعة، متمثلة بقيم مفقودة في متغير أو أكثر، وقد يكون الفقدان هذا ذي تأثير غير فعال على النتائج، كأن نحصل على تقديرات غير متحيزة (Unbiased) للمعلمات، ولها أقل تباين (فعالة Efficient)، وقد يكون مؤثراً إلى درجة كبيرة ويتطلب علاجاً.

سأعرض باختصار طرق معالجة القيم المفقودة، وتتوفر في [5] مقارنة تفصيلية بينها.

هناك نوعان من المعالجة، هما:

2-2-1 الحذف (Deletion).

تتوفر طريقتان للحذف، تستند الأولى على حذف كامل الحالات التي تحتوي على قيم مفقودة، والبقاء على الحالات التامة فقط، وتسمى بطريقة حذف القائمة (Listwise deletion)، أو طريقة الحالة التامة، وتستند الثانية على حذف القيم المفقودة فقط من الحالات، والتعامل مع ما متوفر من البيانات، وتسمى بطريقة الحذف المزدوج (Pairwise deletion)، أو طريقة الحالة المتوفرة، وتفترض أي من الطريقتين الآلية MCAR لفقدان القيم.

2-2-2 التعويض (Imputation).

ويقصد به إحلال قيم تقديرية محل القيم المفقودة. وهو أما أن يكون أحادياً (Single)، أو متعدداً (Multiple). ومن طرق التعويض الأحادي هي التعويض بالمتوسط، والتعويض بالحالة الأقرب (Hot-Deck)، وتعويض الانحدار، وتعويض EM الذي يستخدم خوارزمية Expectation-Maximization لتقديرالقيمة المفقودة. أما أكثر الطرق استخداماً في التعويض المتعدد فهي خوارزمية MCMC (Markov Chain Monte Carlo) المستندة إلى الانحدار الخطي.وتستند طرق التعويض إلى فرضية MAR لفقدان القيم.

إن التعويض المتعدد أفضل من التعويض الأحادي [8,10]، وطريقة الإمكان الأعظم (Maximum likelihood) أفضل من التعويض المتعدد (Multiple imputation) [1] ، وأي منهما تحتاج إلى خبرة عالية في التطبيق.

2-3 صيغ الاختبارات

سأستخدم اختبار مربع كاي لمعرفة ما إذا كانت البيانات مفقودة بشكل عشوائي تام (MCAR)، وهي فرضية العدم له، وهو متوفر ضمن الحزمة الاحصائية SPSS، ويستند إلى الصيغة الآتية [6]:

$$\chi^2_{MCAR} = \sum_{\text{each unique pattern}} (\text{no. of cases in pattern}) * (MD) \qquad (4)$$

بدرجات حرية محدّدة بالصيغة الآتية :

$$DF_{MCAR} = \sum_{\text{each unique pattern}} (\text{no. of nonmissing variables}) - v$$

عندما:

MD: Mahalanobis D^2 of pattern mean.

v : عدد المتغيرات.

وسأستخدم اختبار t في حالتي الفقدان والمشاهدة للقيم، وباستخدام متغير تأشير (Indicator variable) يقوم بتحديد ما إذا كان متغير في حالة فقدان أو مشاهدة لحالة (Case)، وهو أيضاً متوافر ضمن حزمة SPSS، وفيما يلي صيغته [8]:

$$t_{jk} = \frac{\bar{x}^{P}_{jk} - \bar{x}_{k|\text{ variable } j \text{ is missing}}}{\left(\frac{\hat{\sigma}^{P}_{jk}}{n_{jk}} + \frac{\hat{\sigma}_{k|\text{ variable } j \text{ is missing}}}{n_{kk} - n_{jk}} \right)^{1/2}} \qquad (5)$$

عندما :

k المتغيرات الكمية.

j جميع المتغيرات.

n عدد الحالات.

$$\overline{\mathbf{X}}^{P}=\left[\overline{x}_{Ik}^{P}\right]=\left[\sum_{i} x_{ik}/n_{Ik};\quad i\in I(I,k)\right]$$

$$\hat{\sigma}^{P}=\left[\hat{\sigma}_{Ik}^{P}\right]=\left[\left(\sum_{i}\left(x_{ik}-\overline{x}_{Ik}^{P}\right)^{2}/(n_{Ik}-1)\right)^{1/2};\quad i\in I(I,k)\right]$$

3- الإطار الإجرائي

سأستخدم بيانات الملفين telco.sav و telco_missing.sav، المذكورين في الفقرة 1-4 من هذا البحث، وهي خاصة بشركة اتصالات هاتفية، والبيانات مجمعة من عينة عشوائية حجمها 1000. يحتوي الملف الأول على بيانات كاملة عن أشخاص، أما الثاني فهو نسخة ناقصة من الأول؛ لاحتوائه على قيم مفقودة في عدة متغيرات. كتبت بيانات كل شخص في سطر، وهو يمثل حالة (Case)، أما بيانات الأعمدة فهي قيم متغيرات، ويستند بحثنا إلى ستة متغيرات كمية، كالآتي :

اسم المتغير	المعنى
Tenure	عدد الأشهر بالخدمة
Age	العمر بالسنوات
Address	عدد السنوات في العنوان الحالي
Income	دخل الأسرة بالآلاف
Employ	عدد السنوات مع رب العمل الحالي
Reside	عدد أفراد العائلة

يبيّن الجدول (1) قيمتي الوسط الحسابي والانحراف المعياري لكل من المتغيرات، وللحالات الكاملة للبيانات، من الملف telco.sav ، فضلاً عن عدد القيم الكاملة لكل متغير (1000)، وعدد القيم المتطرفة، ونلاحظ عدم وجود قيم مفقودة؛ لأن هذا الملف يحتوي على بيانات كاملة، وسأستفيد منه لأغراض المقارنة فيما بعد.

جدول (1)
إحصائيات عن كامل البيانات من الملف telco.sav .
Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes	
				Count	Percent	Low	High
tenure	1000	35.53	21.360	0	.0	0	0
age	1000	41.68	12.559	0	.0	0	0
address	1000	11.55	10.087	0	.0	0	13
income	1000	77.5350	107.04416	0	.0	0	93
employ	1000	10.99	10.082	0	.0	0	17
reside	1000	2.33	1.436	0	.0	0	35

سأستخدم بيانات الملف telco_missing.sav لاستخراج الإحصائيات ولنفس المتغيرات، وكما هو موضّح في الجدول (2).

جدول (2)
إحصائيات عن البيانات غير الكاملة من الملف telco_missing.sav .
Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes	
				Count	Percent	Low	High
tenure	968	35.56	21.268	32	3.2	0	0
age	975	41.75	12.573	25	2.5	0	0
address	850	11.47	9.965	150	15.0	0	9
income	821	71.1462	83.14424	179	17.9	0	71
employ	904	11.00	10.113	96	9.6	0	15
reside	966	2.32	1.431	34	3.4	0	33

السؤال المهم هنا هو: هل أن البيانات مفقودة بشكل عشوائي تام (MCAR)؟

يمكن استخدام اختبار مربع كاي [6] مباشرة لهذا الغرض (وهو متوفر في معظم الحزم الإحصائية)، وهو الصيغة (4) في البند 2-3، وتستند فرضية العدم له على أن البيانات مفقودة بشكل عشوائي تام، ونتيجة الاختبار مبيّنة في الجدول رقم (3):

جدول (3)

اختبار الفقدان العشوائي التام للبيانات غير الكاملة

EM Means[a]

Tenure	age	address	income	employ	reside
35.62	41.72	11.44	76.6841	11.02	2.32

a. Little's MCAR test: Chi-Square = 187.284, DF = 107, Sig. = .000

نلاحظ أن البيانات ليست مفقودة بشكل عشوائي تام؛ لأن القيمة المعنوية (Significant value) للاختبار هي أقل من 0.05، وهنا يلجأ الباحث ذو الخبرة إلى استخدام التعويض المتعدد أو الإمكان الأعظم لمعالجة المشكلة، ولكن ما الذي يمكن أن يفعله من لا يمتلك تلك الخبرة في هذه الحالة، ولدينا نسب فقدان 15.0% ، 17.9% ، 9.6% ، في المتغيرات address ، income ، employ على التوالي (أكثر من 10% في متغيرين)؟ إن التعويض الكلي للبيانات بالطرق الأحادية يؤدي إلى تقديرات متحيزة (Biased) للمعلمات، وكذلك إلى مشكلة في الأخطاء المعيارية (Standard errors) للتقديرات [3 , 1]؛ لأننا نتعامل مع قيم تقديرية مفروضة بفعل التعويض.

سنرى فيما بعد كيف أن طريقة التعويض الجزئي للقيم المفقودة، وبإحدى طرق التعويض الأحادي، البسيطة التطبيق (مثل التعويض بالمتوسط)، تجعل البيانات في حالة MCAR، ومن ثم جواز استخدام الحذف، وسنرى أيضاً أن سبب مشكلة عدم الفقدان العشوائي التام للبيانات هو ارتباط فقدان القيم في الدخل (income) بمتغيرات أخرى قيد الدراسة.

سنطبق اختبار t في حالتي الفقدان والمشاهدة للقيم، وهو الصيغة رقم (5) في البند 2-3، إذ يبيّن الجدول (4) قيم الاختبار باستخدام متغير تأشير (Indicator variable) يقوم بتحديد ما إذا كان متغير في حالة فقدان أو مشاهدة لحالة (Case)، وتم أيضاً حساب متوسطات المجموعات الجزئية لمتغير التأشير.

جدول (4)

قيم اختبار t ومتوسطات المجموعات الجزئية في حالتي الفقدان والمشاهدة.

Separate Variance t Tests[a]

	tenure	age	address	income	employ	reside
address						
t	,4	.3	.	3.5	1.4	1.0
# Present	819	832	850	693	766	824
# Missing	149	143	0	128	138	142
Mean(Present)	35.68	41.79	11.47	74.08	11.20	2.34
Mean(Missing)	34.91	41.49	.	55.27	9.86	2.21
income	-5.0	-8.3	-3.9	.	-5.9	3.6
t	793	801	693	821	741	792
# Present	175	174	157	0	163	174
# Missing	33.93	40.01	10.67	71.15	9.91	2.39
Mean(Present)	42.97	49.73	14.97	.	15.93	2.02
Mean(Missing)						
employ	-1.0	-.4	-.7	.5	.	-.3
t	877	881	766	741	.	874
# Present	91	94	84	80	904	92
# Missing	35.34	41.69	11.37	71.50	0	2.31
Mean(Present)	37.70	42.27	12.32	67.91	11.00	2.37
Mean(Missing)						

a. Indicator variables with less than 5% missing are not displayed.

نلاحظ أن متوسط العمر في حالة فقدان الدخل هو 49.73 بينما في حالة عدم الفقدان (المشاهدة) هو 40.01، ونلاحظ أيضاً أن الفقدان في الدخل أثر على متوسط متغيرات أخرى، وكل ذلك مؤشر على أن البيانات ليست مفقودة بشكل عشوائي تام. إنها ليست مفقودة بشكل كامل بسبب الفقدان في متغير الدخل، فلو

استبعدنا هذا المتغير من الدراسة، سنلاحظ أن البيانات أصبحت MCAR، كما هو موضّح في الاختبار الوارد في جدول (5)، حيث نلاحظ أن القيمة المعنوية 0.708، أي لا نرفض فرضية العدم في كون البيانات MCAR، عند مستوى معنوية 0.05.

جدول (5)

اختبار الفقدان العشوائي التام للبيانات بعد استبعاد متغير الدخل.

EM Means[a]

Tenure	age	address	employ	Reside
35.61	41.72	11.45	11.05	2.32

a. Little's MCAR test: Chi-Square = 45.049, DF = 51, Sig. = .708

ولكننا في الواقع لا نريد استبعاد متغير الدخل من دراستنا. ومن جهة أخرى، لو حصلنا على بيانات كاملة للدخل؛ فنتوقع أن نحصل على نتيجة معنوية لاختبار MCAR. وللتأكد من ذلك، سأقوم بإحلال قيم متغير الدخل(income) من الملف الكامل البيانات telco.sav وإحلالها محل البيانات الناقصة للمتغير في الملف telco_missing.sav، وكأنما قمنا بجمع البيانات الناقصة عن متغير الدخل، ويتضح هذا الإجراء من الجدول رقم (6) ؛ إذ لا توجد قيم مفقودة لمتغير الدخل .

جدول (6)

إحصائيات عن البيانات غير الكاملة من الملف telco_missing.sav باستثناء متغير الدخل

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes	
				Count	Percent	Low	High
tenure	968	35.56	21.268	32	3.2	0	0
age	975	41.75	12.573	25	2.5	0	0
address	850	11.47	9.965	150	15.0	0	9
income	1000	77.5350	107.04416	0	0	0	93
employ	904	11.00	10.113	96	9.6	0	15
reside	966	2.32	1.431	34	3.4	0	33

وبإجراء اختبار MCAR نحصل على النتيجة المتوقعة، وهي أن البيانات MCAR، وكما هو موضّح من القيمة المعنوية للاختبار في الجدول (7).

جدول (7)

اختبار الفقدان العشوائي التام للبيانات بعد إحلال بيانات تامة لمتغير الدخل فقط

EM Means[a]

tenure	age	address	income	Employ	reside
35.62	41.72	11.45	77.5350	11.02	2.32

a. Little's MCAR test: Chi-Square = 58.411, DF = 67, Sig. = .764

لقد قمنا في الواقع بالتعويض الجزئي عن القيم المفقودة؛ لأننا أحللنا قيماً محل القيم المفقودة لمتغير الدخل فقط، بعد تشخيصنا له بكونه متغير المشكلة في دراستنا، وتركنا القيم المفقودة لبقية المتغيرات كما هي، وتحصلنا نتيجة لذلك على نتيجة معنوية لاختبار MCAR. لقد كان مجموع القيم المفقودة للملف tele_missing.sav وللمتغيرات الستة قيد الدراسة هو 516 (32+25+150+179+96+34)، وبعد الإحلال أصبح المجموع 337، فالإحلال شمل 179 قيمة مفقودة خاصة بمتغير الدخل حصراً.

نستطيع أن نذهب أبعد من ذلك، ونقوم بالتعويض الجزئي للقيم المفقودة لمتغير الدخل نفسه ؛ فنستطيع مثلاً تعويض مجموعة من القيم المفقودة فقط بطريقة عشوائية وباستخدام إحدى طرق التعويض الأحادي البسيطة، مثل طريقة المتوسط أو طريقة الحالة الأقرب أو غيرهما، وسنحصل على نتيجة معنوية لاختبار MCAR ، وكما هو موضّح في الجدول (8) الخاص بالإحصائيات، والجدول (9) الخاص باختبار MCAR.

جدول (8)
إحصائيات عن البيانات من الملف telco_missing.sav بعد التعويض الجزئي للقيم المفقودة لمتغير الدخل

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
tenure	968	35.56	21.268	32	3.2	0	0
age	975	41.75	12.573	25	2.5	0	0
address	850	11.47	9.965	150	15.0	0	9
income	867	72.1326	81.84437	133	13.3	0	64
employ	904	11.00	10.113	96	9.6	0	15
reside	966	2.32	1.431	34	3.4	0	33

جدول رقم (9)
اختبار الفقدان العشوائي التام للبيانات بعد التعويض الجزئي للقيم المفقودة لمتغير الدخل.

EM Means[a]

tenure	age	address	income	employ	reside
35.62	41.72	11.44	74.3177	11.03	2.32

a. Little's MCAR test: Chi-Square = 120.836, DF = 107, Sig. = .170

نلاحظ من جدول رقم (8) أن التعويض شمل 46 (179 - 133 = 46) قيمة مفقودة فقط من مجموع القيم المفقودة لمتغير الدخل (179). وبتعبير أعم، إن التعويض شمل 46 قيمة مفقودة فقط من المجموع الكلي للقيم المفقودة للمتغيرات الستة والبالغ 516 قيمة، أي أن نسبة تعويض بمقدار 9% تقريباً كانت كافية لجعل البيانات MCAR.

في الواقع العملي، غالباً ما نحاول جمع بيانات حقيقية من المستجيبين تحل محل البيانات الناقصة لمتغير الدخل؛ فمن 179 شخص قد نحصل على 46 (9%) قيمة فعلية، تجعل البيانات MCAR، وهذا بالتأكيد أفضل من التعويض بقيم تقديرية.

4- الاستنتاجات وعمل مستقبلي (Conclusions and Future Work)

4 - 1 الاستنتاجات (Conclusions)

عند ملاحظة وجود قيم مفقودة بعد جمع البيانات، علينا أولاً محاولة جمع بيانات حقيقية من المستجيبين تحل محل البيانات الناقصة؛ لأن ذلك أفضل من التعويض بقيم تقديرية، وقد يكفي ما نحصل عليه لجعل البيانات MCAR، إن لم تكن كذلك، وبالتالي نستطيع استخدام الحذف للمعالجة، وهي طريقة في متناول الجميع، ومتوفرة في الحزم الإحصائية، أما عند تعذر ذلك، وعندما لا تكون البيانات MCAR، وفقاً للاختبار، فيمكننا استخدام التعويض الجزئي لجعلها كذلك، وذلك بإحلال قيم مقدرة محل مجموعة من القيم المفقودة وليس جميعها، وبإحدى طرق التعويض الأحادي، السهلة التطبيق، والمتوافرة في معظم الحزم الإحصائية، ثم استخدام الحذف للمعالجة.

إجراء التعويض الجزئي بسيط، واختبار MCAR أيضاً، وكذلك استخدام الحذف للمعالجة، وكل ذلك ممكن التطبيق من قبل الباحث غير المتخصص بمفرده، دون مساعدة خبير في الإحصاء، أما طرق التعويض المتعدد والإمكان الأعظم، فتحتاج إلى خبرة في التطبيق، بدونها لن يحصل الباحث إلا على نتائج غير دقيقة، وإن توفرت هذه الطرق المتقدمة في الحزمة الإحصائية.

4 - 2 عمل مستقبلي (Future Work)

لاحظنا من جدول رقم (8) أن التعويض شمل 46 قيمة مفقودة فقط من مجموع القيم المفقودة لمتغير الدخل (179)، ومن المجموع الكلي للقيم المفقودة للمتغيرات الستة والبالغ 516 قيمة. وبحساب النسبة فإنها تشكل نسبة تعويض 9% تقريباً، وكانت كافية لجعل البيانات MCAR. إن هذه النسبة لا تشكل الحد الأدنى المطلوب من القيم التعويضية، فقد استندت إلى التجربة كما لاحظنا، وليس إلى صيغة إحصائية محدّدة؛ لعدم توفر مثل هذه الصيغة لغاية الآن.

إن العمل على إيجاد صيغة أو معيار نستند إليه كدليل لمعرفة الحد الأدنى لنسبة القيم التعويضية، التي تجعل البيانات MCAR، والذي سيأخذ بالاعتبار طبيعة البيانات الخاضعة للتحليل، يحتاج بالتأكيد إلى مجهود كبير ووقت طويل، وهو ما أنوي القيام به مستقبلاً، فضلاً عن تطبيق طريقة التعويض الجزئي للقيم المفقودة، المقترحة في هذا البحث، على أحجام صغيرة ومتوسطة من البيانات؛ للتأكد من كفاءتها.

المصادر (References)

1- Allison, P.D. (2012)'Handling Missing Data by Maximum Likelihood', SAS Institute, Statistical Horizons, Haverford, PA, USA.
2- Dunning, T. and Freedman, D.A. (2008)'Modeling section effects', London, UK.
3- Haitovsky, Y. (1968) 'Missing data in regression analysis', Journal of the Royal Statistical Society, Series B 30: 67–82.
4- Heckman, J. J. (1979) 'Sample selection bias as a specification error', Econometric 47: 153–161.
5- Little, R. J. A. and Rubin, D. B. (1987) 'Statistical Analysis with Missing Data', New York, Wiley.
6- Little, R.J.A. (1988) 'A Test of Missing Completely at Random for Multivariate Data with Missing Values', Journal of the American Statistical Association83.
7- Little, R.J.A. (1992)'Regression with missing X's: a review', Journal of the American Statistical Association 87.
8- Rubin, D. B. (1987) 'Multiple Imputation for Non-response in Surveys', New York, Wiley.
9- Schafer, J. L. (1997) 'Analysis of Incomplete Multivariate Data', London, Chapman and Hall.
10- Scheffer, J. (2002) 'Dealing with Missing Data', Available online at http://www.massey.ac.nz/~wwiims/research/letters/.
11- Xiaowei Yang, Jinhui Li, and Steven Shoptaw (2011)' Multiple Partial Imputation for Longitudinal Data with Missing Values in Clinical Trials', University of California, https://escholarship.org/uc/item/9733x421.
12- Yang X, Shoptaw S. (2005) 'Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial', *Drug and Alcohol Dependence*; 77: 213-225

.....

.....

.....