طريقة التعويض الكسرية
المعملية لعلاج الفقد بالبيانات الطولية*

عبدالله سليمان**

تعرف الدراسات الطولية بأنها تلك الدراسات التى تهتم بدراسة التغيرات التى تطرأ على ظاهرة من الظواهر عبر فترة زمنية معينة، وبذلك لا تقتصر على وصف الوضع الحالى للظاهرة، بل تدرس الظاهرة فى فترة ما ثم تتابع دراستها لمعرفة التغيرات التى تمر بها مع الزمن والعوامل التى تسبب هذه التغيرات، ولعل هذا هو الفارق الأساسى بين الدراسات الطولية والدراسات المسحية التى يتم فيها دراسة الظاهرة عند لحظة زمنية معينة. والدراسات الطولية يمكن إجراؤها فى مختلف الميادين، فالمدرس حين يتابع سلوك تلميذ ما عبر مرحلة من الزمن فإنه يقوم بدراسة طولية، وكذلك الطبيب الذى يراقب مريضه عبر فترة زمنية معينة. ويتم تطبيق الدراسات الطولية وفقا للخطوات التالية:

١- ملاحظة ظاهرة أو موقف فى فترة ما من الزمن، ووصف هذه الظاهرة كما هى فى ذلك الوقت.

٢- متابعة هذه الظاهرة بعد مرور فترة من الزمن، ووضعها فى ضوء واقعها الجديد والتغيرات التى تمر بها والعوامل التى أدت إلى حدوث هذه التغيرات.

* ملخص رسالة الماجستير، قسم الإحصاء، كلية الاقتصاد والعلوم السياسية، جامعة القاهرة، ٢٠١٤.

** مدرس مساعد ، المركز القومى للبحوث الاجتماعية والجنائية.

المجلة الاجتماعية القومية، المجلد الحادى والخمسون، العدد الثانى، مايو ٢٠١٤.

٣- متابعة دراسة الظاهرة بعد فترات زمنية أخرى، وتحديد العوامل التى أدت إلى تشكيلها فى آخر صورة لها.

وبهذا نجد أن الدراسة الطولية تمتاز عن الدراسة المسحية بعدة مزايا أهمها:

١- أنها أكثر دقة لأنها تُجرى على مجموعة واحدة فقط وتتم متابعة هذه المجموعة نفسها فى فترات زمنية متتابعة.

٢- يمكن أن يتم إجرائها على عينة صغيرة نسبيا مقارنة بالدراسات المسحية.

٣- فى الدراسات الطولية يمكن للمجموعة الضابطة أن تكون هى نفسها المجموعة التجريبية، فعلى سبيل المثال إذا ما أراد طبيب معرفة تأثير عقار ما فإنه يمكن أن يقوم بقياس المتغير محل الاهتمام للعينة قبل إعطاء العقار وبعد إعطاؤه، بينما فى الدراسات المسحية يتم تسجيل قياسات المجموعة الضابطة ومقارنتها بقياسات المجموعة التجريبية، ولا شك أن المجموعة الضابطة والتجريبية فى هذه الحالة يجب أن يكونا متماثلين بالنسبة لجميع العوامل التى يمكن أن تؤثر على المتغير محل الدراسة مما يمثل عبء إضافى على مصمم الدراسة.

ولكن يعاب على الدراسات الطولية أن المشاهدات عرضة للفقد أكثر من الدراسات المسحية وذلك لتعدد مرات تتبع المشاهدات مما يزيد من احتمال فقد مفردة أثناء فترة الدراسة وما يتبعه ذلك من آثار سلبية على نتائج الدراسة. ونظرًا لطبيعة الدراسات الطولية فإن الفقد فى الدراسات الطولية قد يأخذ عدة أشكال منها:

١- الفقد المستمر: وهذا النوع يحدث عند انسحاب المفردة من الدراسة بشكل نهائى فلا يتم تسجيل مشاهدات للمفردة بعد أول قيمة مفقودة لها.

٢- الفقد المتقطع: وهذا النوع ينتج من تغيب المفردة لفترة زمنية أثناء فترة الدراسة وعودتها بعد ذلك، لذا فإن القيم المفقودة تتبع بقيم مشاهدة.

وقد تشمل الدراسة أحد نوعى الفقد أو كلاهما مما يزيد من تعقيد مشكلة الفقد وضرورة استخدام طرق إحصائية متقدمة للحصول على تقديرات غير متحيزة واستنتاجات صحيحة من عينة الدراسة. وقد يحدث الفقد لأسباب ليس لها علاقة

بالدراسة فيكون فقدًا عشوائيًا وقد يكون متأثرًا بقيم المتغير محل الدراسة فيكون فقدًا غير عشوائيًا، لذا فإنه من الأهمية بمكان تقسيم الفقد وفقًا لعلاقته بالمتغير محل الدراسة إلى:

١- الفقد العشوائى التام: وهذا يحدث إذا كانت أسباب الفقد ليس لها علاقة بالمتغير محل الدراسة.

٢- الفقد العشوائى: وهذا النوع ينتج إذا كانت أسباب الفقد مستقلة عن المتغير محل الدراسة لكنها مرتبطة بخصائص مفردات العينة.

٣- الفقد غير العشوائى: وهذا النوع ينتج من كون حدوث الفقد مرتبط بقيم المتغير محل الدراسة.

ولتوضيح الفارق بين أنواع الفقد الثلاثة نفرض وجود فقد في دراسة تهتم بمعرفة درجة تفضيل الأفراد للأحزاب السياسية القائمة، فإن كان الفقد ناتجًا عن امتناع المبحوث عن الإجابة لضيق وقته أو عدم اهتمامه فهو فقد عشوائى تام، وإن كان الفقد مرتبطًا بفئة عمرية أو منطقة سكنية معينة كأن يكون الفقد متركزًا فى الفئة العمرية الأكبر أو متركزًا فى الحضر دون الريف فيعد فقدًا عشوائيًا، أما إن كان الفقد ناتجًا من توجه المبحوث لحزب لا يحظى بتأييد من غالبية المجتمع فيخشى أو يخجل من إظهار تأييده لهذا الحزب فإن الفقد يكون غير عشوائى. وفى حالة الفقد العشوائى والعشوائى التام يمكن تجاهل الفقد واستخدام الطرق الإحصائية التقليدية للحصول على تقديرات غير متحيزة بالاعتماد على القيم المشاهدة فقط، أما في حالة الفقد غير العشوائى فينبغى أخذ الفقد فى الاعتبار للحصول على استنتاجات صحيحة لمجتمع الدراسة، إذ أن تجاهل القيم المفقودة قد يؤدى لنتائج شديدة التحيز لا تعبر عن المجتمع محل الدراسة، لذا فإن الفقد غير العشوائى يعد أكثر أنواع الفقد أهمية وأشدها ضررًا على نتائج الدراسة.

ومن هنا كانت الحاجة إلى استخدام طرق إحصائية متقدمة تأخذ فى الاعتبار الفقد الحادث في البيانات- خاصة إذا كان الفقد غير عشوائى- وتقوم بتعويض

وحساب القيم المفقودة أثناء عملية تقدير المعلمات والوصول إلى استنتاجات للمجتمع محل الدراسة.

الدراسات السابقة

تعد مشكلة الفقد مشكلة تاريخية تناولتها الكثير من الدراسات وتم تقديم العديد من الاقتراحات لعلاجها تباينت من البساطة إلى التعقيد وفقا للسبل والإمكانات المتاحة لكل زمن، فكانت الطرق الكلاسيكية تُستخدم قديمًا لبساطتها وعدم احتياجها إلى حسابات معقدة إلا أنها لم تصلح لعلاج الفقد إلا في حالة الفقد العشوائى التام أو الفقد العشوائى فقط، بينما جاءت العديد من الطرق الحديثة لحل مشكلة الفقد المستمر والمتقطع بغض النظر عن نوع الفقد ومدى ارتباطه بالمتغير محل الدراسة.

وبعد مراجعة التراث العلمى حول علاج مشكلة الفقد للدراسات الطولية، تم تقسيم طرق علاج الفقد إلى:

١- طرق الحذف: وهذه الطرق تعد من أقدم الطرق المستخدمة لعلاج مشكلة الفقد وأشدها بساطة إذ أنها تعتمد على حذف المفردات التى لها بعض القيم المفقودة والاكتفاء بالمفردات كاملة الاستجابة.

٢- طرق الأوزان: وتعتمد هذه الطرق على تحديد المفردات كاملة الاستجابة الأقرب تشابهًا مع المفردات التى لها بعض القيم المفقودة، ثم إعادة حساب أوزان العينة بحيث تكون هذه المفردات ممثلة عن نفسها وعن المفردات ناقصة الاستجابة المشابهة لها.

٣- طرق التعويض: وتعتمد هذه الطرق على تعويض القيم المفقودة بقيم أخرى محسوبة ومن ثم يتم إكمال البيانات المفقودة بالقيم المحسوبة ويتم استخدام طرق التحليل التقليدية للحصول على المقدرات وإجراء اختبارات الفروض المطلوبة.

وتنقسم طرق التعويض إلى طرق تعويض أحادية ومتعددة وكسرية. تعتمد طرق التعويض الأحادية على استبدال كل قيمة مفقودة بقيمة واحدة محسوبة، بينما تقوم طرق التعويض المتعددة باستبدال كل قيمة مفقودة بعدة قيم محسوبة

وبالتالى يكون هناك أكثر من مجموعة بيانات كاملة، فيتم إجراء التحليل الإحصائى لكل مجموعة على حدة للحصول على عدة مقدرات لكل معلمة مراد تقديرها، ثم يتم دمج كل المقدرات فى مقدر واحد يتم استخدامه كمقدر نهائى. أما بالنسبة لطرق التعويض الكسرية فهى تقوم على استبدال كل قيمة مفقودة بعدة قيم محسوبة مع إعطاء قيم كسرية – تُسمى بالوزن الكسرى– لكل قيمة محسوبة، ثم يتم إجراء التحليل الإحصائى مرة واحدة اعتمادًا على جميع القيم المحسوبة والأوزان الكسرية المعطاة.

٤- طرق الإمكان الأكبر: وتعتمد هذه الطرق على نمذجة وتقدير نموذج البيانات المفقودة ومن ثم يتم إكمال البيانات المفقودة عن طريق سحب قيم لها من التوزيع المقدر، لذا فإن طرق الإمكان الأكبر هى فى واقع الأمر أحد انواع طرق التعويض.

أهمية الدراسة

تتمثل أهمية الدراسة فى وجود حاجة ماسة إلى الحصول على مقدرات غير متحيزة فى ظل وجود الفقد غير العشوائي خاصة مع قلة الأساليب المتاحة التى يمكن أن تعالج هذا النوع من الفقد بالإضافة إلى ما تطلبه هذه الأساليب من حسابات معقدة و وقت مبذول وتكلفة كبيرة، وعلاوة على ذلك لا تضمن كثير من هذه الأساليب حدوث تقارب لأقرب قيمة لتقدير الإمكان الأكبر، وبناء على ذلك تسعى الدراسة الحالية إلى إيجاد طريقة يمكن الاعتماد عليها فى الحصول على مقدرات غير متحيزة فى ظل وجود فقد مستمر غير عشوائى فى الدراسات الطولية وتستطيع أيضا معالجة بعض عيوب الطرق القائمة مثل الحسابات المعقدة وعدم حدوث تقارب لأقرب قيمة لتقدير الإمكان الأكبر.

تقسيم الدراسة

فى إطار تحقيق أهداف الدراسة تم تقسيمها إلى خمسة فصول، تناول الفصل الأول منها مقدمة عن الدراسات الطولية وتوضيح الفارق بينها وبين الدراسات المسحية، كما تطرق إلى طرق نمذجة البيانات الطولية وكيفية تقدير معلمات النموذج فى ظل ارتباط مشاهدات الفرد الواحد بالدراسة عبر الزمن واستقلالها عن مشاهدات باقى أفراد الدراسة. وتناول أيضا طرق نمذجة البيانات المفقودة التى تعانى من مشكلة الفقد وأنواع هذه النماذج ومدى صلاحية كل نموذج، كما تم عرض أحد أهم طرق الإمكان الأكبر المستخدمة لعلاج مشكلة الفقد وهى طريقة Expectation- Maximization Algorithm أو EM Algorithm التى تقوم على حساب مقدرات معلمات النموذج باستخدام التكاملات، وأخيرًا تم عرض الأشكال العشوائية من EM algorithm مثل أسلوب Monte Carlo EM Algorithm ،Stochastic EM Algorithm ،Stochastic Approximation EM و Parametric Fractional Imputation، وجميع هذه الطرق تقوم بتقريب التكامل باستخدام طريقة Monte Carlo لتجنب صعوبة حساب تكاملات أسلوب EM Algorithm.

وتناول الفصل الثانى طرق التعويض المختلفة بأنواعها الثلاث مع إبراز مزايا وعيوب كل نوع، حيث تمتاز طرق التعويض الأحادية بالبساطة لكنها لا تصلح للحصول على مقدرات غير متحيزة إلا فى حالة الفقد العشوائى والعشوائى التام، بينما تمتاز طرق التعويض المتعددة بأنها تأخذ فى الاعتبار حقيقة أن القيم المحسوبة قيم صناعية وليست واقعية ولكن يعاب عليها تكرار عملية التحليل الإحصائى بعدد مرات تعويض كل قيمة مفقودة مما يزيد من صعوبة وتعقيد الحسابات وهذا ما تجنبته طرق التعويض الكسرية.

وفى الفصل الثالث تم عرض أسلوب Parametric Fractional (PFI) أو طريقة التعويض الكسرية المعلمية، وهى إحدى الأشكال العشوائية من أسلوب EM Algorithm وتقوم على تطوير أسلوب Monte Carlo EM Algorithm باستخدام الأوزان الكسرية بحيث لا يتم تكرار سحب القيم فى كل دورة للأسلوب

بالإضافة إلى أن طريقة PFI تضمن حدوث تقارب لأقرب قيمة لتقدير الإمكان الأكبر وهذا ما لا يحققه أسلوب Monte Carlo EM Algorithm، كما تم عرض تطبيق طريقة PFI فى الدراسات المسحية والطولية لأنواع الفقد العشوائى التام والعشوائى وغير العشوائى، وتم اقتراح تطبيق طريقة PFI للدراسات الطولية في حالة الفقد المستمر غير العشوائى باستخدام نموذج Selection Model، وأخيرًا تم تقديم طريقة حساب الانحرافات المعيارية للمقدرات الناتجة من أسلوب PFI باستخدام طريقة Jackknife Replication وذلك لاستخدامها فى اختبارات فروض معنوية المقدرات الناتجة.

كما تناول الفصل الرابع تطبيق طريقة PFI على بيانات واقعية لمجموعة من الأبقار وأيضا على بيانات صناعية تم توليدها من التوزيع المعتاد المتعدد، وأخيرا تم عرض النتائج والتوصيات فى الفصل الخامس.

عينة الدراسة

تم تطبيق طريقة PFI على بيانات خاصة بمجموعة **من الأبقار** يبلغ عددهم ١٠٧ بقرة تم تتبع كمية اللبن الناتجة من كل بقرة لمدة عامين، وتهدف هذه الدراسة إلى مقارنة متوسط كمية اللبن **خلال سنتى الدراسة بغية معرفة أيهما أكبر**، إلا أنه بعد **السنة الأولى من الدراسة** أصيب ٢٥ بقرة **بمرض أدى لانخفاض كمية اللبن** المنتجة، لذا تم استبعاد هذه الأبقار من السنة الثانية للدراسة وتم اعتبار قيمهم مفقودة.

وتمتاز هذه البيانات بأنه تم تحليلها مسبقًا فى دراسات سابقة باستخدام أسلوب SEM Algorithm وأسلوب Nelder Mead Simplex Algorithm– أحد طرق الإمكان الأكبر الكلاسيكية التى تحتاج إلى حسابات شديدة التعقيد– مما أمكن مقارنة النتائج المتحصل عليها بواسطة أسلوب PFI مع النتائج السابقة. وقد أشارت نتائج الدراسة إلى أن الفقد فى البيانات فقدًا غير عشوائى، كما أن تقارب قيم مقدرات طريقة PFI وSEM algorithm وNelder Mead Simplex Algorithm تدل على أن

طريقة PFI سلكت مسلكًا مشابهًا للطرق السابقة مما يدل على صلاحيتها للتطبيق فى حالة الفقد المستمر غير العشوائى.

وأيضا تم تطبيق طريقة PFI على ثلاثة مجموعات من البيانات الصناعية قوامهم ثلاثون وخمسون ومائة مفردة على الترتيب مولدة من التوزيع المعتاد المتعدد بعد أن تم صناعة فقدًا مستمرًا غير عشوائى بنسبة ١٥٪ و٤٥٪ من إجمالى البيانات، وكان أهم ما يلاحظ على النتائج أن طريقة PFI تمكنت من الوصول إلى قيم مشابهة للقيم الحقيقية لمعلمات نموذج التوزيع المعتاد المتعدد المستخدم فى توليد البيانات مما يشير أيضا إلى جودة الطريقة وصلاحيتها للتطبيق فى حالة الفقد المستمر غير العشوائى.

أهم النتائج

بناء على ما سبق فإن أهم ما يمكن استخلاصه من هذه الدراسة يمكن إيجازه كما يلى:

١- استخدام طرق التحليل التقليدية لتحليل بيانات تعاني من الفقد غير العشوائى يؤدى إلى مقدرات متحيزة واستنتاجات خاطئة لا تعبر عن الصورة الحقيقية للمجتمع.

٢- عدم صلاحية طرق معالجة الفقد التقليدية فى حالة الفقد غير العشوائى.

٣- جودة طريقة PFI وصلاحيتها للتطبيق فى حالة الفقد المستمر غير العشوائى وقدرتها على الوصول إلى مقدرات غير متحيزة حتى فى حالة معدلات الفقد المستمر غير العشوائى المرتفعة.

٤- أفضلية طريقة PFI على العديد من طرق الإمكان الأكبر القائمة من حيث تسهيل الحسابات وسرعة التقارب.