ВЪЗМОЖНОСТИ ЗА ИЗПОЛЗВАНЕ НА МНОЖЕСТВЕНО ВЪВЕЖДАНЕ ПРИ ИГНОРИРУЕМИ МЕХАНИЗМИ

гл. ас. д-р Деян Лазаров
Бургаски свободен университет

USAGE POSSIBILITIES OF MULTIPLE IMPUTATION UNDER IGNORABLE MECHANISMS

Assis. Prof. Deyan Lazarov, PhD
Burgas Free University

Резюме: *В настоящото изследване се прави кратка демонстрация на възможностите на множественото въвеждане. За целта се симулира база от данни, в която признаците образуват многомерно нормално разпределение. Изкуствено в една от променливите са елиминирани 50% от стойностите, следвайки механизма липсващи напълно случайно стойности. Върху базата с липсващи стойности е приложено множествено въвеждане и получените резултати са в контекста на „подходящото въвеждане“ на Доналд Рубин.*

Ключови думи: *липсващи стойности, множествено въвеждане*

Abstract: *The present research demonstrates the possibilities of multiple imputation as method for dealing with missing data. For this reason a data base is simulated, in which the variables form multivariate normal distribution. For the needs of the research in one of the variables artificially are implement missing values, which covers 50% of the values. Multiple imputation is conducted over data base with missing data and the derived results are in context of the "proper imputation" of Donald Rubin.*

Key words*: missing values, multiple imputation*

Увод

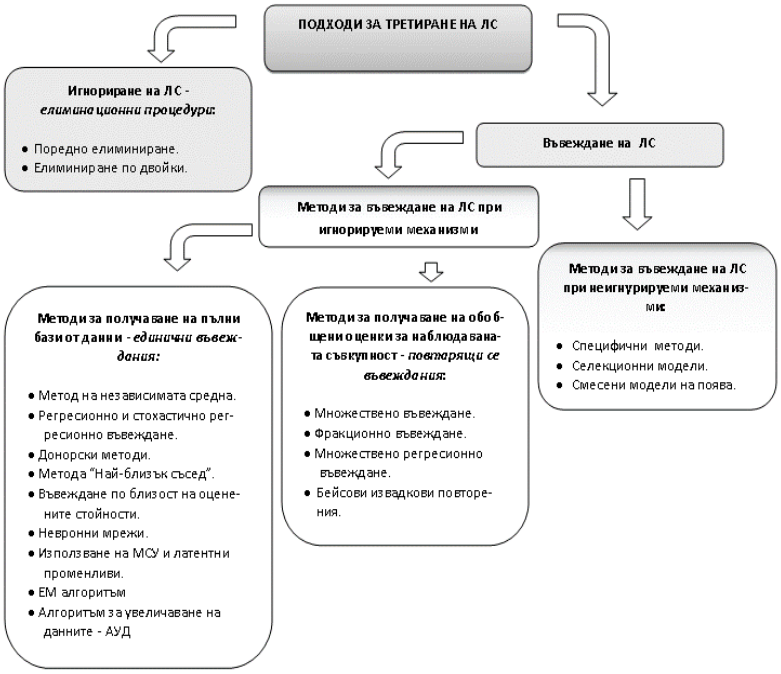
Липсващите стойности (ЛС) са основен проблем в емпиричните изследвания. Те са появяват независимо от това кой и къде провежда изследването. Такива се наблюдават както при репрезентативни изследвания на НСИ[172], така и при индивидуалните изследвания на всеки отделен корпоративен или частен изследовател. Удовлетворяващо е, че проблемът е във фокуса на теоретиците и практиците по света през последните 20 години. Той се оценя като сериозен и предизвиква дискусии и публикации. Например Европейската икономическата комисия към ООН провежда регулярни срещи по въпросите на редактирането на данни и въвеждането на ЛС, като последната е през 2012 г.[173] Публикациите в рамките на тези срещи дават основание да се заключи, че появата на ЛС е повсеместен проблем, независимо от държавата, организацията и финансирането на конкретно изследване. Усилията той да се преодолее е ангажирал умовете и ресурсите на много изследователи и институции.

Подходи за преодоляване на проблема

[172] виж. Изданието на НСИ, „Заетост и безработица – годишни данни“ 2007, 2008, 2009, 2010, 2011, 2012 г.

[173] За повече инфомация виж. http://www.unece.org/index.php?id=3214

В теорията и практиката са разработени редица методи за решаване на проблемите произтичащи от ЛС. Въпреки това е важно изборът на метод да бъде подчинен на механизма на поява на самите ЛС [Лазаров, Д (2014) (дисертационен труд), Rubin (1987)]. На фиг. 1 схематично е представена връзката между възможните подходи относно ЛС – дали те се игнорират в даден статистически анализ или е избран подход за тяхното въвеждане. При избор на изследователя да въведе ЛС то във всички случаи е наложително да се провери механизма на тяхната поява. В случай, че се потвърди хипотезата, че механизма е игнорируем изборът на метод/и се определя от целите на изследователя, дали въвеждането на ЛС се прави за да се получи една пълна база от данни, която да бъде предоставена на други потребители или се цели провеждането на конкретни анализи и получаването на обобщени оценки за изследваната съвкупност. В случай, че хипотезата за игнорируеми механизми бъде отхвърлена, изследователят трябва да подбере някой от съществуващите методи за анализ при неигнорируеми механизми. Тези методи могат да бъдат съчетани с методи при игнорируеми механизми, но едва след като проведения анализ на базата от данни обособи групи от единици, в които появата на ЛС е със случаен характер. В настоящото изследване фокус е поставен върху множественото въвеждане.



Фигура 1. Схематично представяне на подходите за третиране на ЛС и методите, които биха могли да бъдат използвани в зависимост от механизмите на ЛС и целите на анализа

Множествено въвеждане

Най-лансирания метод за анализ на бази данни с ЛС при игнорируеми механизми е множественото въвеждане (МВ). Той компенсира основните недостатъци на единичните въвеждания, а именно че несигурността на въведените стойности

остава недооценена, което рефлектира негативно върху стандартните грешки на оценките.

Основната идея на МВ е да се въвеждат ЛС като се използва подходящ въвеждащ модел, включващ в себе си определена въвеждаща процедура. Тази процедура се повтаря S пъти (S>2), като на всяка повторение се провежда желания анализ, например изчисляване на относителни дялове, оценка на параметрите на основен регресионен модел или др. във всяка от S-те пълни бази данни, получени след въвеждането. На следващ етап резултатите от S-те оценки се обобщават чрез правилата на Rubin (1987). Тази логика е изобразена на фиг. 2. За да работи методът е необходимо да са изпълнени определени правила, попадащи в рамките на понятието „подходящо въвеждане". Доналд Рубин въвежда идеята за *подходящо въвеждане* през 1987 г. за да може в следствие на въвеждаща процедура да бъдат получавани неизместени, ефективни оценки на параметрите на изследваните разпределения, включително техните вариации. Самата идея може да бъде представена по следния начин: Нека X и Y са две променливи и X има ЛС. Нека за да се въведат стойностите на X да се използва стохастична регресия:

първа стъпка $X_i = a + bYi$

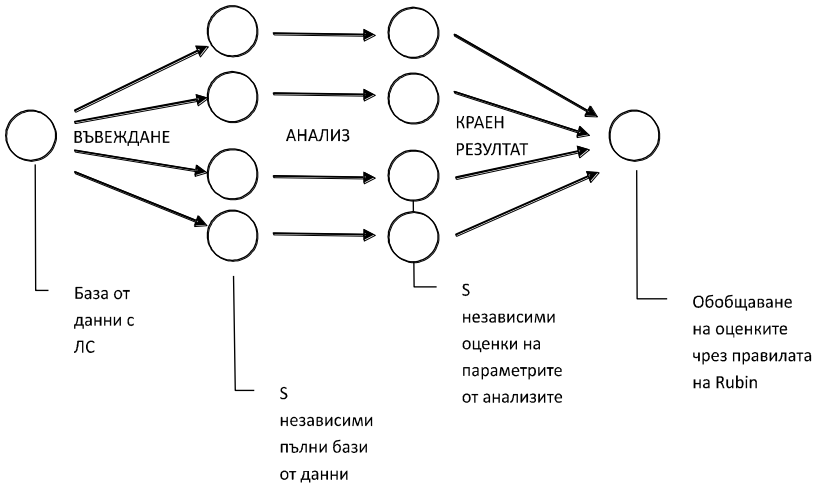
втора стъпка $Xi = a + bYi + S_{X.Y}*u_i$,

където a и b са регресионни коефициенти, $S_{X.Y}$ е остатъчната вариация, а u_i е случайно избран елемент от симулирано нормално разпределение със средна 0 и разсейване $S_{X.Y}$. В този подход на анализ, обаче има една особеност, че третира a, b и $S_{X.Y}$ като действителни параметри на генералната съвкупност, а не като техни оценки. В действителност стойностите на тези параметри са неизвестни и за подходящо МВ всеки въведен вектор от данни трябва да бъде базиран на различен набор от стойности за a, b и $S_{X.Y}$. Тези стойности, също така, трябва да бъдат случайно избрани от Бейсовите постериорни разпределения на параметрите. Прилагането на подходящо МВ използва S-те резултативни бази данни за прилагане на стандартните анализи като накрая резултатите се обобщават в един общ резултат. Различията между индивидуалните резултати при отделните въвеждания се използва за оценка на несигурността причинена от липсващите данни. Така множественото въвеждане може напълно да покрие несигурността по отношение на неизвестните параметри.

Емпиричен анализ

За показване на възможностите на МВ е направена симулация на данни, в които изкуствено са заложени ЛС. Симулираната база от данни съдържа 3 променливи - X1, X2, Y. X1, X2 и Y образуват многомерно нормално разпределение със следните характеристики: cor(X1, X2) = 0,3; cor(X1, Y) = 0,55; cor(X2, Y) = 0,75. Също така X1 има средна аритметична 50 и стандартно отклонение 50. X2 има средна 100 и ст. откл. 60, а Y има средна 200 и ст. откл. 60. В емпиричния анализ за представяне на МВ се използва регресионен анализ със зависима променлива Y и независими променливи X1 и X2. В симулираната база от данни резултатите от този регресионен анализ има следния вид[174] табл.1.

[174] Анализите са направени със софтуерния продукт R.



Фигура 2. Модел на провеждане на множествено въвеждане (адаптация по Enders C., 2010)

Таблица 1. Регресионни оценки от симулираната база от данни при връзка Y~X1+X2

	Оценка	**Ст. грешка**	**t стойност**	**Pr(>|t|)**
Свободен член	94,51	2,65	35,62	<2e-16 ***
X1	0,52	0,03	18,76	<2e-16 ***
X2	0,78	0,02	34,56	<2e-16 ***

Стат. знаичмост (кодове): 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

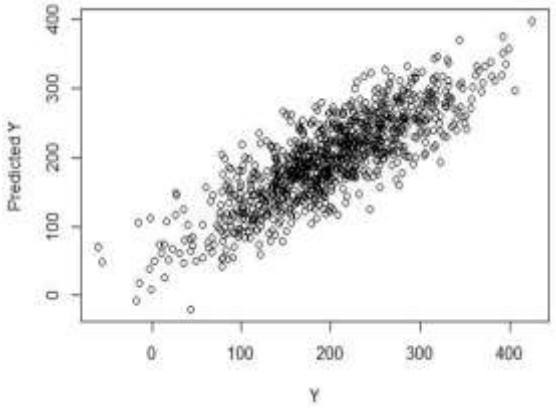
Станд. грешка на остатъците: 42,66 с 997 степени на свобода

Коеф. на детерминация: 0,6812, ажустиран коеф. на детерминация: 0,6806

F-statistic: 1065 с 2 и 997 DF, p-value: < 2.2e-16

Графичния вид на разпределението на оценените значения на Y спрямо действителните стойности могат да се видят на фиг. 3

Този първоначален анализ ще бъде използван за сравнение с резултатите от МВ. За целта са направени следните промени в базата от данни. При две от променливите: X1, X2 са запазени всички значения, докато при Y са премахнати 50% от значенията. Механизмът на тези ЛС е липсващ, напълно случайно.



Фигура 3. Разпределение на оценените значения на Y спрямо действителните при изходната база от данни и връзка Y~X1+X2

МВ се прилага при следните условия. Въвеждат се 5 независими бази данни. Методът на въвеждане в случая е избран да бъде по близост на оценените стойности, но може да бъде и друг [Лазаров, Д (2014) (дисертационен труд), стр. 42]. Върху всяка една от въведените бази от данни е приложен регресионен анализ от типа Y~X1+X2. Обобщаването на резултатите от петте анализа става чрез правилата на Рубин [Gelman и Hill (2007), стр. 542] (табл. 2)

Табл. 2 Обобщени оценки от МВ

	Оценка	**Ст. грешка**	**t стойност**	**Pr(>|t|)**
Св. член	95,71	3,01	31,75	0,00
X1	0,50	0,04	13,12	1,93e-10
X2	0,78	0,03	27,79	0,00

Основната цел на МВ е оценките на ЛС да бъдат третирани като такива, а не като истински, наблюдаеми значения. Това практически означава, че след обобщаването на оценките стандартните грешки (СГ) трябва да бъдат по-големи от действително съществуващите. Това е така, защото СГ обобщават грешките при измерванията и несигурността, която произтича от случайността. Тези резултати се виждат и при направения емпиричен анализ. Като се сравнят резултатите в табл. 1 и 2, колона Ст. грешка се вижда, че след МВ има адекватна (в контекста на Рубин) оценка на несигурността в следствие на ЛС.

Заключение

МВ е очевидно адекватен подход по отношение на третирането на ЛС. В съвременната теория и практика има редица методи за оценка на ЛС, които могат да се използват в контекста на МВ, но почти всички се основават на игнорируемостта на механизмите. В настоящото изследване бе показано предимството на този подход под ЛНС механизъм, когато базата от данни образува многомерно нормално разпределение. За съжаление това е един много ограничен случай и не може да бъде достатъчно показателен за възможностите на МВ. Основна задача пред бъдещите изследвания е разработването на нови методи, които да работят ефективно както под друг тип разпределения, така и под неслучайни механизми.

Литература

1. Лазаров, Д (2014). Някои възможности за въвеждане на липсващи и коригиране на грешни индивидуални данни при статистически изследвания, УНСС-София, (дисертационен труд).
2. Allison, P.D. (2002). Missing Data. Sage University Papers Series on Quantitative Applications in Social Science, 07-136. Thousand Oaks, CA: Sage
3. Enders, C. K. (2010) Applied missing data analysis, The Guilford Press
4. Gelman, A., Hill, J. (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press
5. Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Survey. New York: Wiley.

За контакти

гл. ас. д-р Деян Лазаров
Бургаски свободен университет
e-mail: deyanlazarov@bfu.bg