ВЪЗМОЖНОСТИ ПРИ ВЪВЕЖДАНЕТО НА ЛИПСВАЩИ СТОЙНОСТИ В ЕМПИРИЧНИТЕ ИЗСЛЕДВАНИЯ

Деян Лазаров
Бургаски свободен университет, Бургас

Резюме

В доклада се разглежда една възможност за въвеждане на липсващи стойности (ЛС), която се базира на използването на латентни променливи, както и факторни анализи – първоначален и потвърдителен за оценка на значенията на липсващата променлива. В случая като пример се използва наблюдението на работната сила в България, проведено от НСИ през 2007 г., но изводите и заключенията могат да се приложат върху всички масови изследвания, при които се наблюдават липсващи стойности.

Ключови думи: латентни променливи, липсващи стойности, факторен анализ.
Key words: latent variables, missing values, factor analysis.

Латентните променливи са псевдо реални променливи, които обобщават влиянието на действително наблюдавани променливи (признаци) в дадено изследване. Те най-често се получават като се приложи първоначален факторен анализ[1] върху базата от данни. Посредством този анализ се постигат два резултата. Първият е съкращаване на броя на променливите в базата от данни, като реално наблюдаваните такива се прегрупират в по-малко на брой фактори, наричани латентни променливи. Вторият е, че прегрупираните в един от новите фактори променливи, показват вътрешната съгласуваност и едновременно общо влияние в обяснението на вариацията на цялата база от данни. Силата на влиянието на отделните латентни променливи (нови фактори) може да се изследва посредством структурен или потвърдителен факторен анализи[2] [7, 8].

В настоящото изследване се използват възможностите на латентните променливи като се проследява начина на групиране на наблюдаемите променливи с липсващи стойности във фактори – латентни променливи и впоследствие се оценяват стойностите на тези латентни променливи. Така получените латентни променливи носят обобщеното влияние на променливите с ЛС и лесно могат де се включат във всякакви модели, на базата на които се въвеждат самите ЛС в базата от данни.

Липсващи стойности (ЛС)

Базата данни се разглежда като правоъгълна, образувана от отговорите на всеки един респондент в редовете и въпросите, на които те отговарят, в колоните. За ЛС се приема този случай, при който респондентът притежава значение по даден признак, но не го е посочил или е посочил грешно, както и случаите, при които по други причини то не е нанесено. За един респондент може да има липсващи стойности при повече от един въпрос (признак).

Данни

Специфичен интерес представлява появата на ЛС при заетите лица, т.е. разглеждат се единиците дали положителен отговор на въпроса: „През МИНАЛАТА СЕДМИЦА работили ли сте някаква работа срещу заплащане или друг доход (поне 1 час)?". Друго важно разделение на единиците се направи чрез това дали заетостта е на пълно или непълно работно време. В анализа се включват само единици заети на пълно работно време и така признаците обект на анализ се редуцират до 26, а единиците регистрирали значения по тези признаци 48 529 (Табл. 1). Признаците с ЛС са: Колко часа седмично работите ОБИКНОВЕНО на ОСНОВНАТА РАБОТА? (v14); Колко часа общо сте работили през МИНАЛАТА СЕДМИЦА на ОСНОВНАТА РАБОТА? (v22); Колко часа седмично желаете да работите - общо? (v25).

Механизми на ЛС

Важна част от правилния подход за анализ на ЛС е определянето на механизма на тяхната поява. В литерата се разглеждат три основни механизма [4, 6, 9, 10]. **Липсващи напълно случайно стойности (ЛНС)**, при който появата на самите липсващи стойности може да се

[1] Известен още като проучвателен или изследователски факторен анализ (explorartory factor analysis).
[2] Confirmatory factor analysis

разглежда като случайна извадка от единиците в изследваната база от данни. Това означава, че дори и те да бъдат детерминирани от дадена променлива или признак, той не присъства сред наблюдаваните. Вторият по-малко ограничаващ механизъм е **липсващи случайно стойности (СЛ)**. При него появата на липсващи стойности при даден признак е във функция на някои от наблюдаваните променливи, но не и от самия него. Третия и най-проблемен за анализ механизъм е известен като **не случайно липсващи (НеСЛ)**. При този механизъм се появява зависимост между липсващите стойности и самите значения на признака, при който се наблюдават. По друг начин казано, ЛС са във функция на самите себе си.

Проверката на механизмите на ЛС при Наблюдението на работната сила от 2007 г. е направено в [1, 2]. Проведения анализ еднозначно показва, че ЛС при заетите на основна работа през 2007 г. **не са липсващи напълно случайно**. Има ясна връзка между задавания въпрос и появата на липсваща стойност. Това се потвърждава и от теста на Литъл за ЛНС (Little's MCAR test): Chi-Square = 16327,786, DF = 45, Sig. = 0,000. Статистическата значимост на теста гарантира липсата на пълна случайност при появата на ЛС. Независимо, че делът на ЛС е нисък, това прави последващият анализ интересен и специфичен. Подходът при компенсиране на ЛС трябва да бъде съобразен с различията между отговорилите и неотговорилите и факторната зависимост между задавания въпрос и не получаването на отговори.

Връзката между признаците v14, v22 и v25 е изключително силна (Табл. 1).

Табл. 1. Кроскорелации

	v14	v22	v25
v14	1,000		
v22	0,922	1,000	
v25	0,981	0,898	1,000

Това се проявява и при появата на ЛС. Внимателно разглеждане на данните показва, че вероятността за поява на ЛС при единия признак е свързана с висока вероятност за поява на ЛС и при другите. Практически трите разпределения са много близки (Табл. 2), като се изключи асиметрията. Това дава основание да се предполага, че появата на ЛС, при която и да е променлива, е във връзка със самата променлива, което от своя страна означава, че механизма за поява на липсващи стойности трябва да се разглежда като НеСЛ.

Табл. 2. Основни характеристики на разпределенията на променливите v14, v22, v25

Показатели	V14	V22	V25
Наблюдавани стойности	46596	45987	46762
Липсващи стойности	1933	2542	1767
Средна аритметична	41,36	41,36	41,25
Ст. гр. на сред. аритметична	0,03	0,028	0,032
Медиана	40	40	40
Мода	40	40	40
Стандартно отклонение	6,502	6,068	6,89
Асиметрия	-1,468	0,898	-1,903
Ст. гр. на асиметрията	0,011	0,011	0,011
Ексцес	21,252	8,741	19,865
Ст. гр. на ексцеса	0,023	0,023	0,023
Минимум	0	0	0
Максимум	96	96	96

За алтернативна проверка на този механизъм се използва последователни кластърни модели с нарастващи число на кластърите. При направения анализ се установи, че при групирането на единиците в 8 кластъра в един от тях се получават центрове при променливите с ЛС (v14, v22, v25), значимо различни от останалите. Ако в останалите кластъри центровете съответстват на общите средни при тези променливи, т. е. близки до 41 часа, то в **кластър номер 7**, центровете при тези променливи са със стойности близки до 61 часа (Табл. 3).

Табл. 3. Характеристики на разпределенията с липсващи стойности сред променливите в кластър 7

Брой		Средна аритметична	Стандартно отклонение	Липсващи стойности		Бр. екстремни стойности	
		(Mean)	(Std. Deviation)	Бр.	%	Долна граница	Горна граница
v14	1311	61,3	7,599	973	42,6	53	273
v22	1317	61,38	7,938	967	42,3	52	282
v25	1311	59,88	8,809	973	42,6	140	248
a. Бр. случаи извън границите (Mean–2.SD, Mean + 2.SD).							

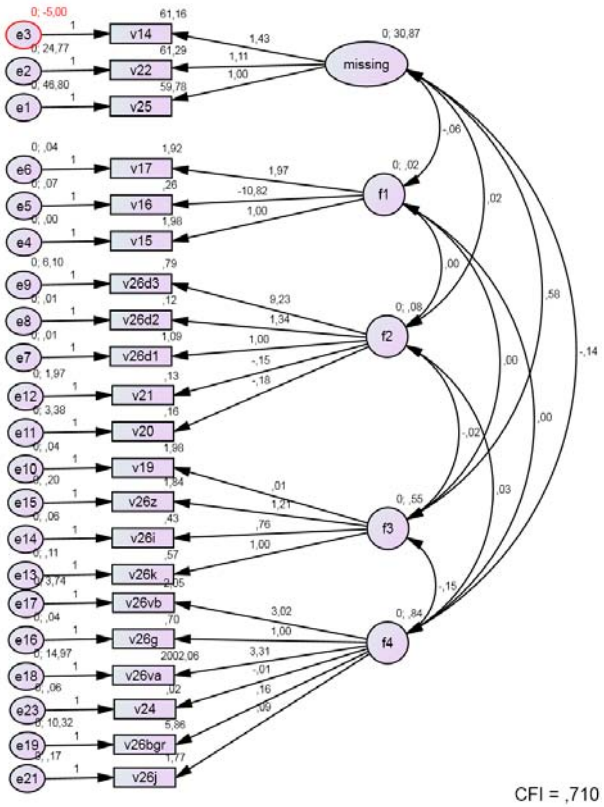
Анализират данните от Клъстер 7

Като пример в доклада ще се използват единиците от Кластър 7 в наблюдението на работната сила. При тях се наблюдава много слаба факторна пригодност (Kaiser-Meyer-Olkin = 0,580). При факторния анализ се изолират 9 фактора със собствени стойности над 1,00, обясняващи 80,556% от вариацията на данните и променливите с липсващи стойности (v14, v22, v25) се групират в един общ фактор, наречен **„missing"**. Това потвърждава силната връзка и съгласуваност при проявата на ЛС при трите променливи. Използва се Вирамакс ротация на факторите и екстракция посредством метода на главните компоненти. Използването на този набор от фактори, обаче, не дава възможност за максимизиране на функцията на максималното правдоподеобие, което налага търсене на „добрия" модел, чрез пренареждане на променливите във факторите и редуциране на самите фактори [4, 8]. В помощ на тази процедура идва възможността да се изпълни обяснителен факторен анализ с помалко на брой фактори от 9. Те се получават на базата на прегрупирането на променливите в по-малко на брой латентни променливи-фактори. Наблюдаваните медиаторни влияния са статистически значими, което подсилва значението на латентна променлива в анализа.

Потвърдителен факторен анализ

При единиците от Кластър 7 се използва потвърдителния факторен анализ за въвеждане на значенията на латентната променлива именувана „missing". В търсене на най-адекватния модел се стигна до модел с 5 латентни фактора (Фиг. 1). За да бъде въвеждането възможно е необходимо всички вариации на параметри в модела да бъдат положителни. Във факторния модел, както се вижда от пътечковата диаграма на граф. 4, вариацията на ненаблюдаваните остатъци e3 при моделирането на променливата v14 е отрицателна (ve3 = -5,00). Това налага използването единствено на Бейсов подход за въвеждане на стойностите на латентната променлива „missing", като се ограничава априорното разпределение на вариацията на e3 само в положителните стойности.

Характеристиките на модела показват слаба адекватност: CFI = 0,710; RMSEA = 0,161; PRATIO = 0,775. Въпреки всичко информацията, която ще пренесе върху "missing" е достатъчна при последващото въвеждане на ЛС при променливите v14, v22 и v25.



Фиг. 1. Модел на факторна връзка за единиците от Кластър 7

Заключение

Използването на латентни променливи е един инструмент, който успешно може да се използва в анализа на ЛС. По този начин се „пренася" информацията на всяка от засегнатите с ЛС променливи без да се налага нейното включване в модела, на базата на който се осъществява въвеждането. Това преодолява в максимална степен опасностите, които може да предизвика колиниарността между независимите променливи и оттам да се опорочи целия анализ. Използването на модели на зависимостите между променливите с ЛС и останалите променливи е в основата на решаването на проблема при НСЛ механизъм. В конкретния случай това означава, че иначе силно корелираните признаци v14, v22 и v25 не се налага да участват в един модел, като тяхното общо влияние е заместено от латентна променлива. Последващите анализи за въвеждане на самите ЛС могат да бъдат базирани на различни модели и подходи, както параметични (базирани на оценка на функцията на максималното правдо-

подобие), така и непараметрични (например невронни мрежи).

Литература

1. Лазаров, Д. *Липсващите стойности при наблюдението на работната сила – 2007 г. в България*. Годишник с научни трудове. БСУ. 2010.
2. Лазаров, Д. *EM или DA или EM и DA*. сп. Бизнес посоки. бр. 1. 2011.
3. Манов. А, *Многомерни статистически методи със SPSS*. УИ „Стопанство". София. 2002.
4. Enders, C. *Applied missing data analysis*. The Guilford Press. 2010.
5. Little, R., Rubin, D. *Statistical analysis with missing data*. Wiley. New York. 1987.
6. Little, R., Rubin, D. *Statistical Analysis with Missing Data*. 2nd ed. Wiley. New York. 2002.
7. MacKinnon, D. *Introduction to Statistical Mediation Analysis*. Taylor & Francis Group. LLC. 2008.
8. Raykov, T., Marcoulides, G. *A First Course in Structural Equation Modeling*. Second Edition. Mahwah. NJ. Lawrence Erlbaum Associates. 2006.
9. Rubin, D. *Multiple Imputation for Nonresponse in Survey*. Wiley. New York. 1987.
10. Scheffer, J. *Dealing with Missing Data*. Research Letters in the Information and Mathematical Sciences. 3. 2002. 153-160.

SOME POSSIBILITIE IN MISSING VALUES IMPUTATION IN EMPIRICAL RESEARCHES

Deyan Lazarov
Burgas Free University, Burgas, Bulgaria

Abstract

In this report is considered an opportunities to impute the missing values. The approach is based on usage of latent variables as well as factor analysis – exploratory and confirmatory to evaluate the items of the latent variable. In this paper as example is used the Labour Force Survey conducted in Bulgaria 2007 by NSI but the conclusions can be spread throughout all statistical surveys with missing values in included variables.