СПЕЦИФИЧНИ ВЪЗМОЖНОСТИ ЗА АНАЛИЗ НА ЛИПСВАЩИ СТОЙНОСТИ В ЕМПИРИЧНИТЕ ИЗСЛЕДВАНИЯ

Деян Лазаров
Бургаски свободен университет

SPECIFIC POSSIBILITIES IN MISSING VALUES ANALYSIS IN EMPIRICAL RESEARCHES

Deyan Lazarov
Burgas Free University

Abstract: *The study considered an option for the introduction of missing values, which is based on the use of the latent variables, and the factor analysis - exploratory and confirmatory for evaluation of the values of the missing variable. As an example of the application of this idea is used the Labor Force Survey in Bulgaria, conducted by the National Statistical Institute in 2007, but the findings and conclusions can be applied to all mass studies in which missing values are observed.*

Key words: *latent variables, missing values, factor analysis.*

Латентните променливи са не реално наблюдавани променливи, които обобщават на по-високо йерархично ниво влиянието на действително наблюдавани променливи (признаци) в дадено изследване. От практическа гледна точка 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 помалко 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 дисперсията на признаците в цялата база от данни. Силата на влиянието и обективното съществуване на отделните латентни променливи (нови фактори) може да се изследва посредством потвърдителен факторен анализ[2] [7,8].

Настоящото изследване цели представяне на възможностите на латентните променливи (ЛП) в анализа на липсващи стойности (ЛС). Една от основните трудности, които произтичат от наличието на ЛС в базите от данни е невъзможността да се прилагат статистическите анализи, а ако такива се приложат много често се достига до грешни заключения. Едно възможно решение на съществуващия проблем е изолирането на ЛП като се проследява начина на групиране на наблюдаемите променливи с ЛС във фактори и оценката на стойностите на тези ЛП. Така получените нови ЛП носят обобщеното влияние на променливите с ЛС и лесно могат де се включат във всякакви анализи и модели, на базата на които се въвеждат дори и самите ЛС в базата от данни.

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

Липсващи стойности

Когато се говори за ЛС, базата данни в която те се появяват 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).

Емпирични данни за апробация на идеята

Апробацията на представената идея се прави върху данните от „Наблюдение на работната сила“, направено от НСИ през 2007 г. Специфичен интерес представлява появата на ЛС при заетите лица, т.е. разглеждат се анкетираните 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 на наблюдение се направи чрез това дали заетостта е на пълно или непълно работно време. В настоящото изследване се включват само лица 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 в обособената база от данни 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 литературата се разглеждат три основни механизма [4, 6, 9, 10]. **Липсващи напълно случайно стойности (ЛНС)**, при който появата на самите липсващи стойности може да се 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 обособената база от данни при Наблюдението на работната сила от 2007 г. е направено в предишни публикации на автора [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 тестът на Литъл за ЛНС (Little's MCAR test): Chi-Square = 16327,786, DF = 45, Sig. = ,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 влиянието на ЛС трябва да бъде съобразен с различията между отговорилите и неотговорилите и зависимостта между задавания въпрос и не получаването на отговори.

Връзката между признаците v14, v22 и v25 е изключително силна (Табл. 2). 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. 1),

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

Таблица 1. Основни характеристики на разпределенията на променливите v14, v22, v25

Показатели	**V14**	**V22**	**V25**
Наблюдавани единици	46596	45987	46762
Единици с ЛС	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

Таблица 2. Кроскорелации

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

Таблица 3. Характеристики на разпределенията с липсващи стойности сред променливите в Клъстер 7

	Брой	Средна аритметична (Mean)	Стандартно отклонение (Std. Deviation)	Липсващи стойности Бр.	%	Брой РОС[18] 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

За илюстриране на възможностите на анализа с ЛП ще се използват единиците от Кластър 7 в „Наблюдението на работната сила“. Изборът е мотивиран и от спецификата на единиците в клъстера и от големия процент на ЛС сред тези единици. Наблюдава много слаба факторна пригодност (Kaiser-Meyer-Olkin = 0,580), което е индикация, че при признаците като цяло липсва добра вътрешна съгласуваност и факторния анализ би бил затруднен. При самия първоначален факторен анализ, с Вирамакс ротация и метод на главните компоненти за екстракция, 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 **„m”**. Това потвърждава силната връзка и съгласуваност при проявата на ЛС при трите променливи. Следваща стъпка в анализа е потвърждаване на получените латентни фактори чрез обясняване на общата дисперсия на данните. Това се прави чрез потвърдителен факторен анализ и използването на модели със структурни уравнения. 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].

Потвърдителен факторен анализ

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. Потвърдителния факторен анализ дава решения в две посоки. Едната посока е да се даде възможност да се въведат директно ЛС на базата на добре обоснован и във висока степен адекватен модел като се използва някой от познатите методи за това [6, 9, 10]. Другата другата посока е същият модел да се използва за оценка на стойностите на ЛП, които от своя страна да се използват в бъдещи анализи. Може да се отбележи, че ЛП могат да се разглеждат като променливи със 100% ЛС. Следователно и тук е желателно да се избере подходящ метод за въвеждане на стойностите на ЛП. Удобство в случая е, че двете задачи могат да бъдат решени едновременно на базата на един модел и с помощта на една и съща въвеждаща процедура.

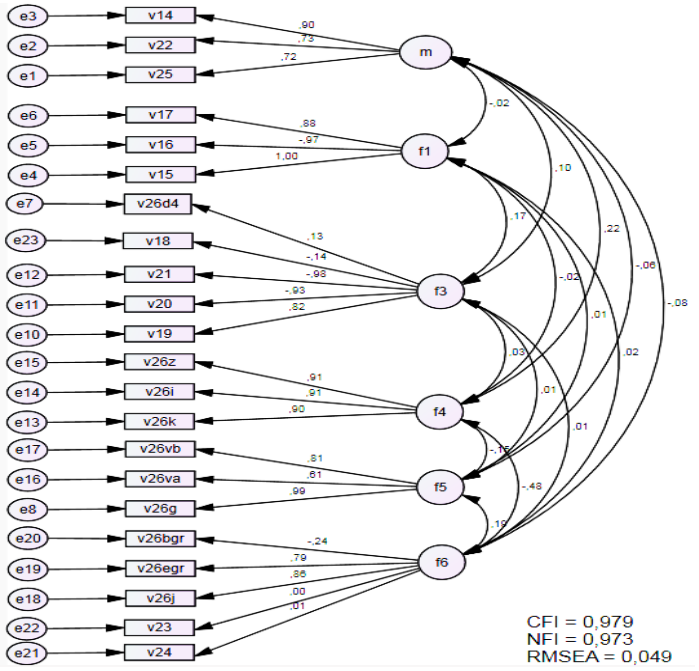
На базата на получените резултати от обяснителния факторен анализ се конструира структурен модел, моделиращ взаимодействието между латентните фактори в базата от данни. Целта е постигане на максимална стойност на съответствие между ем-

[18] РОС – рязко отклоняващи се стойности.

пиричната и моделната ковариационни матрици. В търсене на най-адекватния модел се стигна до модел с 7 латентни фактора (фиг. 1). В модела в правоъгълници са представени променливите, които реално се наблюдават в НРС. В овали (кръгове или елипси) са представени латентните променливи в модела (f1, f2, f3, f4, f5, f6 и m). Както се вижда от фиг. 1, освен факторите получени от обяснителния анализ като латентни фактори в модела се представят и т.н. грешки при оценката на наблюдаемите променливи – от e1 до e23. Тези грешки се получават като остатъчна вариация, която латентния модел не може да обясни. Така тези грешки могат бъдат разглеждани като необяснена вариация или вариация, породена от фактори, не разглеждани в модела. За да бъде въвеждането възможно е необходимо всички параметри в модела вариации на параметри в модела да бъдат положителни. Много често при оценката на моделите се получава остатъчна вариация, при една или повече наблюдавани променливи, с отрицателна стойност.

Тъй като въвеждането на ЛС при променливите v14, v22 и v25, както и при всички латентни фактори се основава на използването на обобщените характеристики в модела, то не е възможно това въвеждане да се осъществи ако има неестествени характеристики на някои от тях. Самото оптимизиране на модела се основава на тези единици от Клъстер 7, при които няма ЛС, които са 1311.

Характеристиките на модела показват много добра адекватност: CFI = 0,979; RMSEA = 0,049; NFI = 0,973; CMIN/DF = 4,201. Това дава основание да се приеме, че информацията, която ще се пренесе върху променливите v14, v22 и v25 и латентните фактори е достатъчна при използването на този модел за анализ на ЛС в базата от данни.



Фигура 1. Модел на факторна връзка за единиците от Клъстер 7

Заключение

Представеното изследване показва, че ЛП са 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 силната връзка (колинеарността) 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 правдоподобие), така и непараметрични (например невронни мрежи).

Литература:

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