ИЗПОЛЗВАНЕ НА ЛАТЕНТНИ ПРОМЕНЛИВИ ПРИ ВЪВЕЖДАНЕТО НА ЛИПСВАЩИ СТОЙНОСТИ

*Деян Лазаров**

Настоящата статия се занимава с анализа на липсващите стойности (ЛС) при масовите изследвания и с една възможност за неговото провеждане. Проблемът на липсващите стойности се състои в това, че значения на признаците в дадено изследване, които трябва да бъдат наблюдавани при отделните единици, всъщност липсват. Тези липсващи стойности не означават само по-малка ефективност на оценките поради редуцирането на размера на базата данни, но също че стандартните методи за анализ на пълни бази данни не могат да бъдат използвани веднага [9]. В случаите на непълни бази данни рискът от вземане на неправилно решение е изключително висок, защото поради липсващите стойности се намалява действието на доверителните интервали, редуцира се силата на статистическите анализи и се получават изместени оценки.

Важна част от правилния подход за аназлиз на ЛС е определянето на механизма на тяхната поява. В литературата се разглеждат три механизма, описани от Доналд Рубин през 1976 година. Първият е **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. Третият и „най-проблемен" за анализ механизъм е известен като **неслучайно 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.

За да се изследва механизмът на появяване на ЛС, се прилага *t* тест на Стюдент при различни дисперсии при контролирано влияние на признаците с ЛС. При този подход, ако резултатите при единиците с ЛС не се различават от тези при единиците без ЛС, механизмът на появата им може да се приеме за ЛНС. В противен случай появата на липсващи стойности е във връзка с променливите

* Главен асистент, Бургаски свободен университет; e-mail: deyanlazarov@bfu.bg .

в базата данни и механизмът е ЛС или НеСЛ. За всяка променлива двойките са формирани на базата на използван индикатор: присъстващ/липсващ (present/missing). Друг метод за проверка на механизма ЛНС е тестът на Литъл за ЛНС (Little's MCAR test), базиран на Махаланобис разстоянието от центъра на групите, образувани от различните модели на ЛС [5].

При отхвърляне на хипотезата, че механизмът е ЛНС, настъпват трудности с еднозначното определяне на действащия механизъм между СЛ и НеСЛ. При механизма СЛ липсващите стойности са разпределени неслучайно по отношение на всички наблюдения, но би трябвало да са случайно разпределени в една или повече подгрупи (подизвадки), определени от значенията на анализирания признак Y_i. При механизма НеСЛ липсващите стойности са неслучайно разпределени по отношение на всички наблюдения и вероятността за появяване на липсваща стойност не може да бъде оценена на базата на променливите в модела. Стандартните подходи за решения в подобни случаи са използването на външни източници, минал опит и предположения за поява на ЛС.

Методика на изследването

В изследването се проучва последният и „най-неудобен" механизъм на ЛС - Неслучайно липсващи (НеСЛ). При него винаги се налага данните да се моделират и на базата на тези модели да се направи самото въвеждане. Преди процеса на моделиране обаче трябва ясно да се определят подгрупите от единици, в които „действат" случайни механизми на поява на ЛС. Ако тази стъпка не е направена коректно, резултатите от процедурата по въвеждане могат да бъдат сериозно опорочени. В изследването се използва клъстеризация по метода K-средни. Предположението е, че ако механизмът е НеСЛ, т.е. има значима връзка между появата на ЛС и значенията на променливата, при която се появяват, то би трябвало да може да се „разпознаят" моделите в базата от данни, при които трябва да има значима клъстерна разлика (най-вече на средните стойности на променливите с ЛС в модела).

След като подгрупите със случайни механизми на поява на ЛС са определени, се пристъпва към моделиране на самите ЛС. За съжаление може да се окаже, че когато има повече от една променлива с ЛС между тях, може да има силна корелационна връзка. Това означава, че те не могат да се използват в един модел, при който няколко от тях да бъдат предиктори или обясняващи променливи по отношение на една зависима променлива с ЛС поради опасността от колинеар-

ност[1]. Това от своя страна означава, че голяма част от информативността в базата от данни по отношение на зависимата променлива отпада и моделът може да бъде зле обусловен, което е в основата на лоши резултати след въвеждащата процедура. За съжаление, тази „жертва" на информативност е крайно нежелана в анализа на ЛС [5, 7, 9, 10]. Напротив, винаги се препоръчва в анализа да се използва цялата налична информация, за да се предпазят променливите от несъществуващи или подценени зависимости след въвеждащата процедура.

На базата на изложените съображения в настоящото изследване се представя възможността за използване на латентни променливи в анализа на липсващи стойности с оглед да се избегне нежеланата зависимост - колинеарност между предикторите в модела за въвеждане на ЛС, и едновременно с това да се запази пълната (или най-малко максималната) информативност в базата от данни. Тази идея е мотивирана от факта, че латентните променливи се получават вследствие на факторен анализ, чрез който се намалява размерността на базата от данни, като в максимална степен се запазва информативността в нея. Така при прилагане на факторен анализ и екстракция на латентни фактори на базата на наличната информация може да се определи към кои латентни фактори се отнасят променливите с ЛС. Вследствие на това може да се създаде структурен модел, чрез който да се въведат стойностите на тези латентни променливи, които впоследствие да се използват като предиктори при въвеждането на самите ЛС. Друго възможно решение е използването на самия структурен модел за въвеждане на ЛС.

С оглед запазване на фокуса на статията върху използването на латентните променливи в процеса на анализ на ЛС въвеждането на самите ЛС няма да бъде разглеждано.

1. Общо описание на признаците и липсващите стойности в тях

Признаци	Брой единици	Средна аритметична (Mean)	Стандартно отклонение (Standard deviation)	Липсващи стойности		Брой екстремни стойности*	
				брой	%	долна граница	горна граница
v14	46596	41.36	6.502	1933	4.0	527	1232
v25	46762	41.25	6.890	1767	3.6	682	1221
v22	45987	41.36	6.068	2542	5.2	1501	1321

*Брой случаи извън границите (Mean - 2*SD, Mean + 2*SD).

[1] Колинеарността/мултиколинеарността е нежелана корелационна зависимост между предиктора/ите (фактора/ите) в даден регресионен модел. Тя води до изместване на оценките на параметрите на регресионната зависимост и оттам до изместване на оценките на ЛС. Повече информация може да се намери в [4, стр. 129].

Като илюстрационен пример в изследването се използва Наблюдението на работната сила в България през 2007 г. и липсващите стойности, които се появяват при него. Използват се данни за цялата 2007 г., а не по тримесечия. В изследването на работната сила, проведено от НСИ през 2007 г., са наблюдавани общо 160 признака, част от които определят домакинството и единицата от домакинството, обект на изследване, чрез демографски характеристики, а друга част се отнасят за заетостта при последната седмица на основна и допълнителна работа, форми на заетост и активност при търсене на работа и др. В настоящото изследване се налага редуциране на признаците поради няколко причини. От една страна, непритежаването на дадено значение на признак води до автоматичното отпадане на въпроси в анкетната карта, задавани на респондента. От друга страна, редица от признаците са технически идентификации на единиците в съвкупността, въведени от изследователския екип, които логически не се отнасят към изследвания обект, а именно липсващите стойности.

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, са 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).*

Липсващи стойности

Базата данни се разглежда като правоъгълна, образувана от отговорите на всеки един респондент, в редовете и въпросите, на които те отговарят, в колоните. **За липсваща стойност (ЛС) се приема този случай, при който респондентът притежава значение по даден признак, но не го е посочил или е го 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). В случая не може да се изследват липсващите данни вследствие на отказ на даден респондент въобще да участва в наблюдението, както и домакинства, които са попаднали в извадката, но не са намерени или са отказали сътруд-

ничество. Може да се допусне, че подобни домакинства има и проблемът произтича основно от неактуалните списъци с информация от Преброяване 2001, върху които се гради вероятностната извадка.

Анализ на механизма на ЛС в Наблюдението на работната сила в България през 2007 година

Проведеният анализ еднозначно показва, че липсващите стойности при заетите на основна работа през 2007 г. **не могат да се определят като липсващи напълно случайно** [1, 2]. 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 = 16 327.786, DF = 45, Sig. = 0.000. Статистическата значимост на теста гарантира липсата на пълна случайност при появата на ЛС. Независимо че делът на ЛС е сравнително нисък, това прави последващия 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 неполучаването на отговори.

2. Кроскорелации

	v14	v22	v25
v14	1.000		
v22	0.922	1.000	
v25	0.981	0.898	1.000

Връзката между признаците v14, v22 и v25 е изключително силна (табл. 2). Това се проявява и при появата на ЛС. Внимателното 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. 3), 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 механизмът за поява на липсващи стойности трябва да се разглежда като НеСЛ.

3. 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
Стандартна грешка на средната 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
Стандартна грешка на асиметрията	0.011	0.011	0.011
Ексцес	21.252	8.741	19.865
Стандартна грешка на ексцеса	0.023	0.023	0.023
Минимум	0	0	0
Максимум	96	96	96

За алтернативна проверка на този механизъм се използват последователни клъстерни модели с нарастващо 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 променливи, т.е. близки до 41 часа, то в **клъстер номер 7 в анализа** центровете при тези променливи са със стойности близки до 61 часа (табл. 4).

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

	Брой	Средна аритметична (Mean)	Стандартно отклонение (Standard deviation)	Липсващи стойности	
				брой	%
v14	1311	61.30	7.599	973	42.6
v22	1317	61.38	7.938	967	42.3
v25	1311	59.88	8.809	973	42.6

При провеждането на клъстерния анализ липсващите стойности са отстранени чрез процедурата елиминиране по двойки[2], т.е. използват се всички налични данни за оценка на разстоянията в клъстерите, за всеки две променливи в

[2] Pairwise delition.

анализа. Така размерът на този клъстер е втори по големина с обем 2 290 единици, като най-големият се състои от 45 951 случая, а всички останали включват по-малко от 90 единици. Тези резултати показват, че в базата данни има единици, които се групират с други подобни и имат различни, специфични характеристики по анализираните признаци (v14, v22, v25) от останалите [1, 2]. Това е основание да се предположи, че ако при тези единици се появят ЛС, то те ще имат различен модел на поява, т.е. **механизмът действително е НеСЛ**.

При разглеждането на ЛС в клъстерите се вижда, че сред единиците от Клъстер 7 има изключително висок процент липсващи значения (табл. 4). Това налага при тяхното въвеждане да се подходи специфично, те да бъдат отделени и техният анализ да се базира на информацията от самия Клъстер 7.

При анализа на ЛС, когато механизмът е НеСЛ, е задължително моделирането на поява на самите ЛС в подгрупи, определени от самите модели на ЛС. В случая един такъв модел е Клъстер 7. Както се вижда, там се наблюдават специфични характеристики на променливите с ЛС. За съжаление, променливите, обременени с ЛС, са силно свързани помежду си и това ограничава възможността две от тях да се използват като предиктори в модела на поява на ЛС при третата. Това има огромен недостатък. В случай че дадена променлива (особено когато тя е от основен интерес в анализа) е изключена от модела на поява на ЛС, има огромна вероятност в процеса на въвеждане на ЛС тя да бъде силно подценена или изкривена зависимостта ѝ с останалите променливи в базата от данни [5, 6, 7, 10, 11]. В този случай е изключително наложително да се намери решение, което едновременно да не компрометира анализа чрез колинеарност или мултиколинеарност между предикторите и да подсигури информативността в базата от данни и зависимостите между променливите в базата от данни вследствие на анализа на ЛС. Решение в тази посока може да се намери в използването на латентни променливи и използването на вътрешната факторна структура на данните в анализа на липсващи стойности. Това ще бъде приложено върху данните само от Клъстер 7, като останалата част от базата данни ще бъде пренебрегната с оглед запазване на фокуса на публикацията.

Определяне на латентната факторна структура при единиците от Клъстер 7

Първоначално се прилага **обяснителен факторен анализ**[3] с цел да се определи факторна структура на данните. Наблюдава се 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

3 Confirmatory factor analysis.

един общ фактор, наречен в случая „missing". Използва се Вирамакс ротация на факторите и екстракция посредством метода на главните компоненти. Резултатите от обяснителния факторен анализ имат до голяма степен ориентировъчен характер. За съжаление, те могат да се разглеждат като необходимо, но не и достатъчно условие са определяне на вътрешната латентна факторна структура на данните.

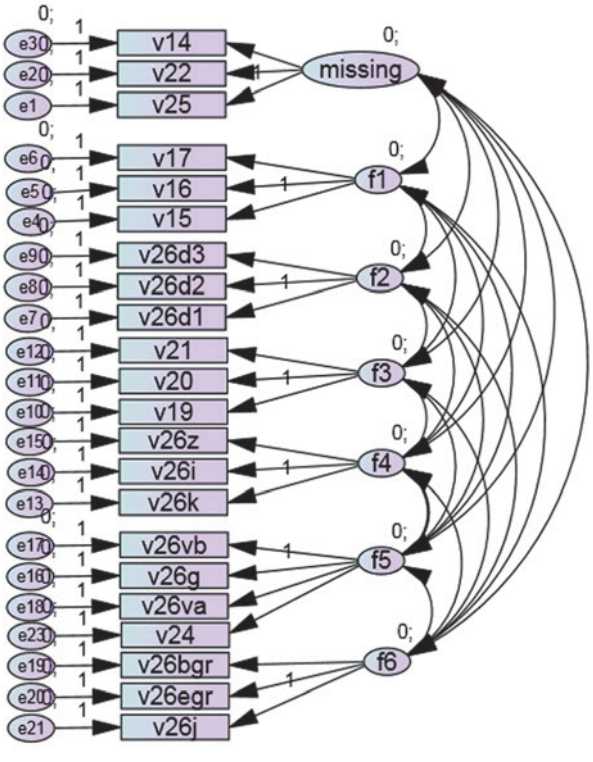
Потвърдителен факторен анализ

На следващата стъпка в анализа се прилага **потвърдителен факторен анализ**[4] при единиците от Клъстер 7. Целта е да се потвърди или разкрие действителната вътрешна структура на данните. Този вид анализ е форма на моделите със структурни уравнения[5] [8]. На базата на потвърдителния модел се получават адекватни оценки за латентните фактори, които да се използват при последващия анализ и въвеждане на самите ЛС. За да се постигне действително адекватен модел, се налагат някои вътрешни модификации на латентната структура и въвеждането на допълнителни релации между променливите в потвърдителния модел [8]. Латентните фактори се редуцират до 7 (от f1 до f6 плюс „missing") и така се достига до изключително добри моделни показатели за адекватност: CFI = 0.982; NFI = 0.981; RMSEA = 0.042[6]. Първите два индекса показват над 98% обяснена вариация на данните, а RMSEA показва по-малко от 5% средна грешка на модела. Илюстративно графично изображение на модела може да се види на фиг. 1. С елипси са отбелязани латентните ненаблюдавани променливи в модела, а с правоъгълници са изобразени наблюдаваните. Вижда се, че всяка наблюдавана променлива е свързана с две латентни променливи. От една страна, с въведените факторни латентни променливи - от f1 до f6 и „missing", а от друга, с променливи, означени с буквата „e" - от e1 до e23. Вторите са т.н. остатъчни вариации в модела или грешки, които се отнасят към всяка наблюдавана променлива и отразяват тази част от вариацията на самите наблюдавани променливи, която не е обяснена с изследвания модел. В описанието на модела се използват също два вида стрелки - еднопосочни, които показват регресионни зависимости (или тегла), и двупосочни, които измерват наличието на ковариации (корелации при използването на стандартизирани оценки). Високата адекватност на модела гарантира, че информацията, която ще се пренесе от латентната променлива

[4] Confirmatory factor analysis.
[5] Structural equation modeling (SEM).
[6] За повече информация относно индексите за оценка на адекватността на модела вж. [8].

„missing", е достатъчна при последващото въвеждане на ЛС при променливите v14, v22 и v25.

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



Заключение

Използването на латентни променливи, обобщаващи влиянията на признаците с липсващи стойности, е 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 до голяма степен опасностите, които може да предизвика колинеарността между независимите променливи и оттам да се опорочи целият анализ. 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 параметрични (базирани на оценка на функцията на максималното правдоподобие), така и непараметрични (например невронни мрежи). Въпреки всичко трябва да се държи сметка за евентуалното нежелано мултиколинеарно взаимодействие при използването на максимално правдоподобни оценки поради включването на едни и същи екзогенни променливи при моделирането на латентната променлива „missing" и моделите за въвеждане на самите ЛС [5]. Обект на бъдеща работа е оценката на тези опасности. Очаква се при ортогонални ротации (както в случая) при факторните модели такива опасности да се тушират. Също така възможността за прилагане на „подходящо"[7] множествено или единично въвеждане на базата на модела с латентни променливи е възможност, която силно мотивира бъдещи изследвания. Не трябва да се пропуска, че наред с потвърдителните факторни модели могат да се използват и други структурни модели, които дават възможност за въвеждане на липсващите данни. Анализът на медиаторните връзки би представил по-добре вътрешните взаимовръзки между променливите и би допълнил анализа на ЛС. Тук трябва да се предположи, че качествата на моделите ще бъдат най-добри, когато и самите те са добре обосновани. Интересна е идеята доколко може да се направи компромис с адекватността на моделите? Може да се предполага, че има определена свобода, която допуска по-ниска адекватност без значима „жертва" на информативност. Обект на бъдеща работа е изследването на тази особеност.

[7] В смисъла на Рубин [6].

ЦИТИРАНА 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. „Бизнес посоки", бр. 1, 2011 г.

3. Манов. А. (2002). Многомерни статистически методи със SPSS, УИ „Стопанство", София.

4. Съйкова, И. (1991). Статистически изследвания на зависимости и други връзки в социално-икономическата област (II част), УИ „Стопанство" София.

5. Enders, C. K. (2010). Applied missing data analysis, The Guilford Press.

6. Little, R. J. A, Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

7. Little, R. J. A, Rubin, D. B. (2002). Statistical Analysis with Missing Data - 2nd ed., New Jersey: Wiley.

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. Schafer, J. L (1997). Analysis of Incomplete Multivariate Data, Chapman & Hall.

11. Scheffer, J. (2002). Dealing with Missing Data, Research Letters in the Information and Mathematical Sciences 3, pp. 153 - 160.

ИСПОЛЬЗОВАНИЕ ЛАТЕНТНЫХ ПЕРЕМЕННЫХ ПРИ ВВЕДЕНИИ ОТСУТСТВУЮЩИХ СТОИМОСТЕЙ

*Деян Лазаров**

РЕЗЮМЕ В исследовании рассматриваются отсутствующие стоимости и предлагается возможность для их анализа с помощью латентных переменных. Первоначально определен механизм возникновения самих отсутствующих переменных, используя для этой цели возможности кластерного анализа. В качестве примера использовано Обследование рабочей силы в Болгарии, проведенное Национальным статистическим институтом в 2007 году. В анализе включены единицы, являющиеся «занятыми» в этот период времени. Выделены три основные признака, при которых проявляются отсутствующие стоимости (ОС). После анализа механизмов возникновения ОС установлено, что они не отсутствуют случайно (НеСО). Применяется последовательный кластерный анализ (К-средних) для обособления двух групп единиц, при которых существует значительная разница в отношении проявления значений различных признаков. Для иллюстрации метода использована одна из групп, названная «Кластер 7». Сначала использован объяснительный факторный анализ, а затем подтвердительный для обособления различных латентных переменных-факторов. Наблюдается также появление одной скрытой переменной, охватывающей все признаки с ОС, названной «missing». На основании полученных результатов предлагаются возможности для использования этой скрытой (missing) переменной в качестве фактора при последующем анализе и введении самих ОС.

* Главный ассистент, Бургасский свободный университет; e-mail: deyanlazarov@bfu.bg .

USE OF LATENT VARIABLES IN THE INTRODUCTION OF MISSING VALUES

*Deyan Lazarov**

SUMMARY The study considered missing values and offers the possibility of their analysis with latent variables. Initially the mechanism of occurrence of missing values by simply using the opportunities of the cluster is determined. As an example, the Labour Force Survey conducted by NSI in 2007 in Bulgaria is used. The analysis covered units that were occupied at that time. Three main attributes that occur in missing values (MV) are isolated. After analyzing the mising data the mechanisms (MV) it was established that it is not missing at random (NMAR). A consistent cluster analysis (K-medium) for adaptation of two groups of units in which there is significant difference in the expression of meanings in different signs is applied. In order to illustrate the method one of the groups named ′Cluster 7′ is used. Explanatory factor analysis is applied first and then confirmation of separation of the various latent variables-factors is also applied. There is also evidence of a latent variable encompassing all signs of MV, called ′missing′. Based on the results possibilities for use of this latent ′missing′ variables a factor in the subsequent analysis and missing data imputatiom.

* Chief Asisitent Burgas Free University, e-mail: deyanlazarov@bfu.bg