АНАЛИЗ НА МАДИАТОРНИТЕ ВЛИЯНИЯ ПРИ ЛИПСВАЩИТЕ СТОЙНОСТИ В НАБЛЮДЕНИЯТО НА РАБОТНАТА СИЛА В БЪЛГАРИЯ

Гл. ас. Деян Лазаров, БСУ

ANALYSIS OF MEDIATION INFLUENCE FOR MISSING VALUES IN LABOUR FORCE SURVEY IN BULGARIA

Deyan Lazarov

***Abstract**: The research investigates the latent influence upon missing values (MV) appearance in Labour Force Survey conducted in Bulgaria 2007 by National Statistical Institute. In previous research [1] was shown that mechanism of MV is MNAR and after K-means clustering procedure the database was split in to two general clusters – Cluster 7 and the rest data. In this research specific analysis is focused on data excluded Cluster 7, and is performed exploratory factor analysis to extract general factors that described internal structure in database. The three variables v14, v22 and v25 (contained MV) are grouped in latent "missing" factor as the result of the initial factor analysis. Three mediation models are performed, using structural equation modeling (SEM) and confirmatory factor analysis, to explain the total, direct and indirect effect of the latent factors in database and variables with MV.*

***Key words:** Mediation Analysis, Exploratory Factor Analysis - EFA, Confirmatory Factor Analysis – CFA, Missing values, Labour Force Survey, Structural Equation Models - SEM.*

Въведение

Настоящото изследване е насочено към анлиз на вървешната структура и взаимовръзки между променливите на базата от данни от наблюдението на работната сила в България през 2007 г. За целта първоначално е приложен обяснителен факторен (Exploratory Factor Analysis – EFA) анализ за определяне на вътрешната факторна сртуктура на данните. След това се прилага потвърдителен факторен анализ за анализ на взаимовръзката между латентните променливи, представящи влиянието на отделните изолирани фактори, като един от тях е латентния фактор на променливите с липсващи стойности. На базата на потвърдителния факторен анализ се оценят преките и косвените, медиаторни влияния на отделните фактори в модела. Анализът се прилага върху две отделни части от базата данни изолирани поради спецификата и по отношение на липсващите стойности (ЛС) и факта, че механизмът на ЛС е неслучайно липсващи, което е показано от автора в [1]. За удобство и краткост в настоящото изследване се представят резултатите само за едната част от базата данни известана като единиците извън Кластър 7, което се обеснява в следващата точка от настоящата публикация.

Наблюдението на работната сила в България - 2007 г. и механизмът на липсващи стойности

В анализа се използват данни от Наблюдение на работната сила в България за цялата 2007 г. Наблюдавани са общо 160 признаци, част от които определят чрез демографски характеристики домакинството и единицата, лицето от домакинството,

обект на изследване, друга част се отнасят за заетоста през последната седмица съответно на основна и допълнителна работа, както и форми на заетост и активност при търсене на работа и др. Част от признаците са изключени от анализа поради няколко причини. От една страна това, че има преходи между въпросите в анкетната карта, т.е ако една единица даде определен отговор на даден въпрос то това води до автоматичното отпадане на въпроси в анкетната карта. От друга страна редица от признаците са технически идентификации на единиците в съвкупността, въведени от изследователския екип, които логически не се отнасят към изследваното явление, а именно липсващите стойности. Интерес представлява появата на ЛС при заетите лица, т.е. се включват лица 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 прави 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. Делът на ЛС при тези признаци е нисък и въпреки това представлява особен интерес за наблюдение поради факта, че механизмът на тяхната поява е НСЛ. Признаците с по-значим дял на 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).*

Появата на липсващи стоности при изследваните променливи следват механизма на не случайно липсващи (НСЛ) стойности, което налага търсенето на модел описващ тяхната поява [1]. След като се прилага процедура на кластъризация по медода К-средни (K-means) се достига до разделяне на данните на две: една по-малка група, наречена Кластър 7 (обем 2290 единици), със средни характеристики близки до 61 часа и една по-голяма група от останалите данни със средни характеристики при променливите с ЛС близки до 41 часа (обем 45951 случая). Заместването на липсващите стойности се прилага при двете групи поотделно. Настоящия анализ се предполага да бъде предхождащ действителното въвеждане на ЛС.

Обяснителен факторен анализ

Първоначално се прилага анализ върху единиците попаднали извън Кластър 7. Както вече беше споменато в настоящата публикация се представят резултати само да тази част от съвкупността. От табл. 1 се вижда, че вътрешната факторна пригодност е слаба, което подсказа, че факторният модел ще бъде лошо обосновен.

Таблица 1: Кайзер-Майер-Олкон тест и Бартлет тест за вътрешна факторна пригодност на данните извън Кластър 7

Kaiser-Meyer-Olkin оценка на извадковата адекватност.		0,611
Bartlett's Test of Sphericity	Approx. Chi-Square	1005395,242
	df	300
	Sig.	0,000

При анализа се използва ортогонална Вирамакс ротация на факторите и екстракция посредством метода на главните компоненти, като при прилагане на неортогонални ротации не се постигат по-добри резултати. Така получените фактори са

некорелирани помежду си. В следвие на екстракцията се обособяват 9 значими фактора, обясняващи 79,62% от вариацията на данните. Елементите на факторите могат да се видят в табл. 2.

Таблица 2: Ротирана компонентна матрица на променливите при единиците извън Кластър 7

Въпроси	Фактори								
	f1	f2	f3	missing	f4	f5	f6	f7	f8
v14				0,965					
v15	0,987								
v16	-0,984								
v17	0,950								
v18									0,961
v19			-0,898						
v20			0,948						
v21			0,946						
v22				0,801					
v23								-0,931	
v24								0,940	
v25				0,942					
v26bgr							-0,674		
v26va									
v26vb						0,936			
v26g						0,935			
v26d1		0,962							
v26d2		0,952							
v26d3		0,908							
v26d4									
v26egr							0,855		
v26j							0,666		
v26z					0,770				
v26i					0,933				
v26k					0,943				

Потвърдителен факторен анализ

При потвърдителния факторен анализ (confirmatory factor analysis - CFA) се използва информацията получена от обяснителния факторен анализ и по-скоро групираните променливи (въпроси) във фактори, които от своя страна се представят като латентни, скрити променливи. Първоначално на анализ се подлагат единиците извън класър 7. Както се вижда от таблица 2, променливите при които се наблюдават ЛС се групират в един общ фактор, наречен “missing”. Друга особеност е обособяването на променливата „v18” в отделен самостоятелен фактор. Това позволява тази променлива да не е свързана с латентна и директно да влияе върху фактора “missing” и съставящете

го променливи. На базата на потвърдителния факторен анализ се оценят преките и косвените (медиаторните) влияния на отделните фактори в модела. Медиаторната променлива е латентната променлива "missing", която представя съгласуваността, взаимовръзката в появата на ЛС при отделните три променливи – v14, v22 и v25[1]. Отделно всеки фактор е директно свързан с всяка една от трите променливи за да може да се оцени прякото влояние.

На граф. 1 е показан модела на факторна взаимовръзка с фокус на влиянието върху v14. За удобство в изследването се нарича **Модел 1**. В овалните форми се представят променливите, които не се наблюдават директно – латентните променливи (f1, f2, f3, f4, f5, f6, f7 и missing), а с правоъгълните тези, които са получени от проведеното наблюдение на работната сила. В Модел 1 факторни променливи са латентните променливи f1,.., f7 и v18, а зависими "missing" и v14, v22, v25. Моделът има не чак толкова добра адекватност: CFI = 0,841; RMSEA = 0,129; PRATIO = 0,703. Това от една страна се дължи на слабата факторна пригодност на данните, а от друга на изключително голямата извадка. В таблица 3 се представя общият, стандартизиран ефект на факторите върху резултативните, ендогенни променливи. Маркираните с „*" коефициенти във всички таблици са незначими. В таблица 4 са представени данните от директния ефект на факторните върху ендогенните проемлниви, а в таблица 5 индиректния, медиаторен ефект на латентната проемнлива „missing" върху v14, v22 и v25.

Таблица 3: Стандартизиран общ ефект от Модел 1

	Фактори								
	f6	**f5**	**f4**	**f3**	**f2**	**f1**	**f7**	**v18**	**missing**
missing	0,033	0,036	0,116	-0,044	0,018	0,236	0,018	0,077	0
v14	**0,03**	**0,033**	**0,108**	**-0,001**	**0,019**	**0,009**	**-0,039**	**-0,026**	**1,077**
v22	0,029	0,032	0,104	-0,039	0,016	0,211	0,016	0,069	0,893
v25	0,029	0,032	0,103	-0,039	0,016	0,21	0,016	0,068	0,889

Таблица 4: Стандартизиран директен ефект от Модел 1

	Фактори								
	f6	**f5**	**f4**	**f3**	**f2**	**f1**	**f7**	**v18**	**missing**
missing	0,033	0,036	0,116	-0,044	0,018	0,236	0,018	0,077	0
v14	**-0,005**	**-0,006**	**-0,017**	**0,046**	**-0,001***	**-0,245**	**-0,058**	**-0,109**	**1,077**
v22	0	0	0	0	0	0	0	0	0,893
v25	0	0	0	0	0	0	0	0	0,889

[1] До подобни заключения се стига и в предходните изследвания на автора. За повече информация виж [1] и [2]

Графика 1: Модел 1 на факторна връзка със зависими величини латентния фактор „missing” и наблюдаваната променлива„v14” при единиците извън Кластър 7

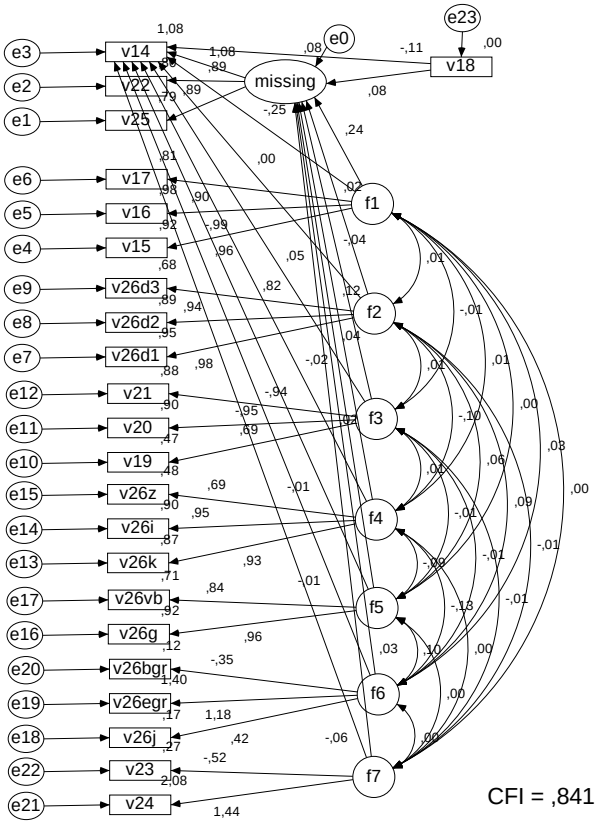


Таблица 5: Стандартизиран индиректен ефект от Модел 1 (медиаторна променлива „missing”)

	Фактори								
	f6	**f5**	**f4**	**f3**	**f2**	**f1**	**f7**	**v18**	**missing**
v14	**0,035**	**0,039**	**0,125**	**-0,047**	**0,02**	**0,254**	**0,019**	**0,083**	**0**
v22	0,029	0,032	0,104	-0,039	0,016	0,211	0,016	0,069	0
v25	0,029	0,032	0,103	-0,039	0,016	0,21	0,016	0,068	0

В **Модел 2** се изследва директния и индиретния ефект на факторите върху v22. Граф. 2 показава структурата на модела. За постигане на максимална адекватност ма модела връзката между f7 и v22 е премахната. Моделът е със характеристики на адекватност: CFI = 0,869; RMSEA = 0,117; PRATIO = 0,707. В таблици 6, 7 и 8 са представени съответно общият, диреткият и медиаторният стандартизирани ефекти върху v22. В Модел 2, както в Модел 1 индогенните променливи са „missing”, v14, v22 и v25, а екзогенни са f1,.., f7 и v18.

Таблица 6: Стандартизиран общ ефект от Модел 2

	Фактори								
	f6	**f5**	**f4**	**f3**	**f2**	**f1**	**f7**	**v18**	**missing**
missing	0,03	0,033	0,108	-0,001	0,019	0,009	0	-0,027	0
v14	0,03	0,033	0,108	-0,001	0,019	0,009	0	-0,027	1,001
v22	**0,033**	**0,034**	**0,115**	**-0,076**	**0,017**	**0,417**	**0**	**0,16**	**0,843**
v25	0,029	0,031	0,101	-0,001	0,018	0,008	0	-0,025	0,94

Графика 2: Модел 2 на факторна връзка със зависими величини латентния фактор „missing" и наблюдаваната променлива„v22" при единиците извън Кластър 7

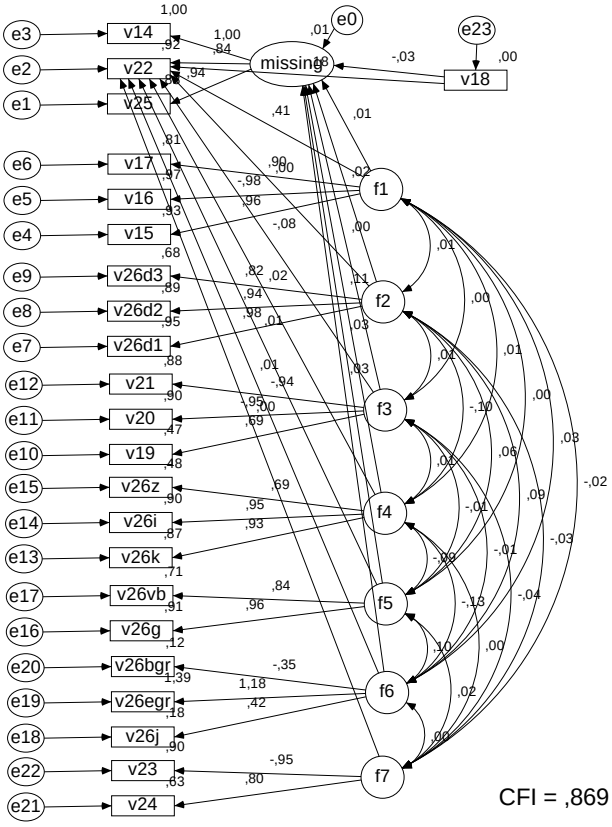


Таблица 7: Стандартизиран директен ефект от Модел 2

	Фактори								
	f6	**f5**	**f4**	**f3**	**f2**	**f1**	**f7**	**v18**	**missing**
missing	0,03	0,033	0,108	-0,001	0,019	0,009	0	-0,027	0
v14	0	0	0	0	0	0	0	0	1,001
v22	**0,007**	**0,007**	**0,024**	**-0,075**	**0,001***	**0,41**	**0**	**0,183**	**0,843**
v25	0	0	0	0	0	0	0	0	0,94

Таблица 8: Стандартизиран индиректен ефект от Модел 2 (медиаторна променлива „missing")

	Фактори								
	f6	**f5**	**f4**	**f3**	**f2**	**f1**	**f7**	**v18**	**missing**
v14	0,03	0,033	0,108	-0,001	0,019	0,009	0	-0,027	0
v22	**0,026**	**0,028**	**0,091**	**-0,001**	**0,016**	**0,007**	**0**	**-0,023**	**0**
v25	0,029	0,031	0,101	-0,001	0,018	0,008	0	-0,025	0

Модел 3 представя влиянието върху v25 (Граф. 3). Адекватността на модела е сравнително слаба, подобно на предходните модели: CFI = 0,824; RMSEA = 0,136; PRATIO = 0,70. В таблици 9, 10 и 11 са представени съответно общият, директният и медиаторният стандартизирани ефекти върху v25.

Таблица 9: Стандартизиран общ ефект от Модел 3

	Фактори								
	f6	**f5**	**f4**	**f3**	**f2**	**f1**	**f7**	**v18**	**missing**
missing	0,03	0,034	0,106	-0,003*	0,019	0,016	-0,041	-0,023	0
v14	0,03	0,034	0,106	-0,003	0,019	0,016	-0,041	-0,023	0,997
v22	0,025	0,028	0,09	-0,003	0,016	0,013	-0,034	-0,019	0,845
v25	**0,028**	**0,031**	**0,097**	**-0,004***	**0,016**	**0,008**	**0,068**	**-0,021**	**0,951**

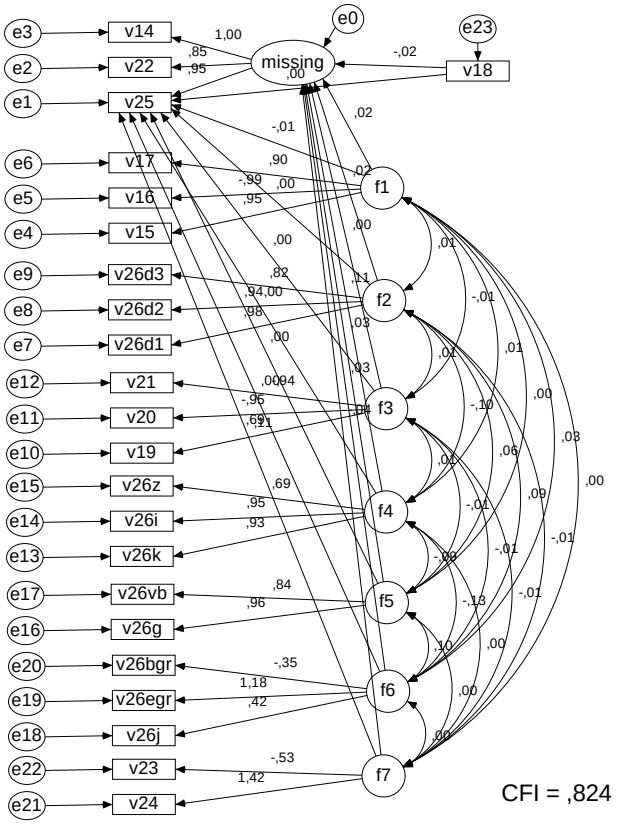
Таблица 10: Стандартизиран директен ефект от Модел 3

	Фактори								
	f6	**f5**	**f4**	**f3**	**f2**	**f1**	**f7**	**v18**	**missing**
missing	0,03	0,034	0,106	-0,003*	0,019	0,016	-0,041	-0,023	0
v14	0	0	0	0	0	0	0	0	0,997
v22	0	0	0	0	0	0	0	0	0,845
v25	**-0,001***	**-0,001***	**-0,004***	**-0,001***	**-0,002***	**-0,007**	**0,107**	**0**	**0,951**

Таблица 11: Стандартизиран индиректен ефект от Модел 3 (медиаторна променлива „missing")

	Фактори								
	f6	**f5**	**f4**	**f3**	**f2**	**f1**	**f7**	**v18**	**missing**
v14	0,03	0,034	0,106	-0,003	0,019	0,016	-0,041	-0,023	0
v22	0,025	0,028	0,09	-0,003	0,016	0,013	-0,034	-0,019	0
v25	**0,028**	**0,032**	**0,101**	**-0,003***	**0,018**	**0,015**	**-0,039**	**-0,021**	**0**

Графика 3: Модел 3 на факторна връзка със зависими величини латентния фактор „missing” и наблюдаваната променлива„v25” при единиците извън Кластър 7



Заключение

В ананлиза на липсващи значения при емпиричните изследвания е важно да се познава моделът на тяхната поява, както и взаимодействието на променливите в базата данни. Това е от основно значение, когато механизмът на поява на ЛС е неслучайно липсващ. В този случай появата на липсващи стойности е във връзка с отделните значения на променливите, при които те се наблюдават. Това означава, че при анализа се налага първо отделянето на различните подгрупи, със специфични прояви на ЛС и едва тогава моделиране на взаимодействията между „засегнатите” промеливи и останалите в базата данни[2].

По тази причина изследването на възможните диретни и индиректтни връзки между променливите с ЛС и останалата база данни, могат само да подобрят тяхното адекватно въвеждане.

[2] На тази основа може да се спомене, че е съвсем реалистична идеята да се използва информация за променливите в наблюдението на работната сила от други тримесечия или години, което се очаква още да подобри анализа. Това предполага да повиши одекватността на описателните модели и оттам да се подобри въвеждането на ЛС.

В настоящият анализ се наблюдават интересни взаимодействия между факторите в базата от данни и променливите с ЛС. И при трите променливи – v14, v22 и v25, се наблюдава значително медиаторно влияние на латентната променлива "missing", която в много от случаите има обратно по знак влияние от директното и общото влияние на факторите. Също така при много слабо или незначимо общо влияние на отделните фактори и при трите модела се наблюдава значимо директно и медиаторно влияние. Това е показателно за моделирането на връзката на променливите с ЛС и останалите проеменливи и фактори в базата данни, което може да има само положително влияние в анализа. От друга страна идеята на латентната променлива "missing" е да се представи формирането на ЛС като общо и свързано между трите променливи, коеот е и в отговор на идеята на механизма на неслучайно липсващи ЛС. Това е една полезна стъпка в посока по-добър анализ на ЛС в наблюдението на работната сила в България.

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. Byrne, B. M. (2010) Structural equation modeling with AMOS: basic concepts, applications, and programming- 2nd ed. Taylor & Francis Group, LLC.
4. Harrington, D (2009). Confirmatory Factor Analysis, Oxford University Press.
5. Little, R. J. A, Rubin, D. B. (2002). Statistical Analysis with Missing Data - 2nd ed., New Jersey: Wiley.
6. MacKinnon, D. P. (2008). Introduction to Statistical Mediation Analysis, Taylor & Francis Group, LLC.