

ЛИПСВАЩИТЕ СТОЙНОСТИ ПРИ НАБЛЮДЕНИЕТО НА РАБОТНАТА СИЛА – 2007 г. В БЪЛГАРИЯ

Гл. ас. Деян Лазаров, БСУ

MISSING VALUES IN LABOUR FORCE SURVEY 2007 IN BULGARIA

Deyan Lazarov

Abstract: *The paper discusses the principles of missing values (MV) patterns and mechanisms. Using Labour Force Survey conducted in Bulgaria 2007 by National Statistical Institute are evaluated the MV mechanisms among employed by permanence of job and is shown that mechanism is not MCAR. In efforts to sift the MAR and MNAR mechanisms K-means cluster procedure is used and two different groups of individuals were found. In this two groups are implemented imputation by EM algorithm that performs very well in this situation.*

Key Words: Missing Values, Missing Mechanism, Missing pattern, Labour Force Survey, Imputation, EM algorithm.

Непрекъснато, в рамките на годината, НСИ провежда репрезентативно наблюдение на домакинствата в страната, известно като НАБЛЮДЕНИЕ НА РАБОТНАТА СИЛА (НРС), и по този начин задоволява нуждата от информация за икономически активното население, заетостта и безработицата в България. В наблюдението участват всички лица на 15 и повече навършени години като на принципа на ротационна извадка се събира информация за редица показатели показващи естеството на явленията и начина на тяхното проявление. Ротационния принцип на извадката се изразява в това, че всяко следващо изследване използва половината от домакинствата от предходното и добавя нови - другата половина, но от същия преброителен участък. От статистическа гледна точка извадката е стратифицирана по области и вид на населеното място, двустепенна гнездова. В първата степен се избират гнезда сред преброителните участъци в България, чрез пропорционален вероятностен подбор. На втората степен се избират домакинства от преброителните участъци попаднали в извадката от първа степен. Обемът на попадналите в извадката домакинства варира в годините като през 2007 г. е бил 18 хил., а от 2008 г. - 19 504.

В настоящия доклад се използват годишни данни от НРС – 2007, предоставени на автора от НСИ с цел научно-изследователска работа и анализ. Всички изводи и заключения са валидни за данните от 2007 г. и не могат да се прехвърлят за предишни и следващи тримесечия и години.

Целта на настоящия доклад е анализ на моделите и механизмите на липсващи стойности в НРС - 2007, който от своя страна ще бъде добра основа за адекватно третиране на самите липсващи данни.

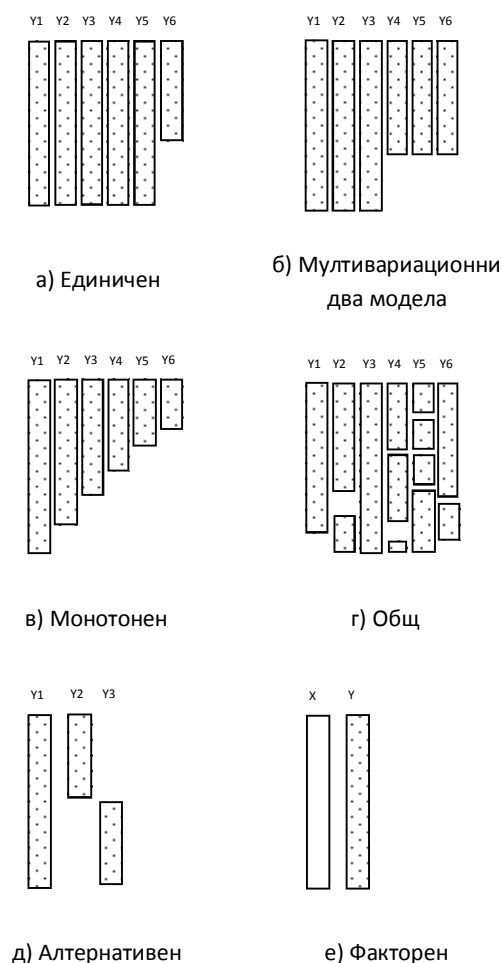
Липсващи стойности.

В настоящия доклад 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 то, по други причини, 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 в следствие на отказ на даден респондент въобще да участва в наблюдението, както и домакинства, които са попаднали в извадката, но не са намерени или са оказали сътрудничество. Авторът

предполага, че подобни домакинства има в значителен обем и проблемът произтича основно от не актуалните избирателни списъци с информация от 2001 г., върху които се гради вероятностната извадка.

Модели на липсващи стойности

Моделите на липсващи стойности могат да се разделят на шест основни групи [2]. Ако в базата данни има единствен признак, при който се наблюдават липси за сметка на напълно въведени значения при останалите признаци се говори за *единичен модел* (Фиг. 1а). В илюстрацията Y_6 е с ЛС, а Y_i за $i = 1, 2, \dots, 5$ са наблюдавани променливи при всички единици. Това е първият установен и описан модел на липсващи стойности. Решения при него се търсят чрез моделиране на променливата с ЛС чрез останалите като задължително условие е добрата балансираност на резултатите по отношение на целите на изследването или наблюдението. Друг случай е когато вместо за един липсва информация за серия на брой признаци – Y_4 , Y_5 и Y_6 (Фиг. 1б). Този модел е известен като *мултивариационни два модела*. Най-често това се свързва с ненамиране на определени единици от извадката и техния отказ да участват в изследването. По този начин имаме тяхна информация за признаците конструиращи извадката като



Фиг. 1 Модели на ЛС

местонахождение, брой лица, възраст и др., но липсва информация за признаците от анкетната карта. Техниката, която се използва в случая за компенсиране на ЛС е притегляне. В настоящата работа няма да може да се изследва този случай, поради база данни, с която разполагаме, където подобни случаи са изключени. *Монотонният модел*

(Фиг. 1в) е типичен за лонгитюдните изследвания, когато дадени единици отпадат по една или друга причина. Така информацията за тях прогресивно намалява при всяка следваща вълна. В НРС също има вълни, но тяхната логика е съвсем различна. Независимо, че половината от домакинствата се запазват при следващата вълна, другата част се подменят случайно, което значително отклонява това наблюдение от логиката на лонгитюдните наблюдения. Типичен и очакван модел на ЛС, в НРС е *общият* (Фиг. 1е). Появата на ЛС има случаен характер и въвеждащите техники са удачно приложими. Специфичен е *алтернативният модел* (Фиг. 1д), при който за едни единици значенията на една променлива се наблюдават, само когато при друга те липсват. *Факторният модел* представя случая, когато в изследването не е наблюдавана променлива, която е в латентна връзка с напълно наблюдавана/и признаци. В примера X липсва, а Y е напълно наблюдавана (Фиг. 1е). Може да се приеме, че X е с 100% липсващи стойности и моделиране на Y с помощта на латентния фактор X и серия строги допускания се прилага като решение в подобни случаи. При НРС, към момента няма да предполагаме наличието на латентни фактори, т.е. разглеждаме факторния модел като възможен.

Механизми на поява на липсващи стойности

Много често, почти автоматично, всеки изследовател е склонен да реагира на проблема с ЛС като взема едно доста крайно и не винаги оправдано решение, че „липсващите стойности са при единици, които по нищо не се различават от единиците регистрирали своите значения на изследваните ги признаци”. Не винаги, особено когато делът на липсващите стойности е голям това е по подразбиране и може да се използва като хипотеза, без да се налага проверка. Приемайки това допускане автоматично се приема идеята, че липсващите стойности са случайна подизвадка или т. нар. „случайно липсващи” от извадката, която е трябвало да бъде наблюдавана.

Нека дефинираме пълната база от данни чрез $Y = (y_{ij})$, която представлява $(n \times K)$ правоъгълна матрица без липсващи стойности, с i -ти ред $y_i = (y_{i1}, \dots, y_{iK})$, където y_{ij} е значението на признака Y_j за i -тата единица.

Нека дефинираме и индикаторът на липсващи стойности – матрицата $M = (m_{ij})$ така, че $m_{ij} = 1$, ако y_{ij} е липсваща стойност и $m_{ij} = 0$, ако y_{ij} е наблюдавана. Т.е:

$$M_{ij} = \begin{cases} 1, y_{ij} - \text{липсват} \\ 0, y_{ij} - \text{наблюдавани} \end{cases}$$

Механизмите на липсващи стойности се описват чрез условното разпределение на M при известно Y , което можем да запишем $f(M|Y, \phi)$, където ϕ отразяват неизвестните параметри. Ако липсващите стойности не зависят от значенията на Y (липсващи или наблюдавани), то е в сила

$$1) \quad f(M|Y, \phi) = f(M|\phi) \quad \text{за всички стойности на } Y \text{ и } \phi$$

данните са **липсващи напълно случайно (ЛНС)**. Нека с $Y_{набл}$ отбележим наблюдаваната част на Y и с $Y_{липс}$ отбележим тази част от Y , която липсва. Механизъм, по-малко ограничаващ от ЛНС е случаят, когато липсващите стойности зависят само от компонента $Y_{набл}$ и не от липсващите стойности. Това може да се запише така:

$$2) \quad f(M|Y, \phi) = f(M|Y_{набл}, \phi) \quad \text{за всички стойности на } Y_{липс} \text{ и } \phi.$$

Този механизъм се нарича **случайно липсващи (СЛ)** данни.

Механизмът е **не случайно липсващ (НСЛ)**, когато разпределението на M зависи от липсващите стойности на Y .

Нека да предположим, че разполагаме с един опростен вариант на разпределение на n единици по значенията на случайната променлива (признак) $Y = (y_1, \dots, y_n)'$. Нека $M = (m_1, \dots, m_n)$, където $m_i = 0$ за единиците, които имат значения на признака, а $m_i = 1$ за липсващите данни. Да предположим, че съвместното разпределение на (y_i, M_i) е независимо при всички единици, т.е. вероятността да е направена регистрация при дадена единица не зависи от значенията на Y или M за останалите единици. Тогава,

$$3) \quad f(Y, M|\theta, \phi) = f(Y|\theta)f(M|Y, \phi) = \prod_{i=1}^n f(y_i|\theta) \prod_{i=1}^n f(M_i|y_i, \phi),$$

където $f(y_i|\theta)$ е плътността на y_i с неизвестни параметри θ , а $f(M_i|y_i, \phi)$ е плътността на разпределение на Бернули за бинарната променлива M_i с вероятност $\Pr(M_i = 1|y_i, \phi)$ y_i да е липсваща. Ако липсващите стойности са независими от Y , то $\Pr(M_i = 1|y_i, \phi) = \kappa$, константа независеща от y_i . В този случай можем да говорим за ЛНС. Ако механизмът зависи от липсващите стойности на y_i , т.е. $\Pr(M_i = 1|y_i, \phi) = f(y_{\text{липс}}, \varpi)$ то той е НСЛ.

Игнорируемост на механизмите на ЛС

Механизмите могат да се нарекат игнорируеми, ако а) са механизъм на случайно липсващи и б) параметрите, които управляват процеса за формиране на данните, като наблюдаеми или липсващи, са независими от параметрите, които трябва да бъдат оценени. Игнорируемостта практически означава, че не е нужно да се моделира механизмът на липсващите стойности, като част от процеса на оценяване.

Въпреки, че от практическа гледна точка, често се слага знак за равенство между ЛС и игнорируемостта поради факта, че на практика условие б) е почти винаги изпълнено, дори и във варианта, когато това не е вярно някои автори твърдят, че методите базирани на пълна игнорируемост работят доста добре. Въпреки всичко трябва да се има предвид, че в значителна степен по-добри резултати биха се получили, ако се моделира механизмът на липсващите данни.

В по-формална форма може да се конкретизира проблема чрез дефинициите дадени от Рубин [2]. Нека отново Y отразява базата данни и $Y = (Y_{\text{набл}}, Y_{\text{липс}})$. Нека с $f(Y|\theta) \equiv f(Y_{\text{набл}}, Y_{\text{липс}}|\theta)$ означим вероятността или плътността на съвместното разпределение на $Y_{\text{набл}}$ и $Y_{\text{липс}}$. Вероятностната плътност на $Y_{\text{набл}}$ може да бъде наблюдавана чрез интегриране на липсващите стойности $Y_{\text{липс}}$:

$$4) \quad f(Y_{\text{набл}}|\theta) = \int f(Y_{\text{набл}}, Y_{\text{липс}}|\theta) dY_{\text{липс}}$$

Авторите дефинират правдоподобие на θ базирано на данните $Y_{\text{набл}}$ да е игнориращо механизмът на липсващи стойности като всяка функция на θ пропорционална на $f(Y_{\text{набл}}|\theta)$:

$$5) \quad L_{\text{игн}}(\theta|Y_{\text{набл}}) \propto f(Y_{\text{набл}}|\theta), \quad \theta \in \Omega_{\theta}$$

Заклучения, относно θ , могат да бъдат базирани на тази функция на правдоподобие $L_{\text{игн}}(\theta|Y_{\text{набл}})$, даваща възможност за игнориране на механизмът на

липсващи стойности. На практика игнорируемите максимално правдоподобни (МП) оценки се получават чрез максимизиране на уравнение (5) по отношение на параметрите θ . Игнорируемите Бейсови оценки за θ , от своя страна, базирани на $Y_{набл}$ се получават чрез обединяването на *a priori* разпределението на $p(\theta)$ за θ и основните заключения, които могат да се направят на апостериорното разпределение:

$$6) \quad p(\theta|Y_{набл}) \propto p(\theta) \times L_{игн}(\theta|Y_{набл})$$

M_{ij} е индикатор на липсващите стойности, показващ кога дадено значение на Y липсва или не. Можем да приемем M като случайна променлива и да представим съвместното разпределение на M и Y . Това разпределение ще има плътност, която може да се представи чрез плътностите на разпределенията на Y и условното разпределение на M при Y със неизвестни параметри ψ . Това разпределение е известно като разпределение на механизма на липсващи стойности и може да се представи в аналитичен вид по следния начин:

$$7) \quad f(Y, M|\theta, \psi) = f(Y|\theta)f(M|Y, \psi), \quad \text{за } (\theta, \psi) \in \Omega_{\theta, \psi},$$

където $\Omega_{\theta, \psi}$ е параметричното пространство на (θ, ψ) .

Механизмът на липсващите стойности е игнорируем от гледна точка на заключения базирани на функция на правдоподобие ако:

а) данните са СЛ и

б) параметрите θ и ψ са независими, в смисъл, че съвместното пространство на параметрите (θ, ψ) е продукт на отделните пространства на θ и ψ .

Механизмът на липсващите стойности е игнорируем от бейсова гледна точка ако:

а) данните са СЛ и

б) параметрите θ и ψ са *a priori* независими, което означава че априорното разпределение има формата:

$$8) \quad p(\theta, \psi) = p(\theta)p(\psi)$$

Неигнорируеми механизми

Ако механизмът на данните не е СЛ, а НСЛ, се твърди, че той е неигнорируем. В този случай се налага той да се моделира за да се получат добри оценки на оценяваните параметри. Един широко разпространен метод, за неигнорируеми липсващи данни е двустъпковият метод за регресионни модели с избор на изместването на зависимата променлива на Хекман. Въпреки всички, трябва много добре да се познава природата на липсващите стойности, за да се правят ефективни оценки на параметрите на разпределенията от данни. Това се налага, защото данните в този случай не съдържат информация за подходящия модел и резултатите са силно чувствителни към избора му.

Тестване на механизмите на липсващи стойности

Основен проблем с липсващите стойности е, че обикновено е неизвестно как е протекъл процеса на формирането им, т.е. механизмът на липсващи стойности остава скрит. Обикновено е необходимо да се правят заключения за разпределението на индикатора на ЛС, които не могат да се проверят [2]. Данните от едно изследване дават

възможност да се тества механизмът на ЛНС, но не и ЛС [1]. Много често изследователите залагат на механизмът ЛС дори и без той да е наистина доказан и действащ. Тук изниква въпроса за грешките, които се получават вследствие на подобно допускане. Според някои автори като Пол Алисън и Габриел Дуран негативното влияние на подобно предположение може да е малко и методите за въвеждане на данни под предположение за механизма СЛ могат да работят не лошо. Това твърдение не би трябвало да се приема абсолютно. Най-малкото то зависи от процента на ЛС в базата данни. Колкото е по-голям дялът на липсващата информация толкова по-важно става правилното идентифициране на механизма на ЛС

Обследването на механизмите на ЛС е не лека задача. При хипотезата на механизма ЛНС има ясна дефиниция за липса на връзка между всички наблюдаваните променливи и появата на липсващи стойности, което ни дава възможност да проверим тази хипотеза като за една променлива (Y_i) с липсващи данни разделим наблюденията при останалите метрирани променливи на две групи, получени при Y_i *набл* и Y_i *липс*. Така ако разполагаме с k -на брой метрирани променливи Y_j ($j = 1, \dots, k$) и $j \neq i$, то за всяка променлива Y_j ще имаме набор от единици със регистрирани значения, които съответстват на Y_i *набл*, които ще означим с $Y_{j|Y_i \text{ набл}}$ и набор от единици и техните значения съответстващи на Y_i *липс*, които ще означим с $Y_{j|Y_i \text{ липс}}$. При статистически значима разлика между средните аритметични на $Y_{j|Y_i \text{ набл}}$ и $Y_{j|Y_i \text{ липс}}$ може да се говори за връзка между Y_i и Y_j . Това дава основание да се отхвърли хипотезата за ЛНС. Друг популярен тест е известен като ЛНС тест на Литъл (Little's MCAR test), представляващ χ^2 тест за механизма ЛНС. Той се основава на сравнението на получените средни стойности при различните подгрупи, образуани от различните модели на липсващи стойности и средните за цялата съвкупност [1]. Ако тестът се окаже незначим то механизма ЛНС може да се приеме за потвърден.

При отхвърляне на хипотезата за механизма ЛНС настъпват трудности с еднозначното определяне на действащия механизъм между СЛ и НСЛ. 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 . Механизмът НСЛ е най-проблемен. При него липсващите стойности са неслучайно разпределени по отношение на всички наблюдения, но вероятността за появяване на липсваща стойност не може да бъде оценена на базата на променливите в модела. Един от подходите в случая е използване на външни източници, минал опит и предположения за поява на ЛС. В настоящата работа ще използваме кластеризация по метода K-means. Предположението е, че ако механизмът е НСЛ, т. е. има значима връзка между появата на ЛС и значенията на променливата, при която се появяват, то би трябвало да може да се „разпознае“ модел в базата данни, при който да имаме значима кластерна разлика (най-вече на средните стойности на променливите в модела).

Наблюдение на работната сила в България 2007 г. – наблюдавани признаци

В изследването са използвани данни за цялата 2007 г., а не по примесечения. Наблюдавани са общо 160 признаци, част от които определят домакинството и единицата от домакинството, обект на изследване, чрез демографски характеристики, 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 настоящия доклад се наложи редуциране на признаците поради няколко причини. От една страна това, че не притежаването 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.

Табл. 1 Общо описание на признаците

Признаци	Бр. единици	Средна аритметична (Mean)	Стандартно отклонение (SD)	Липсващи стойности		Бр. екстремни стойности ^a	
				Брой	%	Долна граница	Горна граница
v14	46596	41,36	6,502	1933	4,0	527	1232
v26egr	48529	5,41	2,603	0	,0	392	0
v25	46762	41,25	6,890	1767	3,6	682	1221
v26bgr	48529	1995,29	86,756	0	,0	91	0
v26va	48529	2,26	6,134	0	,0	0	3999
v22	45987	41,36	6,068	2542	5,2	1501	1321
v21	48529	,05	,838	0	,0	0	239
v20	48528	,04	,706	1	,0	0	202
v26j	48498			31	,1		
v26z	48529			0	,0		
v26i	48529			0	,0		
v26k	48529			0	,0		

В настоящата работа специфичен интерес представлява появата на ЛС при заетите лица, т.е. бяха включени 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 включиха само единици заети на пълно работно време и така признаците обект на анализ се редуцираха до 26, а единиците регистрирали значения по тези признаци 48 529 (Табл. 1). В доклада не са засегнати проблемите на екстремалните стойности. С по-тъмен цвят са означени признаците, при които се наблюдават ЛС. Повече от очевидно е, че делът на ЛС е нисък и представлява интерес за наблюдение в случай, че механизмът на тяхната поява е различен от ЛНС. 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); *Намирането на тази работа свързано ли е с участие на държавно бюро по труда, включително и чрез интернет страницата на Агенцията по заетостта?* (v26g).

Анализ на механизма на ЛС при НРС

За да се изследва механизмът на появяване на ЛС е приложен t тест на Стюdent при различни дисперсии при контролирано влияние на признаците с ЛС (Табл. 2). При този подход ако резултатите при единиците с ЛС не се различават от тези при единиците без ЛС, механизма на появата им може да се прием за ЛНС. В противен случай появата на липсващи стойности е във връзка с променливите в базата данни и механизма е ЛС

или НСЛ. В анализа участват само метрирани признаци, при които може да се изчисли средна аритметична и стандартно отклонение. За всяка променлива двойките са формирани на базата на индикатора присъстващ/липсващ (present/missing).

Видимо е, че при признака v25 – „Колко часа седмично желаете да работите -

Табл. 2 t тест на Студент^a

	v14	v21	v20	v22	v25	v26va	v26d3	v26egr
v14 t	.	,7	1,3	1,1	25,8	-11,3	4,5	-1,4
df	.	2143,4	2235,4	72,1	166,2	40444,9	2144,5	2092,3
P(2-tail)	.	,513	,197	,257	,000	,000	,000	,170
# Present	46596	46595	46596	45914	46595	46596	46596	46596
# Missing	0	1933	1933	73	167	1933	1933	1933
Mean(Present)	41,36	,04	,05	41,37	41,37	1995,09	2,28	5,40
Mean(Missing)	.	,03	,03	39,78	8,35	2000,06	1,71	5,49
v22 t	38,5	-,7	,0	.	43,0	-10,8	6,2	-1,9
df	682,3	2777,6	2883,5	.	829,2	46198,1	2935,2	2829,9
P(2-tail)	,000	,478	,924	.	,000	,000	,000	,060
# Present	45914	45986	45987	45987	45934	45987	45987	45987
# Missing	682	2542	2542	0	828	2542	2542	2542
Mean(Present)	41,78	,04	,05	41,36	41,76	1995,04	2,30	5,40
Mean(Missing)	13,09	,05	,05	.	12,88	1999,78	1,62	5,50
v25 t	.	,4	1,0	-,3	.	-11,2	3,9	-1,3
df	.	1925,7	1994,8	52,0	.	37313,3	1936,4	1898,0
P(2-tail)	.	,684	,321	,735	.	,000	,000	,201
# Present	46595	46761	46762	45934	46762	46762	46762	46762
# Missing	1	1767	1767	53	0	1767	1767	1767
Mean(Present)	41,36	,04	,05	41,36	41,25	1995,11	2,28	5,41
Mean(Missing)	40,00	,03	,03	41,89	.	2000,04	1,76	5,49
v26g t	1,6	12,5	13,0	2,2	1,5	8,5	,0	-,6
df	30,0	48496,0	48497,0	29,3	30,0	30,0	30,0	30,0
P(2-tail)	,126	,000	,000	,033	,146	,000	,975	,531
# Present	46565	48497	48498	45957	46731	48498	48498	48498
# Missing	31	31	31	30	31	31	31	31
Mean(Present)	41,36	,04	,05	41,36	41,25	1996,19	2,26	5,41
Mean(Missing)	39,23	,00	,00	40,53	39,23	580,35	2,23	5,71

a. В анализа не са включени променливи с по-малко от 0.05% ЛС.

общо?” при тази част от единиците, които не са липсващи при v14, средната аритметична е статистически значимо по-голяма от тази при единиците, които липсват при въпрос v14. Аналогично е положението, по отношение на въпрос v14, при признаците v26va – „От кога работите в това предприятие, за този работодател или извършвате самостоятелна дейност? (година)” и v26d3 – “ОБЛАСТ в България”. В Табл. 2 със сиво са маркирани нивата на равнището на значимост при двустранна критична област (P(2-tail)), при които средните се различават значимо при съвкупностите на липсващите и представените случаи.

Проведения анализ еднозначно показва, че липсващите стойности 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 ЛС трябва да бъде съобразен с различията между отговорилите и неотговорилите и факторната зависимост между задавания въпрос и не получаването на отговори. Допълнителен анализ, основан на съпоставката на ЛС при различните признаци показва, че меделът може да се разглежда като множество единични, т. е. няма връзка в появата на ЛС сред различните променливи. Това до известна степен улеснява компенсирането на ЛС при различните променливи. Най-важният извод, който трябва да се направи, е че ще бъде неоправдано механичното елиминиране на ЛС в базата данни. По този начин анализът ще се лиши от специфични подсъвкупности от единици, които значимо влияят на обобщените резултати за признаците.

От Табл. 3 се вижда, че връзката между признаците v14, v22 и v25 е изключително висока. 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. 4).

Табл. 3 Кроскорелации

	v14	v22	v25	v26va	v26d3	v26egr	v20	v21
v14	1,000							
v22	,922	1,000						
v25	,981	,898	1,000					
v26va	,015	,014	,014	1,000				
v26d3	-,021	-,024	-,021	,014	1,000			
v26egr	,024	,033	,021	,015	,086	1,000		
v20	-,005	,060	-,003	,003	-,011	,002	1,000	
v21	-,005	,059	-,002	,003	-,010	,009	,889	1,000

Табл. 4 Основни характеристики на разпределенията на променливите v14, v22, v25

		V14 Колко часа седмично работите ОБИКНОВЕНО на ОСНОВНАТА РАБОТА?	V22 Колко часа общо сте работили през МИНАЛАТА СЕДМИЦА на ОСНОВНАТА РАБОТА?	V25 Колко часа седмично желаете да работите общо?
N	Valid	46596	45987	46762
	Missing	1933	2542	1767
	Mean	41,36	41,36	41,25
	Std. Error of Mean	,030	,028	,032
	Median	40,00	40,00	40,00
	Mode	40	40	40
	Std. Deviation	6,502	6,068	6,890
	Skewness	-1,468	,898	-1,903
	Std. Error of Skewness	,011	,011	,011
	Kurtosis	21,252	8,741	19,865
	Std. Error of Kurtosis	,023	,023	,023
	Minimum	0	0	0
	Maximum	96	96	96

За проверка на модела НСЛ използваме последователни кластерни модели с нарастващи число на кластерите. Началото на анализа е при 2 кластера, а края при 11 кластера. При направения анализ се установи, че при групирането на единиците в 8 кластера в един от тях се получават центрове при променливите с ЛС (v14, v22, v25), значимо различни от останалите. Ако в останалите кластери центровете съответсват на средните при тези променливи, т. е. около 41 часа, то в **кластер номер 7 в анализа**, центровете при тези променливи са със стойност 61 часа. При направения анализ ЛС са отстранени чрез процедурата pairwise, т.е. използват се всички налични данни за оценка на разстоянията в кластерите, за всеки две променливи в анализа. Така размерът на този кластер е втори по големина с обем 2290 единици, като най-големият се състои от 45951 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 други характеристики по анализираните признаци (v14, v22, v25) от останалите. 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 ще е НСЛ.

При анализа на ЛС в кластерите се установи, че сред единиците от кластер 7 има изключително висок процент липсващи значения (табл. 5). Това налага при тяхното въвеждане да се подходи специфично, те да бъдат отделени и техният анализ да се базира на информацията от кластер 7.

Табл. 5 Липсващи стойности сред променливите в кластер 7

	Брой	Средна аритметична (Mean)	Стандартно отклонение (Std. Deviation)	Липсващи стойности		Бр. екстремни стойности ^a	
				Бр.	%	Долна граница	Горна граница
v14	1311	61,30	7,599	973	42,6	53	273
v20	2284	,16	1,839	0	,0	.	.
v21	2284	,13	1,404	0	,0	.	.
v22	1317	61,38	7,938	967	42,3	52	282
v25	1311	59,88	8,809	973	42,6	140	248
v26va	2284	2002,06	4,920	0	,0	104	0
v26d3	2284	,79	3,569	0	,0	.	.
v26egr	2284	5,42	2,598	0	,0	0	0

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

Компенсиране на ЛС

Един от подходите за компенсиране на ЛС в случая е използването ЕМ алгоритъма (Expectation - Maximization Algorithm). Той минава през две основни стъпки, които се повтарят една след друга, докато се достигне до липса на промяна в основните параметри на анализирания променливи. Първата стъпка е оценката на средната и ковариацията на променливата с липсващи стойности на базата на последователни единични или множествени линейни регресии чрез останалите признаци в базата данни. Така се получават първоначални оценки за липсващите значения. Следващата стъпка – максимизиращата – използва първоначално попълнените редове за да оцени наново средните и ковариациите като използва информация за остатъците при регресионните модели. Така получените оценки след тази стъпка се подлагат отново на регресионна оценка – Е стъпка и т.н.

Съобразно направените изводи в предната част се използва следния подход при прилагане на ЕМ алгоритъма. За единиците принадлежащи към кластер 7 се прилага самостоятелен анализ, изхождайки от констатацията, че те имат специфични характеристики. Всички останали единици се обработват заедно. Така след въвеждането на значения при ЛС от кластер 7 дава резултати поместени в табл.6. В табл. 7 са представени сравнителни характеристики за единиците при които са направени въвеждания и останалите единици от кластера. Липсващите стойности при останалите кластери са поместени в табл. 8. Вижда се ЛС са се разделили почти по равно между двете съвкупности.

Табл.6 Характеристики на разпределенията на променливите v14, v22, v25 от кластер 7 след прилагането на ЕМ алгоритъма за въвеждане на ЛС

	N	Размах	Min.	Max.	Средна аритметична		SD	Дисперсия	Асиметрия		Ексцес	
						Ст. грешка				Ст. грешка		Ст. грешка
V14	2284	56	40	96	61,25	0,121	5,806	33,704	1,827	0,051	8,169	0,102
V22	2284	60	36	96	61,28	0,126	6,038	36,460	1,595	0,051	7,898	0,102
V25	2284	56	40	96	59,56	0,141	6,729	45,277	0,723	0,051	6,210	0,102

Табл. 7 Сравнение на средните аритметични между единиците с вмъкнати значения и останалите

Средни аритметични	v14	v22	v25	Стандартни отклонения	v14	v22	v25
Вс. значения	61,30	61,38	59,88	Вс. значения	7,599	7,938	8,809
ЕМ	61,25	61,28	59,56	ЕМ	7,568	7,939	8,779

Табл. 8 Липсващи стойности сред променливите в базата данни без кластер 7

	Брой	Средна аритметична (Mean)	Стандартно отклонение (Std. Deviation)	Липсващи стойности		Бр. екстремни стойности ^a	
				Бр.	%	Долна граница	Горна граница
v14	45285	40,78	5,476	960	2,1	528	44
v20	46245	,04	,755	0	,0	0	214
v21	46244	,04	,652	1	,0	0	179
v22	44670	40,77	4,888	1575	3,4	1871	105
v25	45451	40,71	6,025	794	1,7	682	139

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

Прилагането на ЕМ въвеждане на ЛС към единиците от табл. 8 може да се види в табл. 9, а сравнителни характеристики за единиците при които са направени въвеждания и останалите единици от съвкупността са представени в табл. 10.

Табл.9 Характеристики на разпределенията на променливите v14, v22, v25 от базата данни без кластер 7 след прилагането на ЕМ алгоритъма за въвеждане на ЛС

	N	Размах	Min.	Max.	Средна аритметична		SD	Дисперсия	Асиметрия		Ексцес	
						Ст. грешка				Ст. грешка		Ст. грешка
V14	46245	70	0	70	40,67	5,764	33,224	-4,310	0,011	29,994	0,023	46245
V22	46245	84	0	84	40,35	6,051	36,613	-2,325	0,011	10,809	0,023	46245
V25	46245	76	0	76	40,71	5,976	35,712	-4,154	0,011	28,808	0,023	46245

Табл. 10 Сравнение на средните аритметични между единиците с вмъкнати значения и останалите от базата данни без кластер 7

Средни аритметични	v14	v22	v25	Стандартни отклонения	v14	v22	v25
Вс. значения	40,78	40,77	40,71	Вс. значения	5,476	4,888	6,025
ЕМ	40,67	40,35	40,71	ЕМ	5,813	6,122	6,026

Заклучение

Основен момент в работата е идентифицирането на кластер 7 и изолирането на единиците попадащи в него. Това само по себе си е особен, но работещ подход при тестването и идентифицирането на механизма НСЛ. Предимство в анализа е големият обем на базата данни, която позволява да се изолират дори и малки като дял подсъвкупности, които обаче са обременени с голям процент ЛС. Тази идея, разбира се, има нужда от последващо развитие и подлагане на тестове, за да се определи като масово приложима. Тук беше използван единствено похода на K-means кластеризацията. Разумно е да се конкурира този подход с други, за да се определи оптималния алгоритъм.

От друга страна присъждането на значения на ЛС посредством ЕМ алгоритъма дава изключително добър резултат. Разпределенията се възпроизвеждат добре и една от заплахите на въвеждащите процедури – редуцирането на дисперсията – характерно за моделно базираните процедури, не е на лице. При анализа на кластер 7 и останалите единици тестът на Литъл за ЛНС показва статистическа значимост, което в случая подчерта, че в двете съвкупности механизъмът на ЛС е случайно липсващи. Това от своя страна още веднъж наложи използването на ЕМ алгоритъма като въвеждаща процедура.

Като бъдеща работа е любопитно прилагането на многослоен перцептрон като въвеждаща процедура при механизъм СЛ. Процедурата при определянето на теглата на активиращите функции на невронната мрежа е също итеративно, както при ЕМ алгоритъма, но като цяло моделът на връзката не е линеен за разлика от регресионните

зависимости заложени при ЕМ, което може да бъде значително предимство в общия случай.

Литература:

1. Little, R.J.A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* vol. 83, No 404
2. Little, R.J.A, Rubin, D.B. (2002). *Statistical Analysis with Missing Data* - 2nd ed., New Jersey: Wiley.
3. Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Survey*. New York: Wiley.
4. Rubin, D. (1996). Multiple imputation after 18+ Years. *Journal of the American Statistical Association* 91 (June): 473–89.
5. Scheffer, J. (2002), Dealing with Missing Data, *Research Letters in the Information and Mathematical Sciences* 3, 153-160.
6. Заетост и безработица (годишни данни) 2007, НСИ, София 2008.