Методи за анализ и въвеждане на липсващи индивидуални данни

гл. ас. Деян Лазаров,
Бургаски Свободен Университет

1. ВЪВЕДЕНИЕ

В процеса на всяко емпирично изследване изследователският екип се сблъсква с негативите от появата на липсващи индивидуални сведения за някои от единиците на изследваната съвкупност. Причините могат да бъдат от различен характер, включително да са предизвикани от недостатъци в различни етапи на подготовката и изпълнението на изследването. В съвременната практика се разглеждат основно три класа липсващи данни:

- липса на обхват на съвкупността, от която се избира извадката;
- отказ или невъзможност на единиците, попаднали в извадката, да сътрудничат;
- отказ на единиците, обект на изследване, да дадат отговори на някои въпроси или загуба или пропуск да се регистрира необходимата информация.

Проблемите, възникващи от подобни събития, влияят пряко върху качеството на информацията и точността на оценките. В много случаи, особено при липсващи данни от първите два класа, се прибягва до заместване на единиците с други, които от своя страна имат различни и понякога специфични характеристики, което води до изместване на основните характеристики на съвкупността. Така получените резултати се превръщат в непредставителни за изучаваната съвкупност (Bogdanov, 1988). Проблемът на липсващите стойности се състои в това, че значения, които трябва да бъдат наблюдавани 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" (Rubin, 1987). 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 мощността на статистическите методи и се получават изместени оценки (Demirtas, 2005).

Настоящото изследване засяга проблемът на липсващите данни от трети клас, а именно когато липсват отговори на определени въпроси или загуба или пропуск да се регистрира необходимата информация. То е разделено в двее части като настоящата публикация е първата от тях. Термините липсващи данни и липсващи стойности (ЛС) се приемат за напълно заменяеми и се използват като синоними.

В изследването се обхваща базата, върху която се гради анализа на ЛС, в контекста на механизмите на поява на ЛС, както и някои „по-прости" методи за елиминиране и въвеждане на ЛС. Целта е да се направи един достатъчно обхватен, но без претенции за пълна изчерпателност, обзор на възможностите за анализ при т.н. единични въвеждащи процедури. Представени са техните предимства и недостатъци и генезиса на идеята за множествено въвеждане, предложена в края на публикацията.

2. МЕХАНИЗМИ НА ПОЯВА НА ЛИПСВАЩИ СТОЙНОСТИ

Начинът на проява на ЛС определя механизмите на тяхната поява и се разглежда в контекста на тяхната зависимост от признаци в базата от данни. Доналд Рубин (Rubin, 1976) за първи път формализира теорията, като слага солиден фундамент, чрез точно дефиниране на различни допускания в тази посока.

За представяне на различните механизми е наложително да се въведат някои дефиниции. Нека пълна база от данни е $Y = (y_{ij})$, която представлява $(n \times K)$ правоъгълна матрица без липсващи стойности, с *i-ти* ред $y_i = (y_{i1}, ..., y_{iK})$, където y_{ij} е значението на признака Y_j за *i-мата* единица. Нека **индикатор на липсващи стойности** е правоъгълната матрица $M = (m_{ij})$, така че $m_{ij} = 1$, ако y_{ij} е липсваща стойност и $m_{ij} = 0$, ако y_{ij} е наблюдавана. т.е:

$$M_{ij} = \begin{cases} 1, y_{ij} - \textit{липсват} \\ 0, y_{ij} - \textit{наблюдавани} \end{cases} \quad (2.1)$$

Механизмите на ЛС могат да се опишат чрез условното разпределение на *M* при дадени значения на *Y,* а именно $f(M|Y,\phi)$, където ϕ отразяват неизвестните параметри на разпределението на *M*.

Ако ЛС в базата от данни не зависят от значенията на *Y,* липсващи или наблюдавани, то

$f(M|Y,\phi) = f(M|\phi)$ за всички стойности на *Y*, (2.2)

В този случай ЛС са случайна подизвадка на извадката формираща базата от данни и механизма на тяхната поява се определят като **липсващи напълно случайно (ЛНС).**

Нека $Y_{набл}$ е наблюдаваната част от данните *Y, а* $Y_{липс}$ е тази част от *Y,* за която няма регистрирани стойности, т.е. липсва. Ако появата на ЛС зависи само от $Y_{набл}$ и не от $Y_{липс}$:

$$f(M|Y,\phi) = f(M|Y_{набл},\phi)$$

за всички стойности на $Y_{липс}, \phi$, (2.3)

тогава този механизъм е **случайно липсващи (СЛ)** данни. Механизмът на СЛ стойности е по-малко ограничаващ в сравнение с ЛНС по отношение на условията за съществуване. Може да се каже, че механизма СЛ е по-общ от ЛНС, а също така, че ЛНС е под случай на СЛ.

Особен е случаят, когато разпределението на *M* зависи от ЛС на *Y*. Този механизъм е известен като **не случайно липсващи стойности (НеСЛ).** За представяне на същността на този механизъм нека се използва само един признак в базата от данни. При *n* единици, попаднали в извадката, тяхното разпределение по този признак е $Y = (y_1,...,y_n)'$. В този случай индикаторната матрица е $M = (m_1,...,m_n)$, където $m_i = 0$ за единиците, които са дали своите значения на признака, а $m_i = 1$ за пипващите данни. Съвместното разпределение на (y_i, M_i) е независимо при всички единици, т.е. вероятността да не е направена регистрация при дадена единица не зависи от значенията на *Y* или *M* за останалите единици. Тогава,

$$f(Y,M|\theta,\phi) = f(Y|\theta)f(M|Y,\phi) = \prod_{i-1}^{n} f(y_i|\theta) \prod_{i-1}^{n} f(M_i|y_i,\phi) \quad (2.4)$$

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 механизъм ЛНС. Ако механизмът зависи от липсващите стойности на y_i, т.е $k = f(y_{липс},\varpi)$ то той е НеСЛ.

Механизмът НеСЛ води след себе си сериозни последици, ако не бъде идентифициран като такъв. Почти винаги се наблюдават изместванията на основните характеристики на разпределенията, чиято посока обаче не може да бъде установена без задълбочен анализ.

Доста често в литературата по въпросите на ЛС се срещат и термините игнорируеми (ignorable) и неигнорируеми (nonignorable) механизми.

Игнорируеми и неигнурируеми механизми. Механизмите могат да се нарекат игнорируеми, ако са изпълнение следните условия:

а) механизмите са ЛНС или СЛ и

б) параметрите, които управляват процеса на проява на данните, като наблюдаеми или липсващи, са независими от параметрите, които трябва да бъдат оценени. Формализация на условията за игнорируемост се предлага от Рубин (Little и Rubin, 2002). Игнорируемостта практически означава, че не е нужно да се моделира механизма на липсващите стойности, като част от процеса на оценяване на самите тях (Allison, 2002). От практическа гледна точка, често се слага знак за равенство между механизма СЛ и игнорируемостта (Allison, 2002; Scheffer, 2002; Durrant, 2005; Howell; Frick и Grabka, 2004) поради факта, че условие б) е почти винаги изпълнено. Дори и в случаите, 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 случай, че механизмите са неигнурируеми трябва анализа на ЛС задължително да премине през фаза, която описва, моделира процеса на тяхната поява и едва след това да се премине към

въвеждане. По този начин механизмите на ЛС определят и различните подходи за анализ на самите ЛС.

3. ЕЛИМИНАЦИОННИ ПОДХОДИ В ТРЕТИРАНЕТО НА ЛС

Поредно елиминиране[1] (ПЕ). Това е може би най-разпространения подход за третиране на липсващите стойности в дадена база данни[2]. Този подход е известен още като анализ на пълна база данни. При него решението на проблема с ЛС се намира в отстраняването на всички единици, при които има такива, независимо от признака. Това е доста спонатнно решение и може да се определи също като доста крайно и не винаги оправдано решение. Използвайки подобен подход имплицитно се определя, че механизма е ЛНС и липсващите стойности са при единици, които по нищо не се различават от единиците регистрирали своите значения по изследващите 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. Оказва се, че в повечето ситуации се налагат сериозни усилия за обосновка на тази хипотеза и както отбелязва Пол Алисън (Allison, 2002) „Дори и статистици понякога не са наясно или са несигурни относно това твърдение".

Въпреки всичко поредното елиминиране е приложим подход и има своето място в случай на игнорируеми механизми. Той може да се характеризира с простота и това е основно предимство пред останалите методи за анализ на ЛС. Анализирайки свойствата на оценките получени след прилагане на поредно елиминиране, може да се заключи, че този подход трябва да бъде използван основно само когато механизма на поява на ЛС е ЛНС (Allison, 2002). Тогава ПЕ дава ефективни и неизместени оценки. В случай, че механизма на поява на ЛС е СЛ то трябва да се въведат определени условия за приложимост на подхода. Ако предстои анализ на регресионна връзка от типа $Y = b0 + bi\ Xi + e$, където Xi са независимите променливи в модела, а Y е зависимата променлива, то оценките на параметрите в модела ще бъдатат наизместени ако евентуалните ЛС при Xi ***не зависят от*** Y. В случай, че това условие не е изпълнено, т.е. $Pr(Xmiss) = Pr(Xmiss|Y)$, ***то оценките от регресионния анализ ще бъдат изместени*** (Allison 2002)**.**

Друг основен недостатък при ПЕ произтича от редуцирането на обема на извадката, което независимо от механизма на ЛС неизменно води до намаляване на мощността на критериите при проверката на хипотези, особено когато извадките по дизайн са малки.

Елиминиране по двойки[3] (ЕД) Това е друг подобен подход известен още като „анализ на наличната база[4]". При него отново се елиминират единици с ЛС, но в зависимост от конкретен анализ и участващите в него променливи. За всеки анализ се използват всички единици, за които има регистрирани значения, независимо, че при друг анализ върху същата база от данни може да се използват друг набор от единици с „пълна информация". Например, ако дадена база от данни е получена на базата на извадка с обем n единици, но при различните признаци Yi има различна процент на ЛС, то при определен анализ (да кажем корелационен анализ) включващ част от тези признаци (например Y1, Y3, Y6, Y7 и Y9), то за изчисляване на параметрите на анализа се използва цялата налична информация от единиците при кореспондиращите признаци[5]. Използването на цялата налична информация за оценката на дадени параметри в анализ може да се определи като възможност да се повиши мощността на статистическите анализи в сравнение с ПЕ. Въпреки това ЕД има определени недостатъци, които ограничават неговото използване. ЕД изисква механизъм ЛНС за да може оценките от анализите да не бъдат изместени. Освен това при ЕД има допълнителни, специфични условия за коректно прилагане. Почти при всички оценки от мултивариационните статистическите

[1] Listwise deletion
[2] Complete-case analysis
[3] Pairwise deletion
[4] Available-case analysis
[5] корелацията между Y1 и Y6 се изчислява на базата на единиците, които имат едновременно наблюдавани стойности при двата признака; корелацията между Y3 и Y6 се изчислява на базата на единиците, които от своя страна имат едновременно наблюдавани стойности при тези два признака и т.н.

анализи се използва информация за вече изчислени на преден етап оценки от единичните разпределения. Дори за да се оцени корелацията при два признака се използва информация за средните аритметични и дисперсиите при отделните признаци. Това от своя страна означава, че за да се оценят параметрите на единичните разпределения се използва цялата налична информация от тези разпределения, а за да се оцени параметър на съвместното разпределение се използва наличната съвместна информация. Така се оказва, че следвайки логиката на ЕД всяка оценка може да бъде получена на базата на различна подизвадка от данни с различен обем и това може да доведе до сериозно нарушаване на логиката на самите оценки и оттам до съвсем грешни такива[6].

Липсата на сравнима извадкова основа също води до проблеми с оценките на стандартните грешки. Обема на извадката е основен компонент при оценките на всяка стандартна грешка, а в случая този обем не е ясен при дори една обикновена регресия. Някои софтуерни продукти прилагат осредняване на извадковия обем при подобни оценки, но това често води до подценяване на някои и надценяване на други оценки. Тази осреднена оценка на обема на извадката става значително по-сложна при многомерния анализи. Например при моделите със структурни уравнения обемът на извадката едновременно участва в максимизирането на точността на стандартните грешки и в оценката на адекватността на модела (Enders C., 2010).

Тези недостатъци на ЕД, ограничават неговото приложение в практиката и налагат този подход да бъде пренебрегван за сметка на ПЕ дори.

Като алтернатива на процедурите по елиминиране на ЛС са методите на въвеждане. Тяхната логика е основана на възможността вместо да бъдат отстранени единиците с ЛС, самите ЛС да бъдат „елиминирани" чрез въвеждане на значения. В зависимост от това дали въведеното значение за дадена ЛС е едно или са повече се разграничават различните подходи за единично и множествено въвеждане.

[6] Например може да се получат корелационни коефициенти по-големи от 1 или по-малки от -1.

4. МЕТОДИ ЗА ВЪВЕЖДАНЕ ПРИ ИГНОРИРУЕМИ МЕХАНИЗМИ

4.1 Методи, използващи въвеждане на единични стойности

Това са методи за „запълване" на липсващите данни чрез подходящи стойности с цел да се получи пълна база данни. Методите могат да се разделят основно на два типа – детерминистични и стохастични. Детерминистичните методи се основават на моделиране на липсващите стойности чрез познатите характеристики на базата данни. При тях винаги се получават едни и същи резултати за ЛС при едни и същи характеристики на съвкупността. При стохастичните методи допълнителното добавяне на случайност дава възможност да се получават различни стойности. Обикновено методите за въвеждане използват различни допълнителни променливи, които са в корелация с променливата, при която се наблюдават ЛС и на тази основа се извлича информация за разпределението на тази проеменлива.

Основната причина да се използва въвеждане на данни е желанието да се намали изместването на оценките причинено от наличието на липсващи данни. Очаква се, че след въвеждането базата от данни ще бъде възстановена в състояние на пълна и стандартните статистически методи за анализ ще могат да бъдат нормално използвани. В този процес изборът на процедура за въвеждане се оказва съществен и трябва да бъде съобразен с последяващия анализ на данните, защото ако при въвеждащата процедура не се използва информация от някоя от променливите в базата от данни, то в последващия анализ е много вероятно връзката с нея да бъда подценена. Възможно е, също така, при неизследване на смущаващото влияние на външни променливи със значимо влияние, вместо да се намали, да се увеличи изместването на основните характеристика на съвкупността при използването на методите за въвеждане на ЛС (Kalton и Kasprzyk, 1982; Kalton, 1983; Särndal, Swensson и Wretman, 1992; Little и Rubin, 2002).

Не трябва се забравя обаче, че има сериозна разлика между истинските, действителни стойности и въведените такива. Така, ако се използва детерминистичен метод за въвеждане

на ЛС, се изпада в сериозно подценяване на вариацията в базата от данни. (Rubin 1987). Основата за това подценяване е фактът, че при въведените стойности липсва случаен компонент и на практика за оценката на общата вариация се използва само наличната информация, при единиците с регистрирани значения. Компенсиране на този недостатък на детерминистичните методи се търси чрез допълнителното включване на случаен компонент във въвеждащите процедури.

Друг особено проблемен аспект на оценките на вариацията в базата от данни е, че въведените стойности се третират като реални. Съществува един специфичен компонент на вариацията, който произтича от факт, че ЛС са придружени с огромна несигурност относно техните действителни значения. Стандартните техники за оценка на вариацията са неадекватни в следствие на процеса на въвеждане на ЛС. Стандартните отклонения са нереално малки, доверителните интервали тесни, а емпиричните тестови стойности силно завишени (Rubin 1987; Rao и Shao, 1992). В литературата съществуват някои решения на задачата като специално внимание ще бъде обърнато на множественото въвеждане.

Метод на независимата средна. Методът е основан на въвеждане на стойността на средната аритметична при всички липсващи значения. Вариация на метода е да се използват групови средни, като разделянето на съвкупността на групи да се базира на избрана, обясняваща променлива. Недостатъците на подобен подход или процедура са свързани с факта, че разпределенията на признаците стават „компресирани" т.е., те са лишени от нормалната си вариация и връзките между признаците могат да бъдат силно променени (Kalton, 1983; Lessler и Kalsbeek, 1992; Little и Rubin, 2002). Въпреки всичко, подобни прости методи за въвеждане са много популярни в социалните науки (Jinn и Sedransk, 1989; Allison, 2002), ако и да са крайно неадекватни при решаването на каквито и да са проблеми, независимо от механизма на ЛС. Дори ако изберем най-благоприятния вариант и предположим, че механизма е ЛНС то при въвеждане на ЛС чрез средната аритметична няма да се получи изместване на общата средна за даденото разпределение, но вариацията, асиметрията и ексцесът ще бъдат силно повлияни и изместени. Въпреки всичко в много голяма степен те са единствените стандартни приложения към софтуерните пакети за обработка на бази данни.

Регресионно и стохастично регресионно въвеждане. Друг широк клас от методи за въвеждане на ЛС е регресионното въвеждане или използването на регресионни модели за подобни цели (Little и Rubin, 2002, Durrant G.B., 2005, Enders C, 2010). В теорията може да се срещне като *Прогностичното регресионно въвеждане*, известно още като *детерминистична регресионна зависимост* или още въвеждане на базата на *зависимата средна*. Регресионния модел за въвеждане се изгражда като променливата с ЛС *yi* се регресира (корелира) с допълнителните променливи *xi, при които не се наблюдават ЛС*. Оценените стойности чрез модела се използват за въвеждане на липсващите значения на *Y*. Обикновено линейни регресии се използват за метрирани променливи, докато логистичните регресии се предпочитат при категорийни данни. Недостатъкът на подобен подход е, че нарушава, изменя вида на разпределението на *Y* и корелациите между променливите, които не се използват в регресионния модел за въвеждане. Оказва се, че моделът, на базата на който се прави въвеждането и променливите включени в него, са от изключително значение. Той детерминира последващите анализи и силно променя степента и силата за свързаност на Y с останалите променливи, изменя размера на стохастичните грешки и довежда до по-големи грешки от I род (Enders C, 2010).

За компенсация на тези недостатъци се използва друг подход известен днес като **стохастична регресия**. Методът е известен още като *рандомизирано регресионно въвеждане* и се разбира въвеждане на базата на зависимото разпределение на Y спрямо X_i, но за компенсиране на несигурността (в смисъла на Rubin) на въведените стойности, например при линейна зависимост между Y и X, е чрез добавянето на елементи от разпределението на остатъците към въведените стойности. Остатъците могат да се получат по различни начини. Един е чрез случаен избор от моделирано нормално разпределение, отново

като цяло или на подгрупи. Друг използван подход е разпределението на действителните остатъчните елементи, получени от регресията на променливите при единиците от базата данни, за които има регистрирани значения. Добавянето на остатъците към въведените стойности възстановява вариативността на данните и ефективно елиминира изместването в следствие на стандартното регресионно въвеждане. Стохастичната регресия е единствения подход от т.н. стандартни методи, който дава не изместени оценки на разпределенията при механизъм СЛ (Enders C, 2010).

Предимството на регресионното въвеждане е във възможността да се използват както вариационни така и категорийни признаци. Методът работи изключително добре при метрирани признаци, особено ако те са в значима корелация с екзогенните променливи. Въпреки всичко, въведените стойности са значения получени на базата на модела (с или без корекция за несигурност), а не реално наблюдавани значения, както е в hot deck методите. Те могат да се нарекат псевдо случайни значения – такива които изглеждат реалистични в същия порядък на реалното разпределение, но съвсем не е ясно че могат да се наблюдават в реалността. Друг значим проблем може да се окаже конструирането на регресионната връзка и включените определящи променливи (Schenker и Taylor, 1996). Ако регресионния модел не е добре обусловен, което е проблем на повечето параметрични методи, неговата определяща (прогностична) сила е слаба и това влияе върху оценката на въведените стойности (Little и Rubin, 2002).

Донорски методи[7]**.** Подходът при тези методи е ориентиран към използването на действително наблюдавани в изследването значения на променливите. Така базата от наблюдавани стойности се използва като донор за липсващите значения, които се избират от тях. Подобни методи за вмъкване са известни като *донорски* методи (Kalton и Kasprzyk, 1982; Little, 1986; Lessler и Kalsbeek, 1992). Решаването на задачата минава през определяне на донорските значения. Един от подходите е използването на случайна извадка от значения на дадена променлива. Алтернативен подход е разделянето на пълната база данни на класове и прилагане на случаен избор на донорски значения в самите класове. Подобни класове могат да бъдат определени на основата на многомерно разпределение на наблюдаваните стойности при променливите без ЛС. Предимството на подобен подход е използването на действително наблюдавани „реални" значения на променливите при процеса на вмъкване. Това е основна причина те да се предпочитат особено при работа с категорийни променливи. Донорските методи са не-параметрични или полу-параметрични и нямат ограниченията относно информация за признаковите разпределения. Това е особено важно при разпределения с асиметрия или в случаите, когато данните се получават след закръгляне или когато оценката на честотите на разпределението са заложени в изследователската задача. В следствие на прилагането на hot deck методите вмъкнатите значения ще имат същата форма на разпределенията като наблюдаваните данни (Rubin, 1987). Едно от най-големите ограничения е необходимостта да се работи с големи извадки за да могат да се получат добри резултати. Въпреки, че за въвеждане се използват „истински" значения от базата данни, те не са действителните стойности на отделните единици. Вариацията, породена от този факт, остава подценена и така общата вариация на оценките ще бъде подценена.

Въвеждане чрез метода „Най-близък съсед". Методът „Най–близък съсед", още наричан съответствие по функцията на разстоянията[8] е от типа на донорските методи, където донорът се подбира на базата на минимизиране на функция на разстоянието (Durrant G.B., 2005, Enders C, 2010). В метода се дефинира подходяща характеристика на разстоянията, на базата на външните променливи. За значения на единиците с липсващи стойности се избират наблюдаваните единици с най-малки разстояния до тях, на базата на зададените (избраните) променливи.

[7] Друго много разпространено тяхно име е **Hot Deck методи**

[8] distance function matching

Един от най-елементарните подходи е изборът на една външна променлива X_i (Това в индекса единица ли е или е „i”.) и изчисляването на разстоянията D на базата на нея. За всички респонденти с липсващи стойности се изчисляват разстояния от типа:

$$Dji = |x_{j\ набл} - x_{i\ набл}| \qquad (4.1)$$

(или $Dji = (x_{j\ набл} - x_{i\ набл})^2$), където j определя единицата с липсваща стойност при променливата Y. Липсващата стойност се заместват със значението y_{i*}, където респондента $i*$ е донор за нереспондента j ако $D_{ji*} = \min_{(i)} |X_{j\ наб} - X_{i\ наб}|$. Предимство на метода е, че за въвеждане се използват действително наблюдавани стойности. Друго предимство е възможността да се подредят значенията по даден определящ признак, като по този начин се изолира (контролира) неговото влияние. Трябва да се отбележи, че резултатите са зависими от избрания ред във файла от данни. Shao, J., Wang, H. (2008) показват, че подходът на най-близкия съсед, въпреки че е детерминистичен метод, оценява разпределенията правилно. Някои значения могат да бъдат използвани няколко пъти за въвеждане, ако даден донор има съответствия за повече от една липсваща стойност, а други могат въобще да не бъдат използвани. В този случай вариацията $\widehat{\theta}_{.}(y)$ може да бъде занижена, ако даден донор е използван по-често от друг. За преодоляване на това множественото използване на даден донор може да бъде ограничено до даден конкретен брой пъти. Например, функцията на разстоянието може да бъде определена като:

$$Dji^* = \min\{|x_{jl} - x_{il}|*(1 + \mu t_i)\} \qquad (4.2)$$

, където μ принадлежащо на множеството R^+ е ограничението за всяко използване на донор, а ti е броя на използванията на респондента i като донор. Тук, обаче, отново остава проблема с подценяването на вариацията, породена от несигурността от действителните значения на ЛС.

Въвеждане по близост на оценените стойности.[9] Методът е съчетание на hot-deck въвеждащия подход и модели за въвеждане на данните – регресионни или други (Little 1988, Durrant и Skinner (2005a). В своята опростена форма методът представлява аналог на метода „Най-близък съсед”, където между донора и респондента е определено на базата на оценените, предвидените стойности на y_i (променливата, при която има липсващи стойности) – $\hat{y}_i$ чрез регресионен или друг модел. Въвеждане по близост на оценените стойности е в основата си детерминистичен метод. Следвайки принципите на Рубин въвеждане на стохастичен елемент в метода може да се получи, като от набора от значения, които са „близки” до оценяваните, се избира за въвеждане едно по случаен начин.

Друга форма на метода е hot-deck въвеждане в подкласове, групи формирани на базата на порядъка (размаха) на оценените стойности, чрез модела за предвиждане. Този метод дава възможност за използване на всички донорските значения в подкласовете, което е форма на превенция срещи нежелано редуциране на вариацията на въведените стойности. Донорските значения в подкласовете могат да бъдат избрани с или без връщане, като се очаква, че при варианта без връщане ще се редуцира в по-голяма степен вариацията на нововъведените стойности. Методът на въвеждане по близост на оценените стойности е композитен, съчетаващ в себе си регресионни елементи, елементи от метода „Най-близък съсед” и hot deck въвеждането. До колкото той е полупараметричен метод, независимо, че работи с модел на въвеждане, не е толкова чувствителен от неговото описание и неговата точност, за разлика например от регрессионното въвеждане (Schenker and Taylor, 1996).

Най-общо проблемите на методите, въвеждащи единични стойности, могат да бъдат обобщени по следния начин:

- Подценяване на вариацията на оценките
- Третиране на въведените стойности като наблюдавани и подценяване на срандартните грешки, нереално увеличаване на мощността на критериите и повишаване на риска от грешка от I род.
- Игнориране на допълнителната вариация произтичаща от това, че данните са липсващи, а не наблюдавани.

4.2 Повтарящи се въвеждания

За разлика от обсъжданите до тук методи за

[9] на английски методът е известен като *predictive mean matching imputation*

въвеждане, при които за дадено липсващо значение на признак се използва само едно въведено значение, в теорията и практиката съществува и подхода на множественото или повтарящото се въвеждане. За всяко липсващо значение се използват няколко независими оценки, чрез повторение, най-често, няколко пъти на единичните процедури на въвеждане като получените резултати се обобщават. Методът на множественото въвеждане (МВ), предложен от Рубин през 1987 г., е форма на повтарящи се въвеждания, целяща получаването на няколко независими оценки за всяка липсваща стойност. Идеята в този подход е, че повтарящите се въвеждания, сами по себе си, рефлектират върху несигурността на истинските, но ненаблюдавани стойности. Самите процедури на въвеждане използват максимално правдоподобни оценки (като EM алгоритъм) или бейсов подход (DA алгоритъм). Тези два алгоритъма осигуряват т.н. *подходящо (proper)* множествено въвеждане според Рубин (Rubin 1987). Общата оценка на параметрите на разпределенията и тяхната вариация става чрез единна и опростена техника, което е удобно за всеки изследовател независимо от неговите знания, умения и предпочитания към дадени типове анализи.

В теорията може да се срещне и друг подход за повтарящо се въвеждане основан на повтарящи се единични процедури, например регресионно или hot deck въвеждане. Този подход е популярен като фракционно въвеждане (Durrant и Skinner (2005a), но някои автори го смятат за *неподходящо (improper)* в смисъла на Рубин[10] (Binder и Sun, 1996; Rubin и Little 2002). Поради това използването му не решава основната задача при оценката на вариацията и остава необоснован завишения разход на творчески и технически ресурс.

Множествено въвеждане (МВ). Основната идея на МВ е да се въведат липсващите стойности като се използва подходящи въвеждащ модел и процедура, така че да се компенсира вариацията породена от несигурността за ЛС. Тази процедура се повтаря М пъти (М>2), като на всяка стъпка се въвеждат всички ЛС. Така след приключване на анализа се разполага с М пълни бази данни. След това във всяка от М-те пълни бази данни се провежда желания анализи, например оценка на параметрите на разпределенията, оценка на параметрите на регресионни модели и/или др. На последния етап резултатите от М-те оценки се обобщатват и се изчисляват техните вариации чрез правилата на Rubin (Rubin, 1987).

Методът за да работи е необходимо да са изпълнени определени правила, попадащи в рамките на понятието „подходящо въвеждане”[11]. Доналд Рубин въвежда идеята за подходящо въвеждане през 1987 г. с оглед на това в следствие на въвеждаща процедура да бъдат получавани неизместени, ефективни оценки на параметрите на изследваните разпределения, включително техните вариации. Самата идея може да бъде представена по сления начин: Нека X и Y са две променливи и X има ЛС. Нека за да се въведат стойностите на X да се използва стохастична регресия:

първо $Xi = a + bYi$

втора стъпка $Xi = a + bYi + s_{X.Y}ui$ (4.3)

,където a и b са регресионни коефициенти, $s_{X.Y}$ е остатъчната вариация, а ui е случайно избрано от ND(0; $s_{X.Y}$). Този подход на анализ третира a, b и $s_{X.Y}$ като действителни параметри на генералната съвкупност, а не като техни оценки. В действителност стойностите на тези параметри са неизвестни, но за подходящо множествено въвеждане всеки въведен вектор от данни трябва да бъде базиран на различен набор от стойности за a, b и $s_{X.Y}$. Тези стойности, също така, трябва да бъдат случайно избрани от бейсовите постериорни разпределения на параметрите. Само така множественото въвеждане може напълно да покрие несигурността по отношение на неизвестните параметри.

Същата идея за подходящо множествено въвеждане е обобщена от Schafer (1997) като независими реализации на постериорното разпределение на липсващите стойности - *Нлипс*. Ако $f(H_{липс}|H_{набл})$ е постериорно разпределение на липсващите стойности при даден априорен модел на пълната база данни то може да бъде записано, че:

$$f(H_{липс} \mid H_{набл}) = \int f(H_{липс} \mid H_{набл}, \varsigma)\, f(\varsigma \mid H_{набл})\, d\varsigma \quad (4.4)$$

[10] Съдържанието на понятието „подходящо въвеждане” се изяснява в следващата точка.

[11] Proper imputation

Следователно подходящото въвеждане рефлектира върху несигурността за *Нлипс* при дадените параметри на моделите на пълната база данни ($f(H_{липс} \mid H_{набл}, \varsigma)$) и неизвестните параметри на модела $_{\varsigma}$ ($f(\varsigma \mid H_{набл})$). Различията между индивидуалните резултати при отделните въвеждания се използва за оценка на несигурността причинена от липсващите данни.

Начинът на обобщаване на резултатите може да се представи по следния наин. Нека с $G\hat{}$ означим оценката на вариацията относно средната величина $\theta\hat{}$ като $G\hat{}$. е формула приложима към наблюдаваните и липсващите стойности едновременно. Двете оценки $\theta\hat{}$. и $G\hat{}$. се изчисляват поотделно за всяка от подсъвкупностите на наблюдаваните и липсващите данни. Оценките от *m*-тата новополучена пълна база данни да означим с:

$$\hat{\theta}_{.}^{(m)} = \hat{\theta}(H_{набл}, H_{липс}^{(m)}) \quad (4.5)$$

и

$$\hat{G}_{.}^{(m)} = \hat{G}(H_{набл}, H_{липс}^{(m)}), \text{при m} = 1,\ldots,\text{M}. \quad (4.6)$$

Според формулите на Rubin (Rubin 1987), за получаване на обединената точкова оценка в следствие на множественото въвеждане θ се използва следната осреднителна процедура:

$$\hat{\theta}_{.} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_{.}^{(m)} \quad (4.7)$$

За получаване на общата оценката на вариацията се изчислява две независими оценки. Средната от индивидуалните оценки за отделните M бази данни, наречена *вътрешно-групова вариация (within-imputation variance):*

$$\bar{G}_{.} = \frac{1}{M}\sum_{m=1}^{M}\overline{G}_{.}^{(m)} \quad (4.8)$$

и оценка на вариацията между индивидуалните точкови оценки - θ. Тази част от вариацията ще наречем *между-групова вариация* (*between-imputation variance):*

$$\hat{B}_{.} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_{.}^{(m)} - \hat{\theta}_{.}\right)^2 \quad (4.9)$$

Общата оценка получаваме като се обединят двата компонента на вариацията. Така *пълната вариация* можем да запишем като:

$$\hat{T}_{.} = \bar{G}_{.} + \left(1 + 1/M\right)\hat{B}_{.} \quad (4.10)$$

където $(1 + 1/M)$ е множител за крайност на *M.*

Един от начините за дефиниране на подходящо множествено въвеждане е, че в следствие на използването на формула за пълната вариация, действително се получават неизместени оценки на вариацията. Предимство на МВ е възможността му да дава като резултат пълни микро бази данни, които успешно могат да бъдат използвани от различен кръг потребители с различна подготовка и цели.

Важно място в анализа на ЛС с помощта на МВ заема модела, на базата на който се прави въвеждането. Препоръчва се моделът на въвеждане да не бъде различен от модела на анализ на данните, например ако за анализ на данните ще бъде използван регресионен модел, то е препоръчително същия модел да бъде използван и при МВ. Също така не се препоръчва изключването на променливи при МВ, които ще бъдат използвани при анализа. В противен случай се нарушава връзката между променливите, при които се въвеждат стойности и външните за МВ модел променливи, което се отразява на модела на анализ (Schafer, 1997; Sinharay, Stern and Russel, 2001, Schafer and Olsen, 1998). Carpenter Goldstein (2005) също отбелязват, че вътрешната структура и йерархията на данните също влияе на резултатите от анализа ако не е спазена при въвеждането.

Броят на повторенията M се препоръчва да бъде между 3 и 10, за получаване на реален ефект от въвеждането. МВ има предимство да предлага относително ясна и проста формула за пресмятане на вариацията на различни по съдържание оценките и на базата на различни модели. МВ може да бъде както при множествени разпределения, тека и при различни признаци – метрирани и неметрирани. На практика в момента това е най-дискусионния и препоръчван метод в литературата особено в областта на социалните науки (Rubin 1996, Schafer J 1997 и 1999, Zhang 2003, Schafer и

Olsen 1998, Allison, 2000 и 2001, Sinharay и др. 2001, Schafer и Graham 2002 и др.).

В практиката съществуват различни начини за достигане до подходящо МВ. Много често се залага на Монте Карло алгоритъма на основата Бейсовата статистика за симулиране на липсващите стойности. Това е в основата на алгоритъма[12] за увеличаване на данните, например. По този начин множественото въвеждане се превръща по същество в MCMC подход към анализа на пълната база данни (Rubin, 1996; Schafer, 1997; Lipsitz, Zhao and Molenberghs, 1998). Подобен подход е напълно параметричен, което изисква да се правят заключения за разпределенията на анализираните променливи. Много често това довежда до автоматичния избор на многомерното нормално разпределение за работно, което не винаги е вярно и не винаги гарантира сходим резултат (Horton and Lipsitz, 2001). Значително удобство, последните години, е появата на компютърни програми за изчисляване на марковските вериги, което компенсира запознатостта на голяма част от изследователите със съдържанието на метода.

Друг много разпространен метод за получаване на подходящо МВ е базиран на максимално правдоподобните оценки и е известен като ЕМ алгоритъм. И двата метода – алгоритъма за увеличаване на данните и ЕМ алгоритъма – са разгледани по-подробно в следващата публикация.

5. Заключение

Към днешна дата съществуват значителен брой методи за анализ *и въвеждане* на ЛС. Както стана видно от изложеното до тук всеки един от тях е базиран на обосновани заключения относно механизма на поява на самите ЛС. Това трябва да бъде и начина, по който се избира и съответния метод. За съжаление тази теория не е достатъчно позната на всички, които организират и провеждат емпирични изследвания или обработват подобна информация. Много често проблема на ЛС се игнорира, поради незнание за пораженията, които може да нанесе върху резултатите. Използват се т. н. „заложени по подразбиране" методи за анализ на ЛС в софтуерните продукти, които не винаги (или даже почти никога не) са адекватни в дадената ситуация. Обикновено тези методи са елиминационни, а те както се оказа трябва да бъдат използвани много пестеливо и с огромно внимание. По-доброто познаване на теорията и спецификата на анализа *и възможностите за въвеждане* на ЛС е задължително условие за повишаване на коректността на изводите и заключенията правени на база на събраната емпирична информация. Това важи за всички, които се впускат в това поле, независимо от техните интереси и квалификация и особено за статистиците.

Литература:

1. Afifi, A.A., Elashoff, P. M. (1966). Missing observations in multivariate statistics: Review of the literature, Journal of American Association, 61, 595-604

2. Allison, P.D. (2002). Missing Data. Sage University Papers Series on Quantitative Applications in Social Science, 07-136. Thousand Oaks, CA: Sage.

3. Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. Sociological Methods and Research 28: 301-309.

4. Bogdanov, B. (1988). Izmervane izmeneieto na otshenkite ot nablyudenieto na domakinskite byudzheti, Sotshiologicheski pregled, br. 3.

5. Demirtas, H. (2005) Bayesian Analysis oh Hierarchical Pattern-Mixture Models for Clinical Trails Data with Attrition and Comparisons to Commonly Used Ad-hoc and Model-based approaches, Journal of Biopharmaceutical Statistics, 15: 383–402.

6. Durrant, G. B., (2005) Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review, ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI), University of Southampton,.

7. Durrant, G.B. and Skinner, C. (2005): Using Missing Data Methods to Correct for Measurement Error in a Distribution Function, Survey Methodology

8. Durrant, G.B. and Skinner, C. (2005): Using Data Augmentation to Correct for Nonignorable Nonresponse when Surrogate Data are Available: An Application to the Distribution of Hourly Pay, Journal of the Royal Statistical Society, Series A.

[12] Markov chain Monte Carlo (MCMC)

9. Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM algorithm (whit discussion). Journal of the Royal Statistical Society. B39, 1-38

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

11. Ennett C. M, Frize M, Walker C.R. (2001) Influence of missing values on artificial neural network performance. Medinfo;10(Pt 1):449-53.

12. Fay, R.E. (1996): Alternative Paradigms for the Analysis of Imputed Survey Data, Journal of the American Statistical Association, 91, 434, 490-498.

13. Fay, R.E. (1999), Theory and application of nearest neighbour imputation in census 2000, Proceedings of the section on survey research methods, American Statistical Association 1999, pp. 112-121

14. Frick, J. R., Grabka, M. M. (2004), DIW Berlin, Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income distribution.

15. Ghosh-Dastidar, B. и Schafer, J. L., (2003) Multiple edit/ Multiple imputation for Multivariate Continuous Data. Journal of the American Statistical Association, Dec. 2003, Vol. 98, No. 464, Application and Case Studies

16. Glynn, R. J., Laird, N. M., Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (Ed.), Drawing inferences from self-selected samples (pp. 115–142). New York: Springer-Verlag.

17. Groves, R. M. (1989), Survey Errors and Survey Costs, Wiley -New York

18. Harrington, D (2009). Confirmatory Factor Analysis, Oxford University Press.

19. Hartley, H.O., Hocking, R.R. (1971). The analysis of incomplete data. Biometrics 27, 783-808

20. Horvitz, D.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population. Journal of the American Statistical Association. 47, 663-685.

21. Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models, Annals of Economic and Social Measurement 5, 475-492.

22. Howell D. C., Statistical Home Page, http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

23. Ibrahim, J. G, Chen M., Lipsitz S. R, Herring, A. H., (2005), Missing-Data Methods for Generalized Linear Models: A Comparative Review, Journal of the American Statistical Association; Mar 2005; 100, 469; ABI/INFORM Global, pg. 332

24. Kalton, G, Kish, L. (1981) Two efficient random imputation procedures, In Proc. Survey Res. Meth., p. 146-51. American Statistical Association.

25. Kalton, G. and Kasprzyk, D. (1986) The Treatment of Missing Survey Data. Survey Methodology 12, 1-16.

26. Kim, J. and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika (2004), 91, 3, pp. 559–578*

27. Kim, J. K., Fuller, W. A., Bel, W. R. (2008) Variance Estimation for Nearest Neighbor Imputation for U.S. Census Long Form Data, RESEARCH REPORT SERIES (Statistics #2008-13)

28. Little, R.J.A., Hyonggin A., (2003). Robust Likelihood-based Analysis of Multivariate Data with Missing Values. The University of Michigan Department of Biostatistics Working Paper Series. University of Michigan School of Public Health. Paper 5/2003

29. Little, R.J.A (1997). Biostatistical analysis with missing data. Encyclopedia of Biostatistics (P. Armitage, T. Colton, egs.), London: Wiley

30. Little, R.J.A, Rubin, D.B. (1983a). Incomplete data. Encyclopedia of the Statistical Science 4, 46-53

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

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

33. Little, R.J.A, Schenker, N. (1994). Missing data, Handbook of Statistical Modeling in the Social and Behavioral Sciences (G. Arminger, C.C. Clogg, M.E. Sobel, eds.), pp. 39-75. New York: Plenum.

34. Munnich, R., Rassler, S.,(2004) Variance Estimation under Multiple Imputation, study conducted within the DACSEIS research project (http://www.dacseis.de)

35. Newgard, C. D., Haukoos, J.S., Lewis, R.J. (2006), Missing Data: What Are You Missing? Society for Academic Emergency Medicine Annual Meeting San Francisco, CA. May 2006

36. Orchard, T., Woodbury, M.A. (1972). A missing information principle: theory and applications, Proc. 6th Berkeley Symposium on Mathematics Statistics and Probabilities. 1, 697-715

37. Oudshoorn, K., Buuren, S. v., Rijckevorsel, J. v. (1999): Flexible multiple imputation by chained equations of the AVO-95 Survey, TNO Prevention and Health, TNO report PG/VGZ/99.045

38. Raghunathan, T.E., Lepkowski, J.M. van Hoewyk M., Solenberger P.W. (2001): A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models, Survey Methodology, 27, 85-95.

39. Rubin, D.B. (1976). Inference and missing data (with discussion). Biometrika, 63, 581-592.

40. Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Survey. New York: Wiley.

41. Rubin, D. (1996). Multiple imputation after 18+ Years. Journal of the American Statistical Association 91 (June): 473–89.

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

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

44. Shao, J., Wang, H. (2008) Confidence Intervals Based On Survey Data With Nearest Neighbor Imputation, Statistica Sinica, Vol. 18, pp. 281-297, 2008

45. SPSS White Paper. Missing data: the hidden problem. http://www.spss.com

46. William E. Winkler, Bor-Chung Chen (2001), Extending the Fellegi-Holt Model of Statistical Data Editing, Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001