МЕТОДИ ЗА ВЪВЕЖДАНЕ НА ЛИПСВАЩИ СТОЙНОСТИ ПРИ НЕИГНУРИРУЕМИ МЕХАНИЗМИ

Деян Лазаров
Бургаски свободен университет

MISSING DATA IMPUTATION METHODS UNDER NONIGNORABLE MECHANISMS

Deyan Lazarov
Burgas Free University

Abstract: *In presented research was made an overview of the basic methods for dealing with missing data when nonignorable mechanisms exist. In the beginning different missing data mechanisms are presented. Second part shows several simple methods for Missing data imputation when mechanism is NMAR. In the next sections the basis of the Selection models and the advantages and disadvantages of this models are presented. The final section shows the essence, advantages and disadvantages of the Pattern mixture models.*

Key word*: missing data, missing data mechanisms, NMAR, Selection models, Pattern mixture models.*

Въведение

Подходите при въвеждането на липсващи стойности (ЛС) когато механизмът на формиране на сами липсващи данни не е игнорируем се различават значително от описаните при игнорируемите механизми. Това е в следствие на зависимостта на поява на ЛС и значенията на променливите, при които се появяват тези стойности. В тази ситуация функцията на максималното правдоподобие дава изместени оценки. При игнорируемите механизми ЛС се разглеждат като случайни и нормално разпределени извадки (или подизвадки) по отношение на всички признаци в базата от данни (при ЛНС) или поне в някои подизвадки, образувани от значенията на дадена променлива от базата данни (при СЛ), но без да са свързани с тази променлива. Така ако се приложат методите работещи под хипотезата за игнорируемост на механизма на ЛС се получават изместени оценки.

За по-голяма яснота е наложително да се направи кратък обзор на механизмите на ЛС, както и да се уточни съдържанието на понятията игнорируеми и неигнорируеми механизми.

Нека първоначално се дефинира пълна база от данни чрез $Y = (y_{ij})$, която представлява $(n \times K)$ правоъгълна матрица без липсващи стойности, с *i-ти* ред $y_i = (y_{i1}, \ldots, y_{iK})$, където y_{ij} е значението на признака Y_j за *i-mata* единица. Нека също така се въведе **индикатор на липсващи стойности**, което 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:

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

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

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

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 $Y_{липс}$:

$$f(M|Y,\phi) = f(M|Y_{набл},\phi) \text{ за всички стойности на } Y_{липс},\ \phi, \quad (3)$$

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 по-добро представяне на същността на този механизъм нека се предположи, че в базата от данни има само един признак с ЛС. 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,

$$f(Y,M|\theta,\phi) = f(Y|\theta)f(M|Y,\phi) = \prod_{i-1}^{n} f(y_i|\theta)\prod_{i-1}^{n} f(M_i|y_i,\phi) \quad (4)$$

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 измествания 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 механизмите на ЛС и условията за игнорируемост се предлагат за първи път от Donald Rubin. 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. 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 на самите ЛС. По този начин механизмите на ЛС определят и различните подходи за анализ на самите ЛС.

В настоящото изследване основна цел е представянето на възможностите за анализ и въвеждане на ЛС, когато механизмите са неигнурируеми. В теорията са разработени два класа от модели препоръчвани в подобни ситуации: **селекционен модел (СМ)[1] и Смесени модели на проява на ЛС (СМП)**[2]. И двата типа модели включват взаимодействието на разпределението на данните и вероятностното разпределението индикатора на ЛС, но по различен начин. СМ се състоят от две части – от една страна регресионни уравнения, които предвиждат вероятността за получаване на ЛС и разпределението на данните от друга. СМП моделите от своя страна имат друг подход. Те групират единиците в подгрупи, в които се наблюдават едни и същи модели на ЛС и провеждат последващ анализ във всеки модел поотделно. Трябва да се държи сметка, че и двата подхода имат своите слаби страни. СМ в голяма степен разчитат на непроверими характеристики на разпределенията, докато СМП изискват от изследователя да определи значения на един или повече неизвестни параметри, които на практика не могат да бъдат проверени. За съжаление никога не е известно дали тези предположения и допускания са в сила и от там, дали това което се получава като резултат не е по-лошо от това което може да се получи ако се използва допускането за СЛ механизъм. Това единствено показва, че използването на подобни модели трябва да се прави след особено добро опознаване на базата данни, с логиката на процесите които тя описва и добра запознатост с теорията при въвеждане на ЛС.

---

[1] Selection model (SM)

[2] The pattern mixture models (PMM)

Ad hoc подход за обработка на липсващи данни, когато механизма е НеСЛ

Този подход е имал своята актуалност в периода на последното десетилетие на 20 век и първото десетилетие на 21 век, когато липсата на подходящ софтуер е ограничавал решенията. Изследователите са разчитали на специални подходи за тестване на чувствителността на оценките получавани чрез допускането на СЛ механизъм. Rubin (Rubin, 1987) предлага значително опростен подход за решаване на задачата, която от своя страна може да бъде съвместена с процедурите за множествено въвеждане. Неговото предложение е следното: да се проведе множествено въвеждане[3] приемайки, че механизмът е СЛ, след това да се добави определена константа към значенията на въведените стойности за да се компенсира различията на значенията от грешно избрания механизъм, т.е. евентуално завишаване и подценяване на стойностите. В подобен подход ключов момент е адекватното предположение за средното завишаване или занижаване на липсващите стойности. Друг момент е априорното приемане, че разпределението на единиците с и без ЛС е едно и също по изследваните признаци. Друг подход в същи контекст е този на Cohen, при който се използва не определена предполагаема стойност, а се добавя определена част от стандартното отклонение към всяка въведена стойност (Enders, C 2010: стр. 289).

Rubin дава и други предложения за ad hoc решения. Той предлага добавянето на определена константа към дадена подгрупа от въведени стойности. Целта на подобен анализ е да се представи вариативността на оценките на параметрите по отношение на различните модели и допускания (Enders, C 2010: стр. 290). Въпреки всичко трябва да се *предполагат* какви и колко механизми действат в базата от данни. Грешни заключения относно това биха довели до неочаквани и неизвестни по размер грешки в заключенията.

Теоретични основи на моделирането при механизъм НеСЛ

Нека се върнем към индикаторната променлива *M*, която приема стойност 0 ако респондента е отговорил на дадения въпрос и има запис в базата данни и стойност 1 ако такъв запис липсва. В такъв случай може да се изчисли вероятността (напр. чрез логистична регресия) да се получат ЛС при дадена променлива от базата данни, като за предиктори служат останалите променливи. Rubin показва, че когато механизмът е СЛ параметрите на разпределението на липсващите стойности „...**не са натоварени с никаква „уникална" информация...**" относно параметрите на модела, на базата на който ще се оценят самите ЛС (Rubin, 1987). Поради тази причина, в литературата относно липсващите данни често се описва механизма СЛ като игнориращ ЛС, защото няма нужда да се вземат под внимание параметрите на разпределението на липсващите стойности при прилагане на базираните на максималното правдоподобие анализи (това включва едновременно оценките получени чрез метода на максималното правдоподобие и множественото въвеждане). От друга страна **при механизма НеСЛ има значима връзка между параметрите на разпределението на ЛС** и моделът, на базата на който ще се оценят самите ЛС. От тази гледна точка това вътрешно взаимодействие неминуемо довежда до смущения в оценките на параметрите на модела и оттам на самите ЛС. Целта на НеСЛ моделите е смекчаване или неутрализиране на това взаимодействие, чрез вмъкване на модела на вероятността на поява на ЛС, и оттам поучаването на неизместени оценки

[3] Един от подходите за въвеждане на ЛС. За повече информация виж. Rubin, 1987.

на ЛС. Двата типа модели СМ и СМП го постигат по различен начин. При СМ това се получава като регресионен модел, който оценя вероятността за поява на ЛС, докато при СМП извадката от единици (базата от данни) се стратифицира по отношение на съществуващите модели на ЛС и всеки един от моделите се оцени поотделно. В основата и на двата подхода е съвместното разпределение на данните и вероятността на поява на ЛС. Различията в неговото третиране дава и различието в двата подхода.

Съвместното разпределение на данните от базата и вероятността за поява на ЛС при тях е $p(Y,M)$. СМ моделите представят това разпределение по следния начин:

$$p(Y,M) = p\langle M \,|\, Y \rangle p(Y) \tag{5}$$

Където $p\langle M \,|\, Y \rangle$ е вероятностното условно разпределение на ЛС при дадени *Y,* а $p(Y)$ е разпределението на данните. Вероятностното условно разпределение на ЛС при дадени *Y* показва вероятността да се появи ЛС при лице участвало в изследването и имащо конкретни значения на изучаваните признаци. Разпределението на данните показва вероятността да се появи конкретно значение на признаците в базата от данни.

От своя страна СМП моделите могат да се представят по следния начин:

$$p(Y,M) = p\langle Y \,|\, M \rangle p(M) \tag{6}$$

където $p\langle Y \,|\, M \rangle$ е вероятностното условно разпределение на *Y,* при дадено значение на *M*, а $p(M)$ е разпределението на ЛС. Това представя отново съвместното разпределение на *Y* и *M*, но по обратен начин. Условното разпределение в случая определя вероятността да се наблюдава определено значение на *Y* в подгрупа определена от единици имащи еднакъв модел на ЛС, а разпределението на ЛС $p(M)$ описва различните модели на ЛС. Уравнение (6) представя логиката на РММ, а именно стратифициране на извадката от единици в изследването на базата на моделите на ЛС и провеждане на анализ във всяка отделна страта.

СМ и СМП са взаимосвързани в смисъл, че представят един и същи феномен и се отнасят до алтернативна форма на факторизация на едно и също съвместно разпределение. Въпреки това, двете факторизации лежат на много различни крайни предположения, което означава, че двата модела могат да произведат съвсем различни точкови оценки на параметрите и оттам различни оценки на ЛС. Някои изследователи дават предимство и предпочитат СМП, понеже изискват по-ясни предположения, отколкото СМ (Enders, C 2010).

Класически селекционен модел

Автор на идеята за СМ е Heckman (Heckman 1976) и той ги предлага като възможност за коригиране на изместването в регресионните модели при НеСЛ механизъм в базата от данни. Моделът предложен от Heckman се състои от две части, които комбинират едно основно регресионно уравнение описващи взаимодействията в базата от данни и едно допълнително, отново регресионно уравнение, което оценя вероятността за получаване на ЛС.

За да се представи основния начин на функциониране на СМ нека да се предположи, че се работи с един опростен модел на база от данни състояща се от три променливи Y, X_1 и X_2, като липсващи стойности се наблюдават само при Y. Нека също се предположи, че основното регресионно уравнение, което трябва да бъде оценено е:

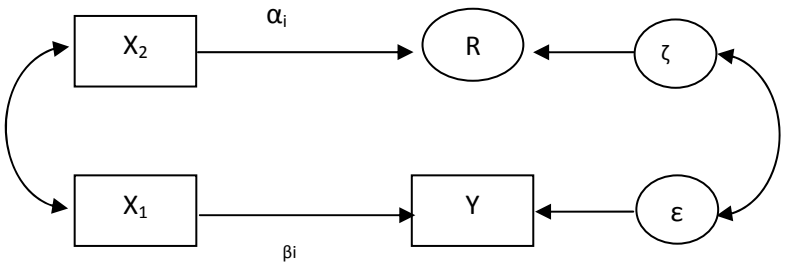
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{7}$$

където β_i са параметрите на модела, а ε са остатъци. При решаването на това регресионно уравнение се предполага, че няма ЛС. Обръщайки се към факторизацията на съвместното разпределение $p(Y, M)$ от урав. (5), то урав. (7) съответства на разпределението на данните $p(Y)$.

Втората част на СМ е регресионно уравнение, което да предвижда вероятността за отговор на респондентите при променливата Y. Класическия СМ дефинира вероятността на поява на ЛС при изходната променлива, като *нормално разпределена латентна променлива*. Случаите, които попадат над определена прагова стойност на тази латентна променлива имат запис на значение по изходната променлива, докато случаите, които попадат под прага имат ЛС. Нека тази латентна променлива се означи с R и тя трябва да се разграничава от индикатора на ЛС - *M*. Тогава може да се запише, че:

$$R = \alpha_0 + \alpha_1 X_2 + \xi \tag{8}$$

където α_i са параметрите на модела, а ζ са остатъците. Трябва да се отбележи, че горните две уравнения могат да съдържат едни и същи предиктори, но е необходимо поне една променлива – фактор да се различава. На фиг. 1 е показана пътечковата диаграма на типичния СМ.



Фигура 2. *Класически селекционен модел*

На диаграмата са изобразени по стандартен начин променливите, които са наблюдавани в базата от данни (представени в правоъгълници) и латентни променливи представени в елипси или кръгове. Единичните стрелки показват регресионни връзки, а двойните ковариационни. Така взаимодействието между ζ и ε показва ковариацията между остатъчните елементи в двете уравнения. Тази ковариация е изключително важна, защото така моделът на ЛС се аждустира за изместването в регресионните коефициенти.

НеСЛ механизмът определя, че данните и вероятността за поява на ЛС имат съвместно разпределение, така че данните носят информация за вероятността за поява на ЛС и обратно. В СМ тази зависимост е описана чрез бивариационни нормални разпределения за остатъчните елементи:

$$\begin{bmatrix} \varepsilon \\ \zeta \end{bmatrix} \sim BN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\varepsilon}^{2} & \sigma_{\varepsilon,\zeta} \\ \sigma_{\zeta,\varepsilon} & \sigma_{\zeta}^{2} \end{bmatrix} \right) \quad (9)$$

където нулевия вектор представя средите на остатъците в двете уравнения, които трябва да са равни на 0, а σ^2_ε, σ^2_ζ и $\sigma_{\varepsilon,\zeta}$ са съответно дисперсиите и ковариацията между остатъците. Ковариацията между остатъците е механизма, чрез който моделът на ЛС (урав. 8) аджустира изместването в основното регресионно уравнение (урав. 7).

Обръщайки се отново към фиг. 1 се вижда, че съществуването на НеСЛ механизма се дължи на ковариацията на остатъците от двете регресионни уравнения. Това може да се използва като инструмент за оценка на механизмите на ЛС. При моделиране на връзките по описания чрез урав. (7) и (8) начин се установи, че ковариацията при остатъците е статистически значимо различна от нула, то моделът би трябвало да е НеСЛ. Въпреки всичко, СМ са изключително чувствителни към допускането на нормалност на разпределенията, единични и множествени. В случай, че някое от разпределенията в базата от данни напусне нормалността, то това повлиява бивариационното нормално допускане за разпределението на остатъците. При не изпълнено условие за нормалност на бивариационното разпределение не може да се гарантира постигането на неизместеност на оценките от модела, включително и оценката на ковариацията на самите остатъци. Затова използването на оценката на ковариацията за оценка на механизма на липсващи стойности е ненадежден подход.

Оценката на СМ може да се направи чрез два подхода. Предложението на Heckman е да се използва двустъпков МНМК, но към днешна дата техническите възможности позволяват използването и изчисляването на функцията на максималното правдоподобие изключително лесно. Използването на методът на максималното правдоподобие от своя страна се гради на допускането, че разпределението на R и Y е бивариационно нормално, което в емпирична ситуация изисква проверка.

Последователността от стъпки при оценка на модела е следната: Първа стъпка е използването на пробит (probit) модел за намирането на регресионната зависимост между R и една или повече обясняващи променливи. Целта е да се получат оценките на вероятността за поява на ЛС. Втората стъпка изисква оценката на основното регресионно уравнение, като в анализа не участват единиците с липсващи стойности (прилага се последователно елиминиране). Включването на оценката на вероятностите като независима променлива в основния регресионен модел коригира изместването на оценките на коефициентите му.

Ограничения на СМ

Общото мнение е, че СМ биха могли да редуцират или напълно елиминират изместването на оценките в следствие на НеСЛ, но при условие, че са изпълнени условията за прилагането им (Enders, C 2010: стр. 296). В действителност тези условия не винаги са налице, което може да доведе до резултати значително по лоши от тези, които могат да бъдат получени ако се прилага анализ при механизъм на СЛ. Начинът по който СМ компенсират изместването на оценките се основава на оцен-

ката на R – вероятността за поява на ЛС. Регресионното уравнение, което описва появата на ЛС трябва да се гради върху ясни и правдиви хипотези. За съжаление, в която и да е реална ситуация няма пълна яснота за причините за поява на липсващите стойности. От друга страна, изборът на променливи за описание на вероятността за поява на ЛС от наличните в базата данни води до това, че оценената променлива R ще бъде сериозно корелирана с тях. Следователно, когато тази променлива се използва за оценка на параметрите на основното регресионно уравнение, в което най-често се използват същите предиктори, то това води до колинеарност[4] между входните променливи. За да се минимизира опасността в подобна ситуация, изследователите препоръчват поне една от независимите променливи в модела за вероятността на поява на ЛС да е различна от тези при модела на основното регресионно уравнение. Това, разбира се, не гарантира винаги успех. Също така остава въпроса, че появата на ЛС не винаги може да се обясни добре само с признаците в базата от данни.

Смесени модели на проява на ЛС (СМП)[5]

Както вече беше отбелязано в уравн. (6), СМП факторизират съвместното разпределение на данните и липсващите стойности по следния начин:

$$p(Y,M) = p\langle Y \mid M \rangle p(M) \tag{10}$$

В класическата си проява СМП оценят дадени определени характеристики на разпределението на данните в подгрупи (напр. средните по даден признак), определени от различните модели на ЛС. След получаването на подгруповите оценки те се осредняват, за да се получи обща оценка за цялата база от данни. Това води със себе си някои важни особености. При определянето на подгрупите може да се окаже, че характеристиките или параметрите при техните разпределения са неоценими или неидентифицируеми, поради липса на информация. За преодоляването на тези затруднения се налага да се въвеждат редица правила за идентифицируемост (Glynn, Laird, Rubin, 1986; Little, 1993; Rubin, 1987; Enders, C 2010). Тези превила въвеждат връзките между неоценимите параметри, породени от наличието на ЛС, с останалите наблюдавани значения при единиците от подгрупата. За да се илюстрират правилата на идентифицируемостта нека:

$$\mu_Y = \beta_0 + \beta_1 \mu_X \tag{11}$$

където μ_Y и μ_X са математическите очаквания на променливите Y, която съдържа ЛС и X. В действителност параметрите на това регресионно уравнение са неоценими поради наличието на ЛС. Това, което може да се оцени е как биха изглеждали параметрите на регресионното уравнение само при единиците без ЛС. В случая ограничението, което се въвежда, е приемането за еднаквост между параметрите на уравнението при единиците с нелипсващи стойности и тези с липсващи стойности, т. е. оценката на параметрите се получава от:

---

[4] Колинеарността е зависимост между два предиктора, а мултиколинеарността между повече от два. Това е явление, което сериозно нарушава правилното оценяване на параметрите в моделите.

[5] Pattern mixture model

$$\mu_{Y(\text{липсв})} = \beta_0 + \beta_1 \mu_{X(\text{липсв})} \quad (12)$$

Следователно, когато се използва, че при X няма липсващи стойности, то може да се намери $\mu_{X(Y\text{липсв})}$ – средната аритметична от стойностите на X при тези значения на Y, които липсват. При хипотезата, че коефициентите β_i се запазват еднакви при $Y_{(\text{липсв})}$, то

$$\mu_{Y(\text{липсв})} = \beta_0 + \beta_1 \mu_{X(Y\text{липсв})} \quad (13)$$

Ако се предположи, че в базата данни има $к$ на брой подгрупи със изразени специфични модели на ЛС този подход се прилага при всяка от тях. За получаването на общите оценки за цялата база от данни се използва притеглено осредняване от отделните групови оценки с тегла обема на подгрупите. При оценката на стандартната грешка на тази оценка се налага използването на т.н. Делта метод.

Ограничения на СМП

Подобно на СМ и СМП базират своите резултати на неоценими параметри. Въпреки, че този тип модели не изискват някакви специфични допускания за разпределенията на променливите се изисква изследователя да прави допускания и да приеме значения на неоценимите параметри. При грешно дефиниране на подгрупата, от която да се извлекат параметрите на подмоделите, се достига до грешни оценки и липса на редуциране на изместването заради механизма НеСЛ. Това налага оформянето на подгрупите с еднакви модели на ЛС да бъде направено по безспорен начин от изследователите. Предимство на анализа е, че всяка подгрупа се третира като отделна база от данни с отделни специфични разпределения и с игнорируем механизъм на поява на ЛС, което позволява да се прилагат различни подходи за описването на моделите на ЛС.

Заключение

В основата на добрия анализ на ЛС и подходящото им въвеждане е познаването на природата на техните механизми на поява. В редица случаи този механизъм е НеСЛ и това създава редица затруднения на изследователите. Дори се срещат мнения, че пренебрегването на този механизъм и използването на методи адекватни при случайни механизми работят добре, което е сериозно оспоримо (Allison, 2002). Гореизложените методи са базови, и от тази гледна точка с малко по-висока степен на универсалност. Върху тях лесно могат да бъдат надградени техни деривати, които в много по-голяма степен биха били адекватни в практическите ситуации, в които може да се попадне. За съжаление към момента трудно може да се говори за абсолютно, универсално решение, така че познаването на теорията на анализа на ЛС е в основата на елиминирането на тези необратими грешки, от които зависят решенията на изследователите. Честа практика е проблема с ЛС да 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 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 намират или се впускат в полето на изследователската работа, независимо от техните интереси и квалификация.

Литература:

1. Allison, P.D. (2002). Missing Data. Sage University Papers Series on Quantitative Applications in Social Science, 07-136. Thousand Oaks, CA: Sage.
2. Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. Sociological Methods and Research 28: 301-309.
3. 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.
4. Durrant, G.B. and Skinner, C. (2005): Using Missing Data Methods to Correct for Measurement Error in a Distribution Function, Survey Methodology
5. 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.
6. 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
7. Enders, C. K. (2010) Applied missing data analysis, The Guilford Press
8. 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
9. 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.
10. 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
11. 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.
12. Hartley, H.O., Hocking, R.R. (1971). The analysis of incomplete data. Biometrics 27, 783-808
13. 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.
14. 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.
15. Kalton, G. and Kasprzyk, D. (1986) The Treatment of Missing Survey Data. Survey Methodology 12, 1-16.
16. Kim, J. and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika (2004), 91, 3, pp. 559–578*
17. 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)
18. Little, R.J.A (1997). Biostatistical analysis with missing data. Encyclopedia of Biostatistics (P. Armitage, T. Colton, egs.), London: Wiley

19. Little, R.J.A, Rubin, D.B. (1983a). Incomplete data. Encyclopedia of the Statistical Science 4, 46-53
20. Little, R.J.A, Rubin, D.B. (1987). Statistical analysis with missing data. New York: Wiley.
21. Little, R.J.A, Rubin, D.B. (2002). Statistical Analysis with Missing Data - 2nd ed., New Jersey: Wiley.
22. 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.
23. 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
24. 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
25. 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
26. 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.
27. Rubin, D.B. (1976). Inference and missing data (with discussion). Biometrika, 63, 581-592.
28. Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Survey. New York: Wiley.
29. Schafer, J.L (1997). Analysis of Incomplete Multivariate Data, Chapman & Hall
30. Schafer, J. (2002), Dealing with Missing Data, Research Letters in the Information and Mathematical Sciences 3, 153-160