Множествено модераторно въвеждане

Деян Лазаров,
Бургаски свободен университет

1. Въведение

Един от основните проблеми на емпиричните изследвания е появата на липсващи стойности (ЛС). Проблемът е във фокуса на изследователите последните десетилетия, но търсенията в посока на адекватни методи и подходи продължават. Проблемите, възникващи от появата на ЛС в базите от данни, влияят пряко върху качеството на информацията и точността на оценките. В много случаи се наблюдава изместване на основните характеристики на съвкупността и получените резултати се превръщат в непредставителни за изучаваната съвкупност. Един от подходите за преодоляване на недостатъците от ЛС е използването на единични и множествени въвеждания. В настоящото изследване се прави нещо по-различно. Към класическия вариант на множествено въвеждане (МВ) се добавя модераторен анализ. Тази идея е насочена към по-пълното описание на взаимодействията между променливите в базите от данни и оттам на по-добрия модел, чрез който да се осъществи въвеждането на ЛС.

2. Множествено въвеждане

Множественото въвеждане е метод създаден да компенсира наличието на ЛС като се използва подходящ въвеждащ модел и определена въвеждаща процедура. Особеност на МВ е, че тази процедура се повтаря S пъти (S>2), като на всяко повторение се провежда желания анализи, например изчисляване на относителни дялове, оценка на параметрите на определен регресионен модел или др. във всяка от S-те пълни бази данни, получени след въвеждането. На следващ етап резултатите от S-те оценки се обобщават чрез правилата на Rubin (1987). Тази логика е изобразена на фиг. 1.

МВ е пряко свързан с понятието „подходящо въвеждане". 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 разсейвания. Самата идея може да бъде представена по следния 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:

$$\begin{aligned} &\text{първо} && X_i = a + bY_i \\ &\text{втора стъпка} && X_i = a + bY_i + s_{X.Y}u_i \end{aligned} \qquad (1)$$

където a и b са регресионни коефициенти, $s_{X.Y}$ е остатъчната вариация, а u_i е случайно избрано от нормално разпределение със средна 0 и разсейване $s_{X.Y}$. 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. Така множественото въвеждане може напълно да покрие несигурността по отношение на неизвестните параметри.

Съдържанието на понятието подходящо въвеждане формално може да се опише като независими реализации на постериорното разпределение на ЛС. Rubin (1987 и 1996) дефинира „подходящото" въвеждане през призмата на честотната перспектива, без да конкретизира даден специфичен параметричен модел. Прилагането на подходящо МВ позволява използването на S-те резултативни бази данни за прилагане на стандартните анализи като накрая резултатите за изследваните параметри се обобщават.

Различията между индивидуалните резултати за параметрите, при отделните въвеждания, се използва за оценка на несигурността причинена от липсващите данни.

Фигура 1. Модел на провеждане на множествено въвеждане

Това може да се опише формално по следния начин. Нека с $\widehat{G}$ означим оценката на вариацията на даден параметър $\hat{\bar{\theta}}$ на разпределението с ЛС. В контекста на МВ се въвеждат S отделни бази от данни. Двете оценки $\hat{\theta}_{\cdot}$ и $\widehat{G}_{\cdot}$ се изчисляват поотделно за всяка от тях. Тези оценки от *i*-тата новополучена пълна база данни ($i = 1,.., S$) може да се запишат като:

$$\hat{\theta}_{\cdot}^{(i)} = \hat{\theta}(H_{\text{набл}}, H_{\cdot\text{липсв}}^{(i)}) \qquad (2)$$

и

$$\hat{G}_{\cdot}^{(i)} = \hat{G}(H_{\text{набл}}, H_{\cdot\text{липсв}}^{(i)}) \qquad (3)$$

Според формулите на Rubin (1987), за получаване на обединената точкова оценка на $\bar{\theta}$ в следствие на множественото въвеждане се използва следната осреднителна процедура:

$$\hat{\theta} = \frac{1}{S}\sum_{i=1}^{S} \hat{\theta}_{\cdot}^{(i)} \qquad (4)$$

За получаване на общата оценката на вариацията се изчисляват две независими оценки. Средната от индивидуалните оценки за отделните S бази данни, наречена *вътрешно-групова дисперсия*[3]*:*

$$\hat{\mathrm{G}} = \frac{1}{\mathrm{S}}\sum_{\mathrm{i}=1}^{\mathrm{S}} \hat{\mathrm{G}}_{\cdot}^{(\mathrm{S})} \qquad (5)$$

и оценка на разсейването между индивидуалните точкови оценки - θ. Тази част от дисперсията ще наречем *между-групова дисперсия*[4]*:*

$$\hat{B}_{\cdot} = \frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\theta}_{\cdot}^{(s)} - \hat{\theta}_{\cdot}\right)^2 \qquad (6)$$

Общата оценка се получава като се обединят двата компонента на разсейването. Така *пълната дисперсия* може да се запише като:

$$\hat{T}_{\cdot} = \bar{G}_{\cdot} + \left(1 + \frac{1}{s}\right)\hat{B}_{\cdot}, \qquad (7)$$

където $(1 + 1/S)$ е множител за крайност.

Един от начините за дефиниране на подходящ метод на множествено въвеждане е, че в следствие на използването на формула за пълната дисперсия, действително се получават неизместени оценки на разсейването. Предимство на МВ е възможността му да дава като резултат пълни 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 моделът, на базата на който се прави въвеждането. Най-общо трябва да се каже, че моделът на въвеждане трябва да подсигурява наличните корелационни и ковариационни зависимости в базата от данни, например ако за анализ на данните ще бъде използван регресионен модел. Също така не се

---

[3] *within-imputation variance*

[4] *between-imputation variance,*анг.

препоръчва изключването на променливи при МВ, които ще бъдат използвани в последствие при анализа. В противен случай се нарушава връзката между променливите, при които се въвеждат стойности и външните променливи за модела на МВ, което се отразява на последващия анализ (Schafer, 1997; Sinharay, Stern и Russel, 2001). Много важен аспект при анализа е спазването на вътрешната структура и йерархията на данните при въвеждането, което силно влияе на резултатите от анализа.

Броят на повторенията S 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 дисперията на различни по съдържание оценките и на базата на различни модели. МВ може да бъде използвано при многомерни разпределения и различни признаци – метрирани и неметрирани (Rubin, 1996; Schafer, 1997; Little и Rubin, 2002; Allison, 2000 и 2001; Sinharay, Stern и Russel, 2001).

В практиката съществуват различни начини за достигане до подходящо МВ. Много често се залага на методи използващи Монте Карло алгоритма (МКА)[5] и дефинирани на основата Бейсовата статистика. По този начин МВ се превръща по същество в МКА подход към анализа на пълната база данни (Rubin, 1996; Schafer, 1997). 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 добър резултат. Значително удобство, последните години, е появата на компютърни програми за пресмятане на МКА, което компенсира запознатостта на голяма част от изследователите със съдържанието на метода.

2. Модераторен анализ

Често в проявата на дадено явление може да се открие, че в зависимост от значенията на дадена променлива определени зависимости за проявяват по различен начин. Например, някои изследователи твърдят, че в зависимост от признака пол има разлика в зависимостта между характеристиките на наетия и неговото заплащане. Също така може да се твърди, че в зависимост от сектора на икономическа дейност зависимостта между предлагания труд и неговата цена е различна. И в двата случая говорим за модераторно влияние. В първия пример модераторът е полът, а във втория икономическият сектор. Модераторният анализ (МА) дава отговор на въпроса дали една или повече променливи са модератори при определена връзка.

Модератор може да бъде качествена (например: пол, раса, клас) или количествена (размер на възнагражденията) променлива, която повлиява посоката и/или силата на връзката между една независима променлива (предиктор) и една зависима (критериална) променлива (Baron, R. M., & Kenny, D. A., 1986). Модераторният ефект често се нарича и ефект на взаимодействието, защото модераторната променлива взаимодейства с връзката между определени предиктори Xi и дадена критериална променлива Y. Тази връзка се променя значимо в зависимост от нивата на модератора. За да се изследва модераторното влияние на модератор Z, при зависимост между една променлива X и една Y, може да се използва регресионен анализ като е необходимо да се генерира нова променлива на взаимодействието,

[5] Markov chain Monte Carlo (MCMC), анг.

която е равна на произведението X*Z. Ако има повече от един предиктор в зависимостта по отношение на Y, например X1 и X2[6], то е необходимо да се намери взаимодействието между X1и Z = X1*Z и X2 и Z = X2*Z.

3. Множествено въвеждане и модераторен анализ

Едно от важните условия за провеждане на успешен анализ на ЛС, включително и МВ, е използването на модел на зависимостта, който в максимална степен отразява взаимодействието между променливите в базата от данни. При условие, че всеки модераторен анализ допълва и разкрива дълбочините на взаимодействието между предиктори и критерии, то използването му в МВ би трябвало да подпомогне анализа на ЛС. Така може да се дефинира следната хипотеза в настоящият анализ:

Ако Z е модератор при връзката между два предиктора X1 и X2 и една критериална променлива Y, то модераторния анализ и прилагането на МВ при отделните групи или категории на Z подобрява резултатите от МВ.

Добавянето на модераторен анализ към МВ може да бъде илюстрирано по следния начин фиг. 2. В следствие на неговото използване и определяне на дадена променлива като модератор, съвкупността от единици би могла да се раздели на подгрупи, в зависимост от значенията на модератора. МВ може да се приложи при всяка една отделна подгрупа като анализа за получаване на отделните оценки се прилага след като данните от отделните подгрупи се обединят, за да се получи отново единна база данни.

4. Симулация на работна база данни и предварителен анализ

За проверка на издигнатата хипотеза се използва симулация на база данни, подчинена на следните зависимости. Генерирани са три променливи X1, X2 и Y по следния начин:

1. За първите 1000 единици тези три променливи образуват многомерно нормално разпределение със следните начални характеристики:

```
cor(x1,X2) = 0,3
cor(X1,Y) = 0,7
cor(X2,Y) = 0,8
```

	X1	X2	Y
Средни аритметични	50	100	200
Стандартни отклонения	50	60	60

[6] Тази зависимост между Y – зависима величина и X1 и X2 - предиктори ще бъде записвана за удобство като Y~X1+X2.

Фигура 2. Модел на провеждане на множествено въвеждане с модераторен анализ

2. За следващите 500 единици тези три променливи образуват многомерно нормално разпределение със следните начални характеристики:

```
cor(x1,X2) = 0,3
cor(X1,Y) = 0,3
cor(X2,Y) = 0,9
```

	X1	X2	Y
Средни аритметични	50	100	200
Стандартни отклонения	50	60	60

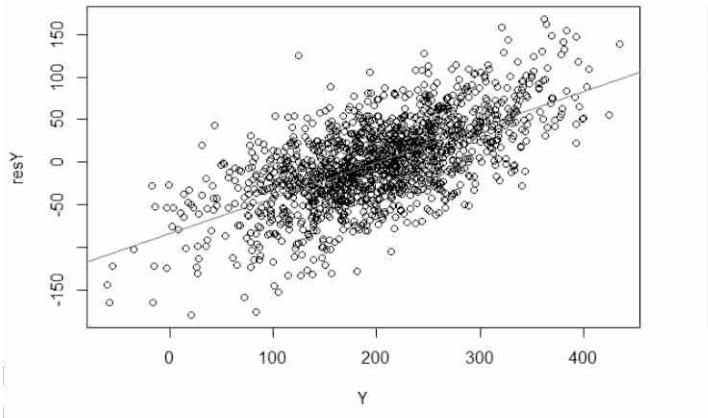
3. Генерира се нова дихотомна променлива Z, която приема значение 1, за първите 1000 единици и значение 0, за последните 500 единици. В зависимост от това, че взаимовръзката между X1,X2 и Y се променя при различните значения на Z, то Z се явява модератор при връзката между предикторите X1 и X2 и критериалната променлива Y.

Първоначално върху създадената база данни (data) е приложен регресионен анализ като коефициентите на регресията са изчислени чрез метода на най-малките квадрати (МНМК)[7]. Резултатите от анализа могат да се видят в табл. 1.

Таблица 1. Регресионни оценки при връзката Y~X1+X2 от симулираната база от данни, без отчитане на влиянието на модератора

```
Коефициенти:
             Оценка    Ст. грешка t-равнище  Pr(>|t|)
(Intercept) 101,30578   2,53183     40,01   <2e-16 ***
X1            0,36336   0,02660     13,66   <2e-16 ***
X2            0,80373   0,02142     37,52   <2e-16 ***
---
Кодове на стат, значимост:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Ст. грешка на остатъците: 49,73 с 1497 степени на свобода, Множествен коеф.
на детерминация:  0,5831, Ажустиран мн. коеф. на детерминация:  0,5825, F-
statistic:  1047 с 2 и 1497 DF, p-value: < 2,2e-16
```

На фиг. 3 е представена зависимостта между симулираните стойности на Y и остатъците от регресионния модел, използван в случая.



Фигура 3. Зависимост между Y и остатъците (resY) при регресионната връзката Y~X1+X2 от симулираната база от данни, без отчитане на влиянието на модератора

[7] В анализа е използван софтуерния продукт R.

При условие, че е известно, че Z е медератор при връзката Y~X1+ X2, то неговото влияние може да се разкрие, като се проведе еднотипен регресионен анализ в двете подсъвкупности от единици при Z=1 и Z=0. Резултатите могат да се видят в табл. 2, за Z=0, и табл. 3, за Z=1.

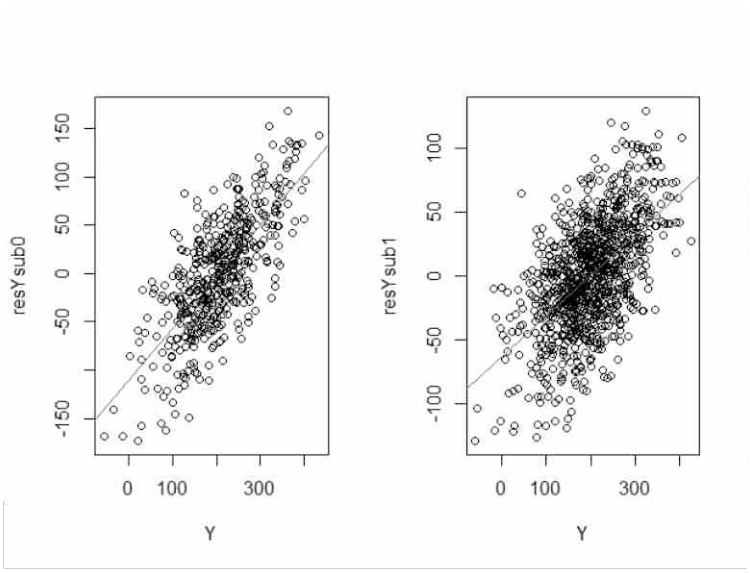
Таблица 2. Регресионни оценки при връзката Y~X1+X2 от симулираната база от данни, при отчитане на влиянието на модератора (Z=0)

```
Коефициенти:
             Оценка    Ст. грешка t-равнище  Pr(>|t|)
(Intercept) 115,25977    5,18142   22,245    <2e-16 ***
X1            0,03957    0,05432    0,729     0,467
X2            0,84483    0,04312   19,595    <2e-16 ***
---
Кодове на стат. значимост:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Ст. Грешка на остатъците: 58,42 с 497 степени на свобода, Множествен коеф.
на детерминация:  0,4653, Ажустиран мн. коеф. на детерминация:  0,4631,
F-statistic: 216,2 с 2 и 497 DF, p-value: < 2,2e-16
```

Таблица 3. Регресионни оценки при връзката Y~X1+X2 от симулираната база от данни, при отчитане на влиянието на модератора (Z=1)

```
Коефициенти:
             Оценка    Ст. грешка t-равнище  Pr(>|t|)
(Intercept) 94,50928    2,65323    35,62     <2e-16 ***
X1           0,52356    0,02791    18,76     <2e-16 ***
X2           0,78240    0,02264    34,56     <2e-16 ***
---
Кодове на стат. значимост:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Ст. грешка на остатъците: 42,66 с 997 степени на свобода, Множествен коеф.
на детерминация:  0,6812, Ажустиран мн. коеф. на детерминация:  0,6806
F-statistic: 1065 с 2 и 997 DF, p-value: < 2,2e-16
```

Графичния вид на зависимостите между значенията на променливата Y и остатъците (resY) от регресионната зависимост Y~X1+X2, при двете подсъвкупности, може да се види на фиг. 3. Коефициентите при регресионните зависимости, както и графичните изображения, показват че при двете подсъвкупности има различие при връзката Y~X1+X2. Нещо повече, при подсъвкупността при Z=0, X1 няма статистически значимо влияние, за разлика от другата подсъвкупност (при Z=1). Що се отнася до графичните изображения, лесно се вижда, че наклона на линията описваща зависимостта между Y и resY при двете подсъвкупности е различен.



Фигура 3. Графичния вид на зависимостите между значенията на променливата Y и остатъците от регресионната зависимост Y~X1+X2 при двете подсъвкупности. Вляво при Z=0, вдясно при Z=1

Модераторният анализ може да бъде проведен и чрез използване на регресионен модел. За целта е необходимо да се генерират две нови променливи на взаимодействието ZX1=Z*X1 и ZX2=Z*X2. В подобен анализ с две независими променливи в едно регресионно уравнение по отношение на Y се включват като предиктори X1 + X2 + Z + ZX1 + ZX2. Резултатите от модераторния анализ е представен в табл. 4. За намирането на коефициентите на регресионния модел се използва МНМК.

Таблица 4. Регресионни оценки при връзката Y~X1+X2+ Z + ZX1 + ZX2 от симулираната база от данни, при отчитане на влиянието на модератора Z

Коефициенти:

	Оценка	Ст. грешка	t-равнище	Pr(>|t|)	
(Intercept)	115,25977	4,29943	26,808	< 2e-16	***
X1	0,03957	0,04508	0,878	0,380	
X2	0,84483	0,03578	23,614	< 2e-16	***
Z	-20,75049	5,25113	-3,952	8,12e-05	***
ZX1	0,48399	0,05511	8,782	< 2e-16	***
ZX2	-0,06243	0,04407	-1,417	0,157	

Кодове на стат, значимост: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Ст. грешка на остатъците: 48,48 с 1494 степени на свобода,
Множествен коеф. на детерминация: 0,6046, Ажустиран мн. коеф. на детерминация: 0,6033 F-statistic: 456,9 с 5 и 1494 DF,
p-value: < 2,2e-16

Анализът показва статистическата значимост на променливите на взаимодействието ZX1, което е в контекста на симулираните значения на базата от данни и може да се приеме като атестация за адекватността на регресионния подход при изследване на модерацията. Но при този подход има и някои опасности, които се крият в това, че новосъздадените променливи на взаимодействието могат да бъдат в силна зависимост със останалите предиктори – X1, X2 и Z.

Това явление е известно като колинеарност (или мултиколинеарност) и има негативно влияние върху резултатите от анализа, получен чрез МНМК. Колинеарността повлиява както самите оценки, така и тяхната стандартна грешка. В настоящия анализ има опасност от подобен проблем поради cor(X1, ZX1)= 0,72 и cor(X2, ZX2)= 0,58. При корелация между два предиктора по-голяма от 0,7 може да се очаква колинеарност. Това съответства на процент обяснена дисперсия между двете променливи от порядъка на 50%. Също така корелация от 0,58 също може да се смята за обезпокоителна. Това се потвърждава и от изчислените индекси на дисперсионни инфлационни фактори (ДИФ)[8] табл. 5. Наличието на ДИФ > 5 показва опасност от колинеарност сред предикторите.

Таблица 5. Дисперсионни инфлационни фактори (ДИФ) на предикторите

при връзката Y~X1+X2+ Z + ZX1 + ZX2

	X1	X2	Z	ZX1	ZX2
ДИФ	3,3228	3,2287	3,9109	4,2868	6,0800

Едно от стандартните решения в подобни ситуации е използването на центрирани значения на предикторите, вместо оригиналните. Центрираните значения се поличават като от оригиналните значения се извади тяхната средна аритметична. Анализът за изследване на модерацията при центрирани данни е показан в табл. 6. В табл. 6 центрираните значения на предикторите са означени със *scale* преди името на променливата.

Таблица 6. Регресионни оценки при връзката Y~X1+X2+ Z+ZX1+ZX2 от симулираната база от данни, центрирани предиктори, при отчитане на влиянието на модератора Z

Коефициенти:

	Оценка	Ст. грешка	t-равнище	Pr(>|t|)	
(Intercept)	201,12042	1,25173	160,674	< 2e-16	***
scale(X1)	0,03957	0,04508	0,878	0,380	
scale(X2)	0,84483	0,03578	23,614	< 2e-16	***
scale(Z)	-20,75049	5,25113	-3,952	8,12e-05	***
scale(ZX1)	0,48399	0,05511	8,782	< 2e-16	***
scale(ZX2)	-0,06243	0,04407	-1,417	0,157	

Кодове на стат, значимост: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Ст. грешка на остатъците: 48,48 с 1494 степени на свобода,
Множествен коеф. на детерминация: 0,6046, Ажустиран мн. коеф.
на детерминация: 0,6033 F-statistic: 456,9 с 5 и 1494 DF,
p-value: < 2,2e-16

След центрирането се виждат, че няма негативни ефекти от колинеарността освен по отношение на свободния член, Анализът показва, че променливата X1 е статистически не

[8] На англ. Variance Inflation Factor (VIF), който се изчислява като $1/(1-R^2)$, където R^2 е коефициента на детерминация между дадения предиктор и останалите предиктори в регресионното уравнение. За повече информация [John Maindonald, W. John Braun (2010), стр. 201]

значим предиктор по отношение на Y в този комплексен модел, както и че Z не модерира зависимостта между Y и X2, Въпреки това Z е модератор по отношение на зависимостта между Y и X1 и по този начин можем да се твърди убедено, че взаимодействието между Y, X1 и X2 е различно при различните значения на Z, а самият регресионен подход е надежден статистически инструмент за изследване на модераторните влияния. Поглеждайки към проблемите на ЛС в базите от данни, този анализ носи допълнителна информация относно структурата на данните и взаимовръзките между променливите. Ако тази допълнителна информация се използва може да се очаква, че последващите въвеждания ще бъдат по-ефективни.

5. Липсващи стойности – Ymiss

За проверка на заложената в изследването хипотеза се симулират ЛС в променливата Y по псевдо-случаен начин[9]. По този начин се отстраняват 50% от значенията на Y. Това означава, че действащият механизъм на поява на ЛС е липсващи напълно случайно (ЛНС) [Rubin, 1987, Lazarov, 2012, 2013]. Първа стъпка в анализа е да се провери как се е повлияла зависимостта между Y~X1+X2 от симулирането на ЛС (табл. 7). При сравнение с резултатите с тези от табл. 1 може да се заключи, че случайността, която се използва за генериране на ЛС довежда до неизместени оценки на регресионните коефициенти. Тъй като проведеният анализ е само върху тези единици, за които има наблюдения по всички признаци, т.е. единиците с ЛС са отстранени, то в случая е приложен подходът поредно елиминиране на ЛС и той както отбелязва P. Alison дава неизместени оценки при МНМК под механизма ЛНС [Alison, 2002], Също така може да се отбележи, че не се наблюдава значима разлика при коефициента на детерминация при двата модела.

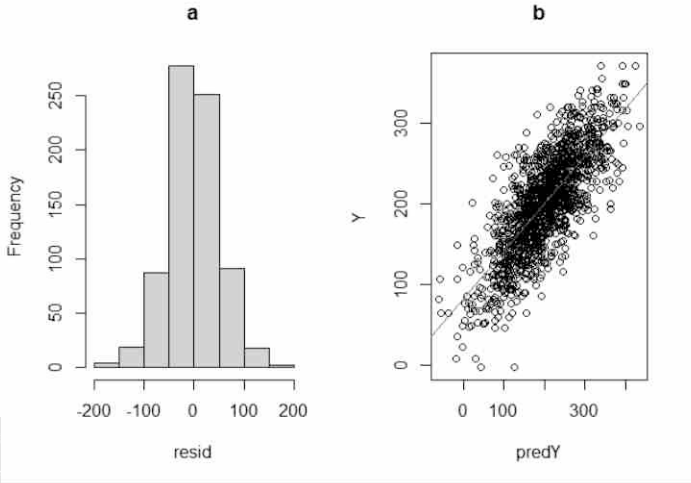
Таблица 7. Регресионни оценки при връзката Y~X1+X2 от симулираната база от данни със ЛС

```
Коефициенти:
             Оценка    Ст. грешка t-равнище  Pr(>|t|)
(Intercept) 100,05956    3,76887   26,55     <2e-16 ***
X1            0,37608    0,03857    9,75     <2e-16 ***
X2            0,80673    0,03025   26,67     <2e-16 ***
---
Кодове на стат. значимост:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Ст. грешка на остатъците: 51,05 с 747 степени на свобода,
Множествен коеф. на детерминация:  0,5751, Ажустиран мн. коеф.
на детерминация:  0,574 F-statistic: 505,5 с 2 и 747 DF, p-value: < 2,2e-16
```

Графичният вид на разпределението на остатъците от регресионната зависимост може да се види на фиг. 4a, Графичния вид показва симетричност на разпределението. Ако се добави и информацията от табл. 8 може да се заключи, че те са нормално разпределени. Зависимостта между оценените стойности и реалните стойности Y в тази регресионна зависимост може да се види на фиг. 4b.

[9] Използва се заложеният в R генератор на случайни числа.

Проведения модераторен анализ показва, че механизма на поява на ЛС, който в това изследване, бе заложен да бъде ЛНС, не нарушава взаимовръзките между модератора, предикторите и критериалната променлива (табл. 9). Модераторът Z продължава да има своята роля във взаимодействието Y~X1+X2 и тази информация може да бъде използвана в последващия анализ на ЛС. Самият модераторен анализ е проведен при хипотезата за наличие на мултиколинеарност и налага центриране на данните.



Фигура 4. Хистограма на остатъците (a) и корелационна зависимост между оценените и реалните стойности (b) при регресионната зависимост Y~X1+X2 при базата данни със симулирани 50% ЛС

Таблица 8. Описателни характеристики на остатъците при връзката Y~X1+X2 от симулираната база от данни със 50% ЛС при Y

N	Средна арит.	Ст. откл.	Медиана	Размах	Асиметрия	Ексцес	Ст. грешка
750	0	50,98	-2,29	345,65	0,016	0,64	1,8

Таблица 9. Регресионни оценки при връзката Y~X1+X2+ Z+ZX1+ZX2 от симулираната база от данни с ЛС, центрирани предиктори, при отчитане на влиянието на модератора Z

Коефициенти:

	Оценка	Ст, грешка	t-равнище	Pr(>|t|)
(Intercept)	200,64528	1,73440	115,686	< 2e-16 ***
scale(X1)	0,01497	0,06113	0,245	0,8067
scale(X2)	0,88289	0,04822	18,311	< 2e-16 ***
scale(Z)	-12,11207	7,20828	-1,680	0,0933 .
scale(ZX1)	0,46868	0,07566	6,195	9,65e-10 ***
scale(ZX2)	-0,09648	0,06077	-1,588	0,1128

Кодове на стат. значимост: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Ст. Грешка на остатъците: 47,4 с 744 степени на свобода,
Множествен коеф. на детерминация: 0,6086, Ажустиран мн. коеф.
на детерминация: 0,6059 F-statistic: 321,3 с 5 и 744 DF,
p-value: < 2,2e-16

6. Регресионно въвеждане при базата данни

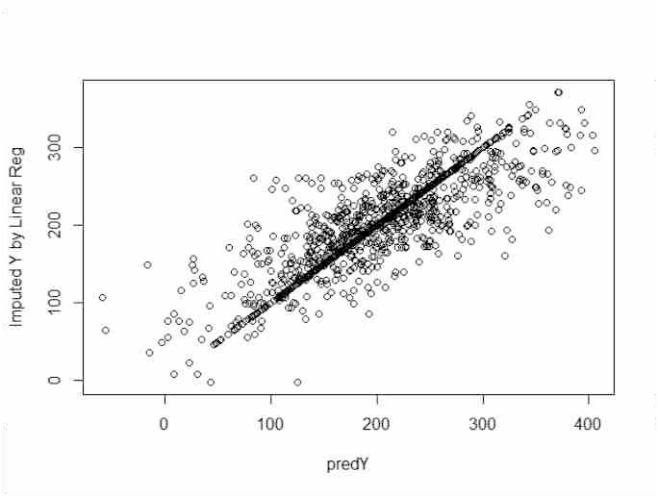
Фокусът на настоящото изследване е върху МВ, провеждано с модераторен анализ, но за разкриване тези възможности е желателно да се започне от малко по-далеч, а именно да се съпостави МВ с единичните въвеждания. За целта е проведено единично въвеждане чрез регресионния модел Y~X1+X2 представен в табл. 7. Така получената регресионна зависимост се използва за предвиждане на липсващите 750 значения на Y в базата данни, след като в нея те бяха отстранени по случен начин. Въпросът е, ако се използва регресионното въвеждане, какъв ще е получения резултат? Преди да се отговори на този въпрос е желателно да се изясни един момент в анализа – как се третират ЛС [Rubin 1987]. Практически всеки един опит те да бъдат въведени ги определя като оценки на реални, но ненаблюдавани значения. Това, че се оценяват ЛС води със себе си и несигурност от този процес, т.е. може да се говори за допълнителна несигурност, която произтича от това, че начина на намиране на ЛС е чрез оценки. В следствие на това резултативните дисперсии на тези признаци трябва да бъдат по-високи, след въвеждането. Това от своя страна означава, че всеки производен показател, като стандартна грешка и тестова характеристика при проверките на хипотези, трябва да бъде различен в сравнение с тези показатели при пълната база от данни. Стандартните грешки трябва да бъдат по-големи, а тестовите характеристики по-малки.

Проведеното регресионно въвеждане на ЛС показва проблемите на метода при анализа на ЛС. Оценките на параметрите на зависимостта (табл. 10) остават неизместени, но за сметка на това стандартните грешки са силно подценени (почти два пъти по-малки), когато се очаква те да бъдат по-големи от тези в табл. 1. Също така t-равнищата са силно надценени, което изкуствено увеличава риска от грешка от I род, т.е. вероятността да се потвърди алтернативната хипотеза, при условие, че е вярна нулевата. Това оказва влияние и на коефициентите на детерминация в модела, които също са надценени.

Таблица 10. Регресионни оценки при връзката Y~X1+X2 от база от данни след въвеждане на ЛС чрез регресионно въвеждане

```
Коефициенти:
             Оценка    Ст. грешка t-равнище  Pr(>|t|)
(Intercept) 100,05956    1,83595   54,50   <2e-16 ***
data1$X1      0,37608    0,01929   19,49   <2e-16 ***
data1$X2      0,80673    0,01553   51,94   <2e-16 ***

---
Кодове на стат. значимост:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Ст. грешка на остатъците: 36,06 с 1497 степени на свобода,
Множествен коеф. на детерминация:  0,7304, Ажустиран мн. коеф.
на детерминация:  0,7301 F-statistic: 2028 с 2 и 1497 DF,
      p-value: < 2,2e-16
```

Графичния вид на зависимостта между въведените значения за ЛС и оценените от регресията може да се види на фиг. 5. Ясно се очертава „подчинеността“ на въведените значения на регресионния модел и произтичащата оттам липса на разсейване,



Фигура 5. Зависимост между въведените значения чрез регресионния модел (по ординатата) и оценените значения от регресионната зависимост Y~X1+X2 при базата данни

7. Стохастична регресия при базата данни

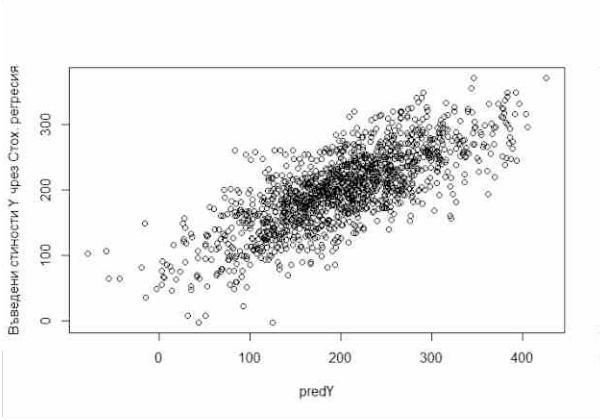
Подход, който до голяма степен преодолява недостатъците на детерминистичното, регресионно въвеждане е т.н. стохастична регресия, описан в т. 1 уравнения (1). За целта се използват характеристиките на разпределението на остатъците при регресионната зависимост Y~X1+X2 за тази част съвкупността, при която няма ЛС. От табл. 8 се вижда, че стандартното отклонение на тези остатъци е 50,98. Въпреки, че в настоящият пример разпределението на тези остатъци е нормално[10], то за добавяне на случаен компонент при въведените стойности се използва симулирано нормално разпределение със средна 0 и стандартно отклонение 50,98. За всяка ЛС се избира случайно[11] стойност от това нормално разпределение, която се добавя. Получените резултати за параметрите на регресионната зависимост Y~X1+X2 показват неизместени оценки, но стандартните грешки в не достатъчна степен отчитат несигурността произтичаща от оценките на ЛС (табл. 11, фиг. 6). Въпреки това стохастичната регресия многократно подобрява резултатите от въвеждащата процедура. Тя разбира се има един недостатък, че въведените стойности могат да не бъдат реално наблюдаеми значения в съвкупността, но в настоящия анализ това няма значимо място, защото изначално се работи със симулирани данни.

[10] Освен данните от описателните характеристики (табл. 6), които показват това, тази хипотеза се потвърждава от Колмогоров-Смирнов теста D = 0.0268, p-value = 0.8645.

[11] Тук отново става дума за псевдо-случаен подход, понеже го прави компютъра през своя генератор на случайни числа.

Таблица 11. Регресионни оценки при връзката Y~X1+X2 от база от данни след въвеждане на ЛС чрез стохастична регресия

```
Коефициенти:
            Оценка    Ст. грешка t-равнище  Pr(>|t|)
(Intercept) 98,66471    2,57949   38,25     <2e-16 ***
data1$X1     0,39730    0,02710   14,66     <2e-16 ***
data1$X2     0,81008    0,02182   37,12     <2e-16 ***
---
Кодове на стат. значимост:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Ст. грешка на остатъците: 50,67 с 1497 степени на свобода, Множествен коеф.
на детерминация:  0,5853, Ажустиран мн. коеф. на детерминация:  0,5847,
F-statistic:  1056 с 2 и 1497 DF, p-value: < 2,2e-16
```



Фигура 6. Зависимост между въведените значения чрез стохастичния регресионен модел (по ординатата) и оценените значения от регресионната зависимост Y~X1+X2 при базата данни

8. Множествено въвеждане при базата данни

Множественото въвеждане се получава като върху дадена база данни с ЛС се направят множество независими въвеждания и в резултат се получат множество алтернативни пълни бази от данни. За целта трябва да има подход, метод на базата на който да се направят тези множество въвеждания. Този метод е желателно да включва всички признаци в базата от данни, за да подсигури запазване на корелациите между тях при въведените значения. Също така е желателно той да въвежда действително наблюдавани значения, а не оценени. Да запазва разпределенията и техните характеристики при отделните признаци, но също така и многомерните разпределения и не на последно място да въвежда значения в границите на размаха на признаците. Един подобен метод, който има потенциала да отговори на всички тези условия е *„Въвеждане по близост на оцененитe стойности“*. Методът е съчетание на донорски въвеждащия подход и модели за въвеждане на данните – регресионни или други (Little 1988, Durrant и Skinner 2005a, Lazarov 2012, 2013). В своята опростена форма методът представлява аналог на метода „Най-близък съсед”[12], където разстоянието между донора и

[12] Методът „Най–близък съсед”, още наричан съответствие по функцията на разстоянията 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 методът се дефинира подходяща характеристика на разстоянията, на базата на външните променливи. За значения на единиците с ЛС се избират наблюдаваните единици с най-малки разстояния до тях, на базата на зададените (избраните) променливи.

респондента е определено на базата на оценените стойности на y_i (променливата, при която има ЛС) – чрез регресионен или друг модел. Въвеждане по близост на оценените стойности е в основата си детерминистичен метод. Следвайки принципите на Рубин е желателно въвеждане на стохастичен елемент в метода, което може да се получи, като от набора от значения, които са „близки" до оценяваните се избере едно по случаен начин.

Друга форма на метода е донорско въвеждане в подкласове, групи формирани на базата на порядъка (размаха) на оценените стойности, получени чрез избран модел за предвиждане. Този метод дава възможност за използване на всички донорските значения в подкласовете, което е форма на превенция срещу нежелано редуциране на разсейването 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 разсейването на нововъведените стойности. Методът на въвеждане по близост на оценените стойности е композитен, съчетаващ в себе си елементи от регресионното въвеждане, както и елементи от методите „Най-близък съсед" и донорското въвеждане. Той е полупараметричен метод и независимо, че работи с модел на въвеждане, не е толкова чувствителен към описанието на модела и неговата точност, за разлика например от регресионното въвеждане (Schenker и Taylor, 1996).

Прилагането на МВ върху цялата база данни без да се държи сметка за съществуващите модераторни взаимодействия дава състоятелни оценки по отношение на регресионните коефициенти в уравнението Y~X1+X2 (табл. 12). Броя на въведените бази данни е 5, а обобщените оценки се получават по формулите на Рубин (форм. 4-7). В таблицата с est са означени стойностите на осреднените оценки на регресионните коефициенти, със se са означени техните стандартни грешки, получени чрез формули 6 и 7 за оценка на разсейването на оценките. С t е означена съответстващата емпирична стойност от t –разпределението на Студент, която има df степени на свобода, с която може да се провери хипотезата за статистическата значимост на коефициентите на регресия. Проверката може да се направи и чрез Pr(>|t|) равнището, което показва вероятността да бъде вярна нулевата хипотеза[13] при наличната информация, Lo95 и hi95 са долната и горната граница при 95% доверителен интервал на коефициентите на регресията. МВ дава неизместени оценки с относително адекватна оценка на несигурността в следствие на въвеждането. Въпреки това относителния дял на промяната на оценката на стандартната грешка на коефициента на X2, остава значително по-ниска в сравнение с тази на X1. Стандартната грешка на X1 е почти двойно по-висока от тази в табл. 1 и същата пропорция се очаква при X2. Това може да се дължи на неотчитането на модераторните влияния в МВ. За да се провери тази хипотеза се провежда МВ в двете подсъвкупности по отношение на значенията на Z. Вече беше доказана хипотезата за съществуващо модераторно влияние в базата от данни след симулирането на ЛС (табл. 9), което дава основание за този анализ. В таблица 13 са представени данните от проведеното независимо МВ в отделните бази данни по отношение на значенията на Z. Броя на въвежданията е 5, методът е по близост на оценените стойности и оценките се получават по формулите на Рубин.

[13] Нулевата хипотеза съдържа условието коефициентът на регресия да бъде равен на 0.

Таблица 12. Регресионни оценки при връзката Y~X1+X2 от база от данни след въвеждане на ЛС чрез МВ без отчитане на модерацията

	est	se	t	df	Pr(>|t|)	lo 95	hi 95
(Intercept)	101,567	5,143	19,75	7,31	1,309192e-07	89,510	113,624
X1	0,386	0,044	8,86	11,33	1,986075e-06	0,290	0,481
X2	0,795	0,029	27,81	26,74	0,000000e+00	0,736	0,853

Таблица 13. Регресионни оценки при връзката Y~X1+X2 от база от данни след въвеждане на ЛС чрез МВ без отчитане на модерацията, Разделени съвкупности в зависимо от значенията на Z

При Z=0

	est	se	t	df	Pr(>|t|)	lo 95	hi 95
(Intercept)	117,591	8,545	13,761	10,662	0,000	98,710	136,473
X1	0,025	0,069	0,367	35,365	0,716	-0,115	0,165
X2	0,834	0,056	14,813	28,733	0,000	0,719	0,949

При Z=1

	est	se	t	df	Pr(>|t|)	lo 95	hi 95
(Intercept)	94,359	5,249	17,977	7,085	0	81,977	106,740
X1	0,512	0,051	10,095	8,104	0	0,395	0,629
X2	0,792	0,030	26,385	21,516	0	0,729	0,854

Въведените бази данни и изчислените обобщени оценки показват неизместени оценки на регресионните коефициенти и адекватно отчитане на несигурността относно оценките на ЛС, Това дава основание да се провери дефинираната основна хипотеза в изследването като се използва описаната чрез фиг. 2 методика. Следователно първо да се използват подсъвкупностите по отношение на Z за да се направят независимите въвеждания в базите от данни. На следваща стъпка тези бази данни се обединяват, като двете половини на съответстващите си такива да се събират в една. Накрая, да се използват формули 4-7 за да се получат обединените оценки за регресията. Резултатите са поместени в табл. 14.

Таблица 14. Регресионни оценки при връзката Y~X1+X2 от база от данни след въвеждане на ЛС чрез МВ без отчитане на модерацията, Обединени съвкупности след МВ в зависимост от значенията на Z

	est	se	t	df	Pr(>|t|)	lo 95	hi 95
(Intercept)	102,086	3,940	25,91	13,306	8,890666e-13	93,595	110,578
X1	0,345	0,037	9,26	20,567	8,853998e-09	0,267	0,422
X2	0,810	0,040	20,13	8,335	2,316735e-08	0,718	0,902

Проведеният анализ показва адекватна и пропорционална оценка на несигурността при оценките на регресията при двата предиктора (виж табл,1 и табл, 13). Оценките се запазват неизместени, емпиричните характеристики са по-ниски и грешките от първи род са

минимизирани. Това дава основание на се заключи, че заложената хипотеза в изследването е доказана, а именно че когато в базата данни има модераторно влияние на дадена променлива Z, то МВ съобразено с това модераторно влияние дава по-добри резултати в сравнение с МВ без отчитане на модератора.

За още една проверка на хипотезата е и модераторният регресионен анализ, включващ в себе си отчитане на взаимодействието между Z и X1 и Z и X2. В случай, че се проведе МВ без да се отчита влиянието на Z, т.е. без то да бъде проведено поотделно в двете подсъвкупности, се наблюдава появата на значимо влияние на X1 (табл. 15) , което липсва при изходните данни (табл. 4). Също може да се забележи, че влиянието на Z е статистически незначимо, за разлика от това при изходните данни, а грешките в недостатъчна степен отчитат несигурността,

Таблица 15. Регресионни оценки при връзката Y~X1+X2+ Z+ZX1+ZX2 чрез МВ, центрирани предиктори, без отчитане на влиянието на модератора Z

	est	se	t	df	Pr(>|t|)	lo 95	hi 95
(Intercept)	200,552	1,709	117,369	17,004	0,000	196,947	204,157
scale(X1)	0,174	0,054	3,197	34,496	0,003	0,063	0,284
scale(X2)	0,881	0,043	20,618	37,401	0,000	0,795	0,968
scale(Z)	-7,158	6,391	-1,120	32,425	0,271	-20,168	5,853
scale(ZX1)	0,234	0,076	3,065	15,977	0,007	0,072	0,396
scale(ZX2)	-0,057	0,051	-1,118	53,465	0,269	-0,158	0,045

От друга страна ако МВ се проведе поотделно в двете подсъвкупности определени от значенията на Z и после тези подсъвкупности се обединят в една, резултатите са значително по-добри (табл. 16). При този подход влиянието на X1 се запазва незначимо, както е при изходните данни. И при този подход се наблюдава незначимо влияние на Z и това е един недостатък на метода. Грешките на оценките са адекватно по-големи, отчитайки несигурността на ЛС. Това дава основание да се предпочете подходът за въвеждане чрез отчитане на влиянието на Z, пред класическия вариант на МВ.

Таблица 16. Регресионни оценки при връзката Y~X1+X2+ Z+ZX1+ZX2 чрез МВ, центрирани предиктори, при отчитане на влиянието на модератора Z

	est	se	t	df	Pr(>|t|)	lo 95	hi 95
(Intercept)	199,391	1,982	100,595	154,631	0,000	195,476	203,307
scale(X1)	-0,056	0,077	-0,728	51,984	0,470	-0,211	0,098
scale(X2)	0,455	0,055	8,331	387,979	0,000	0,348	0,562
scale(Z)	-3,465	8,815	-0,393	63,304	0,696	-21,078	14,148
scale(ZX1)	0,281	0,100	2,819	31,975	0,008	0,078	0,484
scale(ZX2)	-0,067	0,067	-1,000	375,701	0,318	-0,200	0,065

9. Заключение

Проведеното изследване е показателно по отношение на възможностите на МВ. Този подход е желателен за да може да се получават адекватни оценки на стандартните грешки на параметрите при анализите. Важен компонент при тези оценки, когато се работи с ЛС, е несигурността породена от тяхното оценяване. В настоящото изследване се сравниха два подхода за МВ. Единият, който може да се нарече класически, при който въвеждането се провежда в цялата база данни наведнъж и друг, алтернативен, който се базира на модераторен анализ и може да се нарече модераторно МВ. Също така се разгледаха и резултатите от регресионно и стохастично регресионно въвеждане. Последните два метода са от типа единични въвеждания и резултатите от тях имат значителни недостатъци по отношение на оценките на разсейванията на оценките на базата на въведените стойности. Резултатите от изследването показаха, че ако анализът покаже наличие на модератор в базата данни е желателно МВ да се базира на разделяне на съвкупността в зависимост от значенията на този модератор. Това се обосновава и от смисъла на модератора, а именно че при различните негови значения връзката между предикторите и критериалната променлива е различна. Издигната хипотеза за подобряване на резултатите от МВ, когато се отчете модераторнито влияние се доказа. Подобряването на резултатите се изразява в пропорционално и адекватно повишаване на стандартните грешки, което се изисква заради несигурността при оценките на ЛС. В противен случай те се третират като реални значения, а не като въведени, т. е. оценени. Въпреки това е необходимо да се положат още усилия по отношение на подхода. Оказа се, че когато базата от данни се раздели по значенията на модератора неговото влияние като предиктор може да се обезличи и не може да се компенсира от въвеждащата процедура. Това определено е недостатък, който се забеляза в настоящото изследване. За целта са необходими повече симулации, за изследване на тази особеност. Този недостатък се наблюдава и при класическия подход на МВ, и затова може да се твърди, че модераторното МВ продължава да има предимство, поне в настоящата постановка на задачата. Също така е наложителен анализ и с други методи за въвеждане, освен използвания по близост на оценените стойности.

Използвана 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. Baron, R, M,, & Kenny, D, A, (1986), „The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations“, Journal of Personality and Social Psychology, 51, 1173–1182,
4. Durrant, G,B,, Skinner, C, (2005a): Using Missing data Methods to Correct for Measurement Error in a Distribution Function, Survey Methodology
5. Durrant, G,B,, 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. Enders, C, K, (2010) Applied missing data analysis, The Guilford Press
7. John Maindonald, W, John Braun (2010) Data Analysis and Graphics Using R: An Example-Based Approach, Cambridge University Press
8. Lazarov, D (2013), Specifichni vazmozhnosti pri analiza na lipsvashti stoinosti pri empirichnite izsledvaniya, Znanieto - tradicii, inovacii, perspektivi - BFU, pp, 514-519
9. Lazarov, D (2012), Metodi za analiz na lipsvashti stoinosti, Chast I, Spisanie “Bizines posoki”, No, 2/2012, pp, 84-95
10. Little, R, J, A, (1988), A test of missing

completely at random for multivariate data with missing values, Journal of the American Statistical Association,83, 1198–1202,
11.Little, R,J,A, Rubin, D,B, (1987), Statistical analysis with missing data, New York: Wiley,
12. Little, R,J,A, Rubin, D,B, (2002), Statistical Analysis with Missing data - 2nd ed,, New Jersey: Wiley,
13. Rubin, D,B, (1987), Multiple Imputation for Nonresponse in Survey, New York: Wiley,
14. Rubin, D, (1996), Multiple imputation after 18+ Years, Journal of the American Statistical Association 91 (June): 473–89,
15. Sinharay, S,, Stern, H, S,, & Russell, D, (2001), The use of multiple imputation for the analysis of missing data, Psychological Methods, 6, 317–329,
16. Scheffer, J, (2002), Dealing with Missing data, Research Letters in the Information and Mathematical Sciences 3, 153-160
17. Schenker, N,, Taylor, J, (1996), "Partially Parametric Techniques for Multiple Imputation," Computational Statistics and Data Analysis, 22, pp, 425–446