# SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

## SLOVAK STATISTICS and DEMOGRAPHY

1/2025
ročník/volume 35

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 3

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 41 – 65

Dátum vydania/Publication date: 15. január 2025/January 15, 2025

**Roman PAVELKA, Štatistický úrad Slovenskej republiky**

# MOŽNOSTI POUŽITÍ METOD MNOHONÁSOBNÉ IMPUTACE V PROSTŘEDÍ SYSTÉMU SAS

# POSSIBILITIES OF USING MULTIPLE IMPUTATION METHODS IN THE SAS SYSTEM ENVIRONMENT

## ABSTRAKT

Chybějící hodnoty představují u většiny statistických analýz komplikace. Pozorování s nevyplněnými hodnotami proměnných (nazývané také jako neúplné případy) jsou ze statistických analýz zjištěných pozorování ve většině statistických softwarů implicitně vyřazovány. Použití úplných případů (pozorování s kompletně vyplněnými hodnotami u všech proměnných) je sice jednoduché, ale zpravidla vykoupené ztrátou informace v důsledku vyloučení neúplných případů. Navíc vyloučení neúplných pozorování ze statistických analýz také ignoruje potenciální systematické rozdíly mezi odhady a skutečnými hodnotami a výsledná statistická inference nemusí být použitelná na sledovanou populaci všech statistických jednotek (všech případů), zejména v podmínkách menšího počtů kompletních případů. Z tohoto důvodu je proto důležité analyzovat nejen data pozorovaná, ale i rozpoznat mechanismus chybění (statistický model resp. rozdělení pravděpodobnosti) neúplných dat s cílem jejich vhodného doplnění přijatelnými (imputovanými) hodnotami. Ačkoli existuje mnoho různých metod pro práci s nekompletními daty, jednou z nejvíce významných metod řešení otázek neúplnosti zjištěných dat se stala metoda mnohonásobné imputace. I v prostředí analytického systému SAS použití metody mnohonásobné imputace představuje jednu z možností řešení neúplných dat, a právě toto bude náplní předkládaného článku.

## ABSTRACT

Missing values represent a complication in most statistical analyses. Observations with incomplete values of variables (also called incomplete cases) are implicitly excluded from statistical analyses of the detected observations in most statistical softwares. The use of complete cases (observations with completely filled values for all variables), is simple, though usually redeemed by the loss of information due to the exclusion of incomplete cases. Moreover, excluding incomplete observations from statistical analyses also ignores potential systematic differences between estimates and actual values, and the resulting statistical inference may not be applicable to the population of interest for all statistical units (all cases), especially in conditions of fewer complete cases. For this reason, it is therefore important to analyse not only the observed data, but also to recognise the 'missing' mechanism (statistical model or probability distribution) of incomplete data in order to appropriately fill them with acceptable (imputed) values. Although there are many different methods for working with incomplete data, the multiple imputation method has become one of the most important methods for dealing with the incompleteness of the observed data. Even in the SAS analytical system environment, the use of the multiple imputation method represents one of the options for dealing with incomplete data, and this is the content of the present paper.

## KLÍČOVÁ SLOVA

analytický systém SAS, inference, mechanismus chybění, mnohonásobná imputace

**KEY WORDS**
analytical SAS system, inference, mechanism of missingness, multiple imputation

## 1. ÚVOD DO PROBLEMATIKY MNOHONÁSOBNÉ IMPUTACE

Programový systém SAS při řešení otázek neúplnosti zjištěných dat uplatňuje více strategií a metod. Některé procedury SAS uplatňují metodu všech dostupných případů, tj. do analýzy jsou zařazeny všechny případy (zpravodajské jednotky) s použitelnou informací. Například procedura CORR (analyzuje míry těsnosti závislosti proměnných) odhaduje průměr proměnné tak, že použije všechny případy ve sledované proměnné ignorujíc možné chybějící hodnoty v ostatních proměnných. Podobný způsob procedura uplatňuje i v rámci odhadů korelace, kde jsou začleněny vyplněné dvojice hodnot u vybraných sledovaných proměnných bez ohledu na hodnoty ostatních proměnných. I když se může jednat o lepší využití dostupných údajů, výsledná korelační matice nemusí být pozitivně definitní[1].

Jinou strategií programového systému SAS v otázce chybějících údajů je jednoduchá imputace, při níž je každá chybějící hodnota nahrazována jedinou hodnotou. Například imputace realizovaná (i podmíněným) průměrem, kdy se každá chybějící hodnota analyzované proměnné nahrazuje (ne)podmíněným průměrem z ostatních vyplněných hodnot této proměnné. Imputace průměrem doplňuje chybějící hodnoty jako by byly hodnoty, známy v rámci analýzy kompletních případů. Jednoduchá imputace však nezohledňuje nejistotu při predikci neznámých chybějících hodnot (Rubin 1987, s. 13) a výsledné odhadované rozptyly odhadů parametrů jsou vychýlené (zkreslené) směrem k nule.

Nejvýznamnější imputační metodou v programovém systému SAS je metoda mnohonásobné imputace. Namísto doplnění jedné hodnoty za každou chybějící hodnotu se každá chybějící hodnota nahradí množinou přijatelných hodnot, které představují nejistotu ohledně imputované hodnoty (Rubin, 1976, 1987). Soubory mnohonásobně imputovaných dat se pak analyzují pomocí standardních postupů pro úplná data. Lineární kombinací výsledků analýz imputovaných dat se odhadne výsledná doplněná hodnota včetně odhadu rozptylu této nahrazené hodnoty.

Mnohonásobnou imputaci v programovém systému zajišťuje procedura MI, která svou funkcionalitou vytváří soubory imputovaných dat pro neúplná (i mnohorozměrná) data. Procedura aplikuje metody, které zahrnují vhodnou variabilitu imputovaných hodnot napříč zvoleného počtu imputací. Systém SAS nabízí více metod mnohonásobné imputace, jejichž volba závisí na typu imputované proměnné (kvalitativní proměnná nebo kvantitativní proměnná) a způsobu uspořádání neúplných dat (jedná se o tzv. vzory chybění) a na dalších předpokladech uvedených dále. Statistická inference imputovaných údajů se v systému SAS realizuje pomocí procedury MIANALYZE, která slouží k odhadům imputovaných hodnot včetně odhadů jejich variability.

Metody mnohonásobné imputace v prostředí analytického systému SAS se nesnaží odhadnout každou chybějící hodnotu pomocí simulovaných hodnot. Mnohonásobná imputace se naopak provádí generováním náhodného výběru chybějících hodnot s předem definovaným rozdělením, který umožňuje platné statistické závěry

[1] *Symetrická matice $\boldsymbol{A}$ řádu $n$ se nazývá pozitivně definitní, jestliže pro každý nenulový sloupcový vektor $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ platí $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$.*

adekvátně odrážející nejistotu způsobenou chybějícími hodnotami (SAS Institute Inc., 2023, s. 6407).

## 2. MNOHONÁSOBNÉ IMPUTACE – HISTORIE, PRINCIP A PŘEDPOKLADY

### HISTORIE MNOHONÁSOBNÉ IMPUTACE

Použití mnohonásobné imputace (v dalším textu zkratkou jako „MI“) ke zpracování neúplných údajů poprvé navrhl Rubin v roce 1978 ve svém článku (Rubin, 1978). V souvislosti s rozsáhlými výběrovými šetřeními, v nichž měly být údaje shromážděné v jedné studii použity potenciálně velkým počtem výzkumníků pro řadu různých analýz, přístup k chybějícím údajům rozvinul dále Rubin ve své práci (Rubin, 1987). Metoda MI však v té době zůstávala neznámá, a tedy nevyužívaná především kvůli nedostatku vhodných výpočetních zařízení. S příchodem rychlejších počítačů v posledních desetiletích se však MI stala poměrně populární v kontextu výběrových i nevýběrových šetření (Rubin, 1996; Schafer, 1997a; Schafer & Olsen, 1998). MI také byla dobře rozpracována v řadě článků, které porovnávaly přístupy k řešení chybějících dat v rámci modelování strukturálních rovnic (např. Newsom, 2023). Dalším důvodem popularity MI v poslední době je skutečnost, že po nástupu algoritmu EM[2] koncem 70. let 20. století začali statistici považovat chybějící hodnoty za zdroj variability, kterou je třeba zprůměrovat (místo toho, aby chybějící hodnoty považovali za obtíž). Toto průměrování MI dokáže provést velmi jednoduchým způsobem.

### PRINCIP MNOHONÁSOBNÉ IMPUTACE

MI je v podstatě rozšířením myšlenky jednoduché imputace, kdy je každá chybějící hodnota nahrazena souborem $m > 1$ přijatelných hodnot, aby se vytvořilo $m$ zřejmě úplných souborů dat. Těchto $m$ datových souborů se pak analyzuje pomocí standardních statistických metod a výsledky se kombinují pomocí technik navržených Rubinem (Rubin, 1987). Tímto způsobem se získají odhady parametrů a standardní chyby, které adekvátně zohledňují nejistotu způsobenou chybějícími hodnotami dat.

Ve většině aplikací stačí k dosažení uspokojivých výsledků pouhých tři až pět imputací. Podle (Rubin, 1987, s. 147) platí, že efektivita odhadů založených na $m$ imputacích je přibližně

$$RE = \left(1 + \frac{\lambda}{m}\right)^{-1} \tag{2.1}$$

kde RE označuje relativní efektivitu MI, $\lambda$ vyjadřuje podíl chybějící informace pro odhadované parametry a $m$ je předem určený počet imputací. Pro ilustraci při podílu chybějící informace $\lambda = 0{,}3$ jsou k dosažení relativní efektivity $RE = 91\%$ potřebné pouze tři imputace a k dosažení relativní efektivity $RE = 94\%$ je potřeba pět imputací.

Statistická inference pomocí MI zahrnuje tři různé fáze:

1. Chybějící údaje se doplní $m$-krát, aby se vytvořilo $m$ úplných souborů dat.
2. $m$ úplných datových souborů se analyzuje pomocí standardních postupů.

*[2] EM-algoritmus, který vznikl v sedmdesátých letech dvacátého století, je jeden z často používaných algoritmů ve statistice. EM-algoritmus v prvním, tzv. E-kroku nahradí logaritmus věrohodnostní funkce její podmíněnou střední hodnotou při daných hodnotách data a parametrů a ve druhém kroku (tzv. M-kroku) hledá odhady parametrů maximalizací této podmíněné střední hodnoty.*

3. Výsledky analýz z $m$ úplných datových souborů se zkombinují za účelem vyvození statistických závěrů.

## PŘEDNOSTI MNOHONÁSOBNÉ IMPUTACE

Jak již bylo zmíněno, MI byla vyvinuta v kontextu rozsáhlých průzkumů, v nichž mají být údaje shromážděné v rámci jednoho šetření použity potenciálně velkým počtem výzkumných pracovníků pro řadu různých analýz. V tomto kontextu je ideální, když mnohonásobná imputace (neúplných) dat může být realizována autorem statistického šetření (který má obvykle přístup k více informacím než jednotliví uživatelé) a všichni uživatelé pak mohou analyzovat výsledné úplné soubory dat pomocí standardního statistického softwaru.

I když toto prvotní zpracování zjištěných dat platí i pro metody jednoduché imputace, většina metod jednoduché imputace má svá omezení. I kdyby bylo možné chybějící hodnoty imputovat metodami jednoduché imputace tak, aby nedošlo ke zkreslení rozdělení proměnných a vztahů mezi proměnnými, imputované datové soubory získané nahrazením každé chybějící hodnoty nějakým bodovým odhadem by stále nedokázaly zohlednit nejistotu chybějících údajů, a tudíž by podhodnocovaly variabilitu datového souboru. V důsledku toho by standardní chyby odhadů parametrů byly podhodnoceny a míra chyb typu I pro jakýkoli test hypotézy by byla vyšší než předem stanovená hladina významnosti testu (tj. test by byl pozitivně zkreslený). Pomocí MI lze kombinovat výsledky z řady analýz úplných údajů a řešit tak nejistotu způsobenou chybějícími hodnotami.

Velkou výhodou metody MI oproti ostatním odhadům, které se mohou také vhodně vypořádat s odhady parametrů v podmínkách chybějících dat, je jednoduchost metody pro většinu praktických situací. Metody odhadu založené na věrohodnostní funkci jsou specifické pro daný imputační problém a mohou vyžadovat zcela odlišné výpočetní postupy pro integraci chybějících údajů pro různé modely aplikované na stejný soubor dat. Naproti tomu v případě MI mohou stejné soubory imputovaných dat používat různí uživatelé pro různé typy analýz pomocí jakéhokoli populárního statistického softwaru, aniž by se tito uživatelé museli zabývat řešením problému chybějících údajů.

## NEZBYTNÉ PŘEDPOKLADY MNOHONÁSOBNÉ IMPUTACE

Mnohonásobné imputace se vytvářejí na základě předpokladu určitého imputačního modelu, na jehož správnosti závisí úspěch či neúspěch této imputační metody. Podle (Russell, Sinharay, 2001) aspekty analýzy neúplných dat, pro které jsou v MI vyžadovány určité předpoklady, jsou:

- pravděpodobnostní model pro hodnoty dat,
- apriorní rozdělení pro parametry modelu dat a
- mechanismus neodpovědí.

Ve výběrovém souboru složeném ze zjištěných údajů o rozsahu $n$ jednotek jsou zpravidla definovány dva druhy proměnných. První druh zahrnuje proměnné, které popisují charakteristiky jednotek zajímavé pro účely statistického zjišťování, a to plně pozorovatelné vysvětlující proměnné (kovariáty) $\boldsymbol{X}$ a vysvětlované (závislé na vysvětlujících proměnných) proměnné $\boldsymbol{Y}$, které jsou předmětem výzkumného zájmu. Sledované proměnné $\boldsymbol{Y}$, které u každé jednotky reprezentuje náhodný vektor $p$

proměnných $\boldsymbol{Y}_i = (Y_1, \ldots, Y_p)$, jsou náhodné a $\boldsymbol{y}_i = (y_1, \ldots, y_p)$ představuje realizaci $\boldsymbol{Y}_i$ pro $i$-tou respondující jednotku z výběrového souboru rozsahu $n$, $i = 1, \ldots, n$.

**Datový model**. Prvním a nejdůležitějším krokem při realizaci doplnění chybějících dat metodou mnohonásobných imputací za chybějící hodnoty v souboru dat je potřebný předpoklad pravděpodobnostního modelu, který vztahuje úplná data $\boldsymbol{Y}$ (tj. kombinaci pozorovaných hodnot $\boldsymbol{Y_{obs}}$ a chybějících hodnot $\boldsymbol{Y_{mis}}$) k množině odhadovaných parametrů. Na vybraném pravděpodobnostním modelu je založen výpočet pozorované věrohodnostní funkce na základě pozorovaných hodnot $\boldsymbol{Y_{obs}}$. Na rozdíl od četnostních metod hledání maxima věrohodnostní funkce, odhadované chybějící hodnoty nejsou považovány za konstanty, nýbrž v procesu mnohonásobné imputace je odhadováno rozdělení chybějících hodnot $\boldsymbol{Y_{obs}}$. Pomocí tohoto pravděpodobnostního modelu a apriorního rozdělení odhadovaných parametrů (viz následující části textu) se najde prediktivní rozdělení $p(\boldsymbol{Y_{mis}}|\boldsymbol{Y_{obs}})$ pro chybějící hodnoty $\boldsymbol{Y_{mis}}$ podmíněné pozorovanými hodnotami $\boldsymbol{Y_{obs}}$ .Následně se z tohoto prediktivního rozdělení vygenerují imputované hodnoty. Metoda mnohonásobné imputace je tedy založena na tzv. bayesovském modelování, které je podrobněji popsáno například v článku (Pavelka, 2024a).

Předpokládaný pravděpodobnostní model by měl zahrnovat všechny znalosti o procesu, který zjištěná data generují. Pro spojité proměnné je nejvhodnějším modelem vícerozměrný normální model. Jednou z klíčových výhod je, že tento pravděpodobnostní model je výpočetně zvládnutelný. Ukazuje se (viz např. Schafer, 1997, s. 147-148, a tam uvedené odkazy), že vícerozměrný normální model poskytuje celkem přijatelné výsledky i v případě, že proměnné jsou binární nebo kategoriální, přičemž imputace se provádějí za předpokladu normálního modelu a imputované hodnoty se pak zaokrouhlují na nejbližší kategorii. Pokud existuje proměnná, která se nezdá být normálně rozdělena, lze ji transformovat na normální proměnnou a imputované hodnoty transformovat zpět na původní škálu. Mezi další modely, které analytici dat používají, patří logaritmicko-lineární model pro kategoriální proměnné, směs logaritmicko-lineárního modelu a vícerozměrného normálního modelu pro smíšené soubory spojitých a kategoriálních dat a hierarchický lineární model (Bryk & Raudenbush, 1992).

**Apriorní rozdělení**. Statistický přístup používaný k provedení metody vícenásobné imputace založené na modelu je obvykle bayesovský, a proto je nutné určit apriorní rozdělení parametrů pro provedení analýz. Apriorní rozdělení a věrohodnostní datový model poskytnou prediktivní rozdělení $p(\boldsymbol{Y_{mis}}|\boldsymbol{Y_{obs}})$ pro chybějící hodnoty podmíněné pozorovanými hodnotami $\boldsymbol{Y_{obs}}$, ze kterého lze generovat imputace. Obvykle se pro pohodlí při provádění MI používají tzv. neinformační apriorní rozdělení[3]. Kvůli subjektivitě spojené s volbou prioritních rozdělení byly bayesovské metody někdy kritizovány. Volbou neinformativního apriorního rozdělení lze subjektivitě zabránit. U mnoha analýz dat na apriorních (předběžných) rozděleních téměř nezáleží, protože i při středně velkých velikostech vzorků dává každé rozumné apriorní rozdělení

*[3] Apriorní rozdělení vyjadřuje prvotní znalosti o odhadovaných parametrech ještě předtím, než se analyzují pozorovaná data. Neinformativní apriorní rozdělení se užívá za podmínek, kdy nemáme téměř žádné předběžné znalosti o parametrech. Pokud apriorní rozdělení přímo nevyplývá z předběžných znalostí, často se používají konjugovaná rozdělení. Konjugovaná rozdělení zajišťují, že funkcionální forma výsledného aposteriorního (prediktivního) rozdělení je stejná jako rozdělení apriorní. Apriorní rozdělení sehrává významnou úlohu zejména v případech, kdy je vzorek pozorovaných dat příliš malý.*

v podstatě stejné výsledky. Pokud je velikost vzorku malá, je dobré před vyvozením závěrů provést analýzu při různých rozděleních předpovědí a zjistit, zda se výsledky změní.

**Mechanismus chybění**. MI je metoda založená na modelu, který předpokládá, že chybějící údaje jsou náhodné (v literatuře označované jako *Missing At Random* – zkratkou MAR). Podle (Pavelka, 2024b) při náhodném chybění rozdělení chybějících údajů závisí pouze na zjištěných (pozorovaných) hodnotách. Předpoklad MAR umožňuje využít vztahy mezi proměnnými, které jsou zřejmé z pozorovaných dat, k získání imputovaných hodnot pro chybějící pozorování. Například ve studii, v níž jsou dostupné informace o mnoha proměnných na pozadí (kovariáty $\boldsymbol{X}$) s chybějícími hodnotami pouze pro jednu proměnnou $\boldsymbol{Y}$, předpoklad MAR (chybějící hodnota proměnné $\boldsymbol{Y}$ závisí pouze na $\boldsymbol{X}$) povede ke zjištění, jak $\boldsymbol{Y}$ závisí na $\boldsymbol{X}$ z úplných případů. Tento vztah závislosti (nejčastěji regresní přímka spolu s reziduální standardní chybou) lze použít k předpovědi chybějících hodnot $\boldsymbol{Y}$ z odpovídajících hodnot $\boldsymbol{X}$. V mnohých analýzách nelze platnost předpokladu MAR ověřit a nedají se vyloučit ani jiné neměřitelné faktory, které způsobily chybějící hodnoty sledované proměnné.

**Matematická idea metody MI**. Jako první imputace neúplných dat metodou MI navrhl Rubin (Rubin, 1987). Rubinem navržený přístup je bayesovské povahy a zpopularizoval jej Schafer (Schafer, 1997), který poskytl podrobné algoritmy pro vytváření mnohonásobných imputací v různých situacích. Základní myšlenkou MI je získat prediktivní rozdělení chybějících hodnot vzhledem k pozorovaným datům. Predikční rozdělení $p(\boldsymbol{Y_{mis}}|\boldsymbol{Y_{obs}}, \boldsymbol{X})$ lze zapsat jako

$$\begin{aligned} p(\boldsymbol{Y_{mis}}|\boldsymbol{Y_{obs}}, \boldsymbol{X}) &= \int_{\boldsymbol{\theta}} p(\boldsymbol{Y_{mis}}, \boldsymbol{\theta}|\boldsymbol{Y_{obs}}, \boldsymbol{X}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(\boldsymbol{Y_{mis}}|\boldsymbol{Y_{obs}}, \boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{Y_{obs}}, \boldsymbol{X}) d\boldsymbol{\theta} \end{aligned} \tag{2.2}$$

Chybějící hodnoty se zpravidla imputují ve 2 krocích. Nejprve se generuje hodnota parametru z pozorovaného aposteriorního rozdělení $p(\boldsymbol{\theta}|\boldsymbol{Y_{obs}}, \boldsymbol{X})$. Generování hodnoty parametru vyžaduje provedení bayesovské analýzy chybějících dat zpravidla simulací markovských řetězců (Gelman, Carlin, Stern, Dunson, Vehtari & Rubin, 1995). Vygenerovaná hodnota parametru se využije k simulaci vektoru chybějících dat z podmíněného aposteriorního rozdělení $p(\boldsymbol{Y_{mis}}|\boldsymbol{Y_{obs}}, \boldsymbol{X}, \boldsymbol{\theta})$.

**Imputační model a analytický model MI**. Klíčovým rysem metody MI je oddělení modelu použitého k získání imputací od modelu použitého pro analýzu souboru dat. Ačkoli imputace obvykle provádí osoba, která shromáždila data, konečnou analýzu může provádět mnoho dalších uživatelů, kteří sdílejí soubor dat. Osoba, která data sbírá, má samozřejmě mnohem lepší znalosti o datech a je pravděpodobně nejlepší osobou pro provedení imputací. Budoucí uživatelé (datoví analytici) mohou použít k provedení konečné analýzy jakoukoli techniku úplných dat (která je k dispozici v jakémkoli standardním statistickém softwaru) a nemusejí si dělat starosti s mechanismem chybějících údajů, protože mají v podstatě přístup k souborům úplných údajů.

Aby analýzy dat doplněných metodou MI poskytovaly uspokojivé výsledky, musí imputační model být kompatibilní s modely uplatněnými v analýzách kompletních dat. Například podle (Schafer & Olsen, 1998) pokud je zájmová proměnná $\boldsymbol{Y}$ imputována

podle normálního modelu zahrnujícího proměnnou $\boldsymbol{X}_1$ a při analýzách kompletních dat po imputaci výzkumník k předpovědi proměnné $\boldsymbol{Y}$ použije lineární regresní model obsahující vysvětlující proměnné $\boldsymbol{X}_1$ a $\boldsymbol{X}_2$ (kdy proměnná $\boldsymbol{X}_2$ nebyla součásti imputačního modelu), odhadovaný koeficient pro vysvětlující proměnnou $\boldsymbol{X}_2$ může být vychýlený směrem k nule. Možné vychýlení odhadu regresního koeficientu pro vysvětlující proměnnou $\boldsymbol{X}_2$ je zapříčiněno tím, že doplňované hodnoty byly imputovány modelem neuvažujícím vztah mezi vysvětlovanou proměnnou $\boldsymbol{Y}$ a vysvětlující proměnnou $\boldsymbol{X}_2$. Z tohoto důvodu je vhodné do imputačního modelu zahrnout co nejvíce vysvětlovaných proměnných.

## 3. MOŽNOSTI SYSTÉMU SAS PŘI IMPUTACI HODNOT METODOU MI

### STRUČNÝ PŘEHLED PROCEDURY MNOHONÁSOBNÉ IMPUTACE MI

V programovém systému SAS je procedurou mnohonásobné imputace procedura MI, která vytváří soubory mnohonásobně imputovaných dat pro neúplné $p$-rozměrné vícerozměrné proměnné. Používá metody, které zapracovávají vhodnou variabilitu napříč $m$ imputacemi. Volba metody imputace závisí na tzv. vzorech chybějících údajů a na typu imputované proměnné. Vzor chybějících údajů ilustruje konfiguraci pozorovaných hodnot a jejich chybění v zjištěném statistickém souboru (Kang, 2013). Za účelem zjištění vzoru chybění lze řádky a sloupce datové matice seřadit tak, aby vznikly charakteristické vzory chybějících údajů.

Monotónní vzor chybění vzniká v případě, pokud v datovém souboru s proměnnými $(\boldsymbol{Y}_1^T, \dots, \boldsymbol{Y}_p^T)$ existuje pro $i$-tou zpravodajskou jednotku určitá proměnná $Y_{ij^*}$, od které jsou všechny následující proměnné nepozorované, tj. nepozorované proměnné $Y_{ik}$ jsou pro $k > j^*$, $k = j + 1, \dots, p$. U datových souborů s monotónními chybějícími vzory lze proměnné s chybějícími hodnotami postupně imputovat pomocí vysvětlujících proměnných (plně pozorovaných kovariát) vytvořených z odpovídajících souborů předchozích proměnných. K imputování chybějících hodnot pro spojitou proměnnou lze použít regresní metodu (Rubin, 1987, s. 166-167), metodu prediktivní shody průměrů (Raghunathan, Lepkowski, Van Hoewyk a Solenberger, 2001) nebo metodu propensitního skóre (Rubin, 1987, s. 124, 158; Ling, Montez-rath, Mathur, Kapphahn, Desai, 2020). Imputování chybějících hodnot pro klasifikační proměnnou lze realizovat metodou logistické regrese, pokud má klasifikační proměnná binární, nominální nebo ordinální odpověď, nebo metodou diskriminační funkce, pokud má klasifikační proměnná binární nebo nominální odpověď.

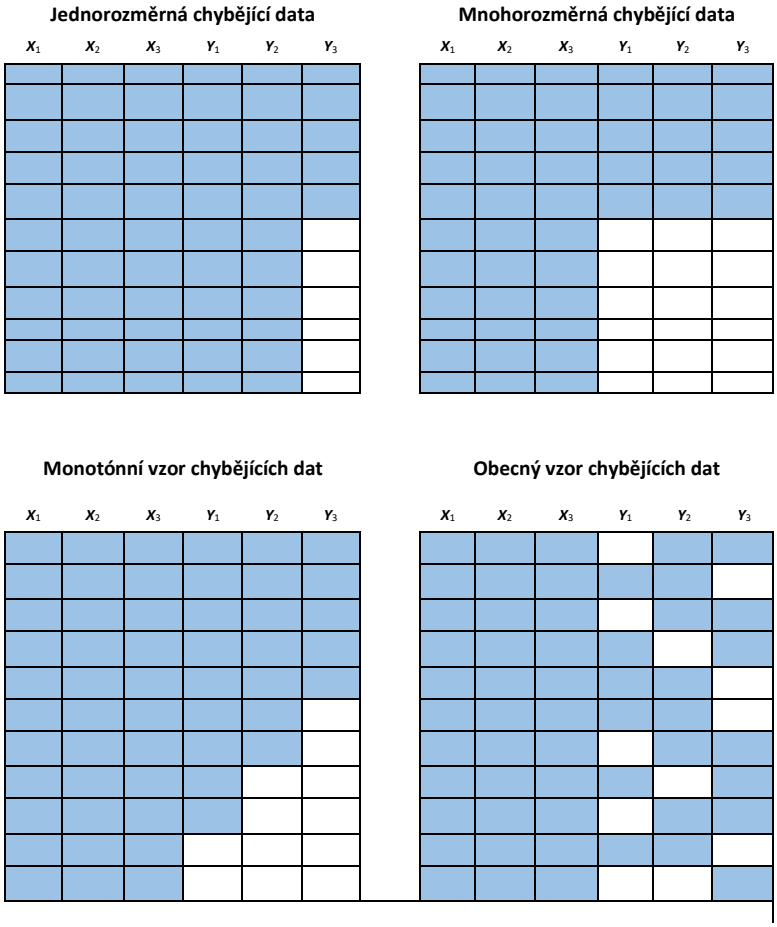
Pro datové soubory s libovolnými chybějícími vzory lze k imputování chybějících hodnot použít některou z následujících metod: metodu Markovova řetězce Monte Carlo (dále zkratkou MCMC) (Schafer, 1997), která předpokládá vícerozměrnou normalitu, nebo metodu plně podmíněné specifikace (v anglickém originálu Fully Conditional Specification – zkratkou FCS) (Brand 1999; van Buuren, 2007)[4], která předpokládá existenci společného rozdělení pro všechny proměnné. Metodu MCMC lze použít k imputování buď všech chybějících hodnot, nebo jen tolika chybějících hodnot, aby imputované datové soubory měly monotónní vzor chybějících údajů. Dosažením monotónního vzoru chybějících dat se zvětšuje flexibilita při výběru imputačních modelů, jako je například metoda monotónní regrese, která nepoužívá

*[4] Přístup FCS v mnohonásobné imputaci spočívá v imputování dat podle jednotlivých proměnných zadáním imputačního modelu pro každou proměnnou.*

Markovovy řetězce. Pro každou imputovanou proměnnou je možné také zadat jinou sadu vysvětlujících proměnných (kovariát). Metoda FCS nevychází z explicitně určeného vícerozměrného rozdělení pro všechny proměnné, ale používá samostatné podmíněné rozdělení pro každou imputovanou proměnnou. Imputace metodou FCS zahrnuje dvě fáze: fázi předběžného vyplnění daty, po níž následuje fáze imputace. Ve fázi předběžného doplňování se chybějící hodnoty pro všechny proměnné doplňují sekvenčně přes proměnné, které se berou jedna po druhé. Takto vyplněné hodnoty poskytují výchozí hodnoty pro tyto chybějící hodnoty ve fázi imputace. Ve fázi imputace se chybějící hodnoty pro každou proměnnou postupně (po několik iteracích) imputují. Podobně jako v případě metod pro datové soubory s monotónními chybějícími vzory je možné k imputování chybějících hodnot využít výše uvedené metody.

Ukázka 4 speciálních vzorů chybění neúplných dat je ilustrováno na obrázku č. 1. Vybarvené části představují pozorované údaje, zatímco prázdné části označují chybějící údaje.

***Obrázek č. 1: Příklady vzorů neúplných dat***




***Zdroj: [van Buuren, 2007, s. 225]***

Shrnutí metod mnohonásobné imputace procedurou MI je uvedeno na obrázku č. 2. Podrobnější informace ke každé z výše uvedených imputačních metod jsou uvedeny především v originální dokumentaci programového systému SAS (SAS Institute Inc., 2023, s. 6407-6409).

***Obrázek č. 2: Shrnutí možných metod imputace pomocí procedury MI podle [14]***

| Vzor chybění | Typ imputované proměnné | Typ vysvětlující proměnné (kovariáty) | Doporučená metoda imputace |
|---|---|---|---|
| Monotónní | Spojitá | Libovolná (obecná) | Monotónní lineární regrese<br>Monotónní predikovaná shoda v průměru<br>Monotónní propensitní[1] skóre |
| Monotónní | Klasifikační (ordinální) | Libovolná (obecná) | Monotónní logistická regrese |
| Monotónní | Klasifikační (nominální) | Libovolná (obecná) | Monotónní diskriminační funkce<br>Monotónní logistická regrese |
| Libovolný (obecný) | Spojitá | Spojitá | MCMC celková imputace<br>MCMC monotónní imputace |
| Libovolný (obecný) | Spojitá | Libovolná (obecná) | FCS lineární regrese<br>FCS predikovaná shoda v průměru |
| Libovolný (obecný) | Klasifikační (ordinální) | Libovolná (obecná) | FCS logistická regrese |
| Libovolný (obecný) | Klasifikační (nominální) | Libovolná (obecná) | FCS diskriminační funkce<br>FCS logistická regrese |

*[1] Odhad pravděpodobnosti přiřazené každému pozorování u proměnné obsahují chybějící hodnoty, že pozorování je chybějící.*

***Zdroj: vlastní zpracování podle (SAS Institute Inc., 2023)***

## MONOTÓNNÍ METODY PRO SOUBORY DAT S MONOTÓNNÍM CHYBĚNÍM

Pro datové soubory s monotónními vzory neúplnosti dat je možné využít k imputaci chybějících hodnot proměnných tzv. monotónní metody. Monotónní metoda vytváří mnohonásobné imputace postupným imputováním chybějících hodnot proměnných sekvenčně jednu po druhé. Například jsou-li v proceduře MI do příkazu VAR zahrnuté proměnné $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_p$ (v tomto pořadí), monotónní metoda mnohonásobné imputace postupně simuluje generování chybějících hodnot pro proměnné $\boldsymbol{Y}_2, \dots, \boldsymbol{Y}_p$ postupně pro proměnnou $\boldsymbol{Y}_2$ až po proměnnou $\boldsymbol{Y}_p$. Potom jsou chybějící hodnoty imputovány generováním přijatelných hodnot z následující sekvence:

$$\begin{aligned} \boldsymbol{\theta}_2^{(*)} &\sim p\left(\boldsymbol{\theta}_2 \middle| \boldsymbol{Y}_{1(obs)}, \boldsymbol{Y}_{2(obs)}\right) \\ \boldsymbol{Y}_2^{(*)} &\sim p\left(\boldsymbol{Y}_2 \middle| \boldsymbol{\theta}_2^{(*)}\right) \\ &\dots \\ &\dots \\ \boldsymbol{\theta}_p^{(*)} &\sim p\left(\boldsymbol{\theta}_p \middle| \boldsymbol{Y}_{1(obs)}, \dots, \boldsymbol{Y}_{p(obs)}\right) \\ \boldsymbol{Y}_p^{(*)} &\sim p\left(\boldsymbol{Y}_p \middle| \boldsymbol{\theta}_p^{(*)}\right) \end{aligned} \tag{3.1}$$

kde $\boldsymbol{Y}_{j(obs)}$ je množina pozorovaných hodnot $\boldsymbol{Y}_j$. Symbol $\boldsymbol{\theta}_2^{(*)}$ označuje množinu simulovaných parametrů pro podmíněné rozdělení proměnné $\boldsymbol{Y}_j$ za podmínky kovariát konstruovaných z proměnných $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_{\mathrm{j}-1}$ a $\boldsymbol{Y}_j^{(*)}$ je množina imputovaných hodnot $\boldsymbol{Y}_j$.

Pro počáteční proměnnou $\boldsymbol{Y}_1$ nejsou chybějící hodnoty monotónními metodami doimputovány. Chybějící údaje proměnných $\boldsymbol{Y}_2, \dots, \boldsymbol{Y}_{\mathrm{j}-1}$ nejsou doimputovány všude

tam, kde chybí hodnota proměnné $\boldsymbol{Y}_1$. Imputace pro každou následující proměnnou $\boldsymbol{Y}_\mathrm{j}$ je konstruována z proměnných předešlých $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{\mathrm{j}-1}$, které jsou použity pro statistické modelování. Statistický model pro proměnnou $\boldsymbol{Y}_\mathrm{j}$, z něhož je generována přijatelná hodnota pro imputaci proměnné $\boldsymbol{Y}_\mathrm{j}$, je vytvářen na základě předcházejícího statistického modelu. Vhodnou monotónní metodu lze samostatně zadat pro jednotlivě každou imputovanou proměnnou. Pokud není příkazy v proceduře MI zadána konkrétní metoda, systém použije metodu výchozí. Výchozí metodou pro imputaci spojité proměnné je metoda lineární regrese, pro imputaci klasifikační proměnné je výchozí metodou metoda diskriminační funkce. Procedura MI také umožňuje pro každou imputovanou proměnnou určit sadu kovariát, které jsou vytvořeny z jejích předchozích proměnných. Pokud není pro imputovanou proměnnou uvedena sada kovariát, jako kovariáty se použijí všechny předchozí proměnné v seznamu VAR.

Imputaci chybějících hodnot monotónními metodami pro spojitou proměnnou je možné realizovat pomocí regresní metody, metody prediktivní střední shody nebo metody propensitního skóre; imputace pro klasifikační proměnnou s binární nebo ordinální odezvou je realizována pomocí metody logistické regrese; metodu diskriminační funkce lze použít pro imputaci klasifikační proměnnou s binární nebo nominální odezvou.

## PLNĚ PODMÍNĚNĚ SPECIFIKOVANÉ (FCS) METODY IMPUTACE PRO SOUBORY DAT S LIBOVOLNÝM (OBECNÝM) VZOREM CHYBĚNÍ

U datového souboru s libovolným (obecným) vzorem neúplných dat systém SAS nabízí použití metody FCS k imputování chybějících hodnot pro všechny proměnné za předpokladu existence simultánního rozdělení těchto proměnných (Brand 1999; van Buuren, 2007). Každá imputace metodou FCS zahrnuje 2 fáze (SAS Institute Inc., 2023, s. 6452): fázi předběžného vyplňování a následnou fázi imputací.

Ve fázi doplňování (inicializační) se chybějící hodnoty vybraných proměnných doplňují postupně přes všechny tyto proměnné – sekvenčně jedna proměnná za druhou. Chybějící hodnoty proměnných se doplňují pomocí zadané metody nebo metody výchozí, pokud metoda doplňování hodnot nebyla specifikována příkazy procedury MI. Doplněné hodnoty slouží jako výchozí pro imputaci chybějících hodnot ve fázi imputace.

Ve fázi imputace se chybějící hodnoty pro každou proměnnou imputují pomocí zadané metody a použitých kovariát při každé iteraci. Pokud opět není metoda imputace zadána, použije se výchozí metoda pro proměnnou a zbývající proměnné se použijí jako kovariáty, pokud není množina kovariát určena příkazy procedury MI. Po provedených počátečních iteracích (počet počátečních iterací se určuje příkazem procedury MI pomocí volby NBITER=) se chybějící hodnoty v proměnných nahrazují imputovanými, a to postupně jedna proměnná po druhé.

Důležitým faktorem při imputaci procedurou MI je pořadí imputovaných proměnných. Procedura MI řadí proměnné do procesu imputace tak, jak tyto proměnné byly seřazeny příkazem VAR procedury MI. Například bylo-li pořadí $p$ proměnných v příkaze VAR procedury MI dáno jako $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_\mathrm{p}$, potom ve stejném pořadí jsou tyto proměnné použity v doplňovací i imputační fázi.

Fáze inicializačního doplňování nahrazuje chybějící hodnoty vyplněnými hodnotami u každé proměnnou. To znamená, že u $p$ proměnných $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{\mathrm{p}}$ (v tomto pořadí) se chybějící hodnoty doplní pomocí posloupnosti:

$$\begin{aligned}
\boldsymbol{\theta}_1^{(0)} &\sim p\left(\boldsymbol{\theta}_1 \middle| \boldsymbol{Y}_{1(obs)}\right) \\
\boldsymbol{Y}_{1(*)}^{(0)} &\sim p\left(\boldsymbol{Y}_1 \middle| \boldsymbol{\theta}_1^{(0)}\right) \\
\boldsymbol{Y}_1^{(0)} &= \left(\boldsymbol{Y}_{1(obs)}, \boldsymbol{Y}_{1(*)}^{(0)}\right) \\
&\ldots \\
&\ldots \\
\boldsymbol{\theta}_p^{(0)} &\sim p\left(\boldsymbol{\theta}_p \middle| \boldsymbol{Y}_1^{(0)}, \ldots, \boldsymbol{Y}_{p-1}^{(0)}, \boldsymbol{Y}_{p(obs)}\right) \\
\boldsymbol{Y}_{p(*)}^{(0)} &\sim p\left(\boldsymbol{Y}_p \middle| \boldsymbol{\theta}_p^{(0)}\right) \\
\boldsymbol{Y}_p^{(0)} &= \left(\boldsymbol{Y}_{p(obs)}, \boldsymbol{Y}_{p(*)}^{(0)}\right)
\end{aligned} \tag{3.2}$$

kde $\boldsymbol{Y}_{j(obs)}$ je množina pozorovaných hodnot $\boldsymbol{Y}_j$, $\boldsymbol{Y}_{j(*)}^{(0)}$ je množina doplňovaných hodnot proměnné $\boldsymbol{Y}_j$ a symbol $\boldsymbol{Y}_j^{(0)}$ označuje množinu pozorovaných a doplňovaných hodnot v proměnné $\boldsymbol{Y}_j$. Označení $\boldsymbol{\theta}_j^{(0)}$ ilustruje množinu parametrů podmíněného rozdělení proměnné $\boldsymbol{Y}_j$ při daných předcházejících proměnných $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{\mathrm{j}-1}$. Tímto způsobem je pro každou proměnnou $\boldsymbol{Y}_j$ s chybějícími hodnotami odhadován imputační model, na jehož základě se odhaduje model i pro následující proměnné.

Hodnoty vyplňované v inicializační fázi jsou ve fázi imputace nahrazovány hodnotami imputovanými v každé iteraci postupně pro každou proměnnou. To znamená, že u $p$ sledovaných proměnných $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{\mathrm{p}}$ se chybějící hodnoty nahrazují sekvencí hodnot v iteraci $t+1$, tj.

$$\begin{aligned}
\boldsymbol{\theta}_1^{(t+1)} &\sim p\left(\boldsymbol{\theta}_1 \middle| \boldsymbol{Y}_{1(obs)}, \boldsymbol{Y}_2^{(t)}, \ldots, \boldsymbol{Y}_p^{(t)}\right) \\
\boldsymbol{Y}_{1(*)}^{(t+1)} &\sim p\left(\boldsymbol{Y}_1 \middle| \boldsymbol{\theta}_1^{(t+1)}\right) \\
\boldsymbol{Y}_1^{(t+1)} &= \left(\boldsymbol{Y}_{1(obs)}, \boldsymbol{Y}_{1(*)}^{(t+1)}\right) \\
&\ldots \\
&\ldots \\
\boldsymbol{\theta}_p^{(t+1)} &\sim p\left(\boldsymbol{\theta}_p \middle| \boldsymbol{Y}_1^{(t+1)}, \ldots, \boldsymbol{Y}_{p-1}^{(t+1)}, \boldsymbol{Y}_{p(obs)}\right) \\
\boldsymbol{Y}_{p(*)}^{(t+1)} &\sim p\left(\boldsymbol{Y}_p \middle| \boldsymbol{\theta}_p^{(t+1)}\right) \\
\boldsymbol{Y}_p^{(t+1)} &= \left(\boldsymbol{Y}_{p(obs)}, \boldsymbol{Y}_{p(*)}^{(t+1)}\right)
\end{aligned} \tag{3.3}$$

kde $\boldsymbol{Y}_{j(obs)}$ je množina pozorovaných hodnot $\boldsymbol{Y}_j$, $\boldsymbol{Y}_{j(*)}^{(t+1)}$ je množina imputovaných hodnot proměnné $\boldsymbol{Y}_j$ v iteraci $t+1$. Symbol $\boldsymbol{Y}_{j(*)}^{(t)}$označuje množinu doplněných hodnot proměnné $\boldsymbol{Y}_j$ v iteraci $t+0$, příp. množinu imputovaných hodnot $\boldsymbol{Y}_j$ v iteraci $t$ ( $t > 0$), $\boldsymbol{Y}_j^{(t+1)}$označuje množinu jak pozorovaných, tak i imputovaných hodnot proměnné $\boldsymbol{Y}_j$ za

iterace $t+1$. Symbol $\boldsymbol{\theta}_j^{(t+1)}$ označuje množinu simulovaných parametrů podmíněného rozdělení proměnné $\boldsymbol{Y}_j$ při daných proměnných $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{\mathrm{j}-1}, \boldsymbol{Y}_{\mathrm{j}+1}, \ldots, \boldsymbol{Y}_{\mathrm{p}}$.

Při každé iteraci je pro každou proměnnou s chybějícími hodnotami odhadován specifikovaný imputační model na základě pozorovaných hodnot pro tuto proměnnou, které mohou zahrnovat pozorování s imputovanými hodnotami pro jiné proměnné. Na základě odhadnutého imputačního modelu je odhadován nový model a poté použit k imputaci chybějících hodnot pro následující imputovanou proměnnou. Kroky se opakují dostatečně dlouho na to, aby výsledky spolehlivě simulovaly přibližně nezávislý výběrový soubor chybějících hodnot pro imputovanou datovou sadu. FCS metody imputace používané ve fázi vyplňování a fázi imputace jsou podobné odpovídajícím monotónním metodám pro monotónní chybějící data. K imputaci chybějících hodnot FCS metodami systém SAS využívá pro spojitou proměnnou metodu regrese nebo metodu prediktivní střední shody, metodu logistické regrese pro klasifikační proměnnou s binární nebo ordinální odezvou a metodu diskriminační funkce pro klasifikační proměnnou s binární nebo nominální hodnotou.

## METODA MCMC PRO MNOHOROZMĚRNÉ NORMÁLNÍ DATA S OBECNÝM VZOREM CHYBĚNÍ

Metoda Markovova řetězce Monte Carlo (MCMC) vznikla ve fyzice jako nástroj pro zkoumání rovnovážných rozdělení interagujících molekul. Ve statistických aplikacích se používá ke generování pseudonáhodných výběrů z mnohorozměrných anebo jinak obtížně řešitelných rozdělení pravděpodobnosti prostřednictvím Markovových řetězců. Markovův řetězec je posloupnost náhodných veličin, v níž rozdělení každého prvku závisí pouze na hodnotě předchozího prvku.

Při simulaci MCMC se konstruuje dostatečně dlouhý Markovův řetězec na to, aby se rozdělení simulovaných prvků ustálilo na stacionárním rozdělení. Opakovaným simulováním kroků Markovova řetězce se generují náhodné výběry ze zájmového rozdělení. Podrobnější informace o metodě MCMC jsou uvedeny například v práci (Schafer, 1997).

Metoda MCMC se používá jako metoda pro zkoumání aposteriorních rozdělení v bayesovské inferenci. To znamená, že prostřednictvím MCMC lze simulovat celé společné aposteriorní rozdělení neznámých veličin a získat na základě simulace odhady aposteriorních parametrů, které jsou předmětem zájmu. V bayesovské inferenci jsou informace o neznámých parametrech vyjádřeny ve formě aposteriorního rozdělení pravděpodobnosti. Toto aposteriorní rozdělení se vypočítá pomocí Bayesovy věty,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \tag{3.4}$$

V mnoha problémech s neúplnými daty není aposterioriorní rozdělení $p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}, \boldsymbol{X})$ z pozorovaných dat jednoduše řešitelné a nelze jej snadno simulovat. Pokud je však vektor pozorovaných hodnot $\boldsymbol{Y}_{obs}$ rozšířen o odhadnutou nebo simulovanou hodnotu chybějících dat $\boldsymbol{Y}_{mis}$, je mnohem snazší toto aposteriorní rozdělení s úplnými daty simulovat, tj. aposteriorní rozdělení $p(\boldsymbol{\theta}|\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs}, \boldsymbol{X})$. Za předpokladu (Takahashi,

2017), že data pocházejí z vícerozměrného normálního rozdělení, lze toto rozšíření[5] dat aplikovat na bayesovskou inferenci s chybějícími daty opakováním následujících kroků:

**Imputační krok** (***I*-krok**) simuluje chybějící hodnoty vzhledem k odhadnutému střednímu vektoru a kovarianční matici pro každé pozorování nezávisle. Pokud se proměnné s chybějícími hodnotami pro $i$-té pozorování, $i = 1,2, \dots, \mathrm{n}$, označí jako $\boldsymbol{Y_{i(mis)}}$ a proměnné s pozorovanými hodnotami jako $\boldsymbol{Y_{i(obs)}}$, pak *I*-krok náhodně vybírá hodnoty pro $\boldsymbol{Y_{i(mis)}}$ z podmíněného rozdělení pro $\boldsymbol{Y_{i(mis)}}$ vzhledem k $\boldsymbol{Y_{i(obs)}}$.

**Aposteriorní krok** (***P*-krok**) simuluje za daného rozšíření neúplných dat na data úplná aposteriorní vektor středních hodnot a kovarianční matici. Tyto nové odhady se pak použijí v dalším *I*-kroku. Pokud neexistují předběžné informace o odhadovaných parametrech, jako apriorní rozdělení se používá neinformativní apriorní rozdělení. Lze však využít i jiné informativní apriorní rozdělení. Například blíží-li se kovarianční matice matici singulární, může apriorní informace o kovarianční matici pomoci stabilizaci odvozování vektoru středních hodnot.

To znamená, že s aktuálním odhadem parametru $\boldsymbol{\theta}^{(r)}$ v $r$-té iteraci imputační *I*-krok generuje $\boldsymbol{Y}_{miss}^{(r+1)}$ z podmíněného rozdělení pravděpodobnosti $p\left(\boldsymbol{Y}_{miss}\middle|\boldsymbol{Y_{obs}}, \boldsymbol{X}, \boldsymbol{\theta}^{(r)}\right)$ a aposteriorní *P*-krok náhodně vybírá $\boldsymbol{\theta}^{(r+1)}$ z podmíněného rozdělení pravděpodobnosti $p\left(\boldsymbol{\theta}\middle|\boldsymbol{Y_{obs}}, \boldsymbol{Y}_{miss}{}^{(r+1)}, \boldsymbol{X}\right)$. Oba kroky se iterují dostatečně dlouho, aby výsledky spolehlivě simulovaly přibližně nezávislé náhodné výběry chybějících hodnot pro soubor mnohonásobně imputovaných dat (SAS Institute Inc., 2023, s. 6455). Tím se vytvoří Markovův řetězec $\left(\boldsymbol{Y}_{miss}^{(1)}, \boldsymbol{\theta}^{(1)}\right), \left(\boldsymbol{Y}_{miss}^{(2)}, \boldsymbol{\theta}^{(2)}\right), \dots$, jehož rozdělení konverguje k $p(\boldsymbol{Y}_{miss}, \boldsymbol{\theta}|\boldsymbol{Y_{obs}}, \boldsymbol{X})$. Za předpokladu, že iterace konvergují ke stacionárnímu rozdělení, je cílem simulovat tzv. přibližně nezávislé náhodné výběry chybějících hodnot z tohoto rozdělení.

Metoda MCMC je užitečná zejména v případech, kdy se soubor neúplných dat blíží monotónnímu chybějícímu vzoru. V tomto případě metoda potřebuje do datového souboru imputovat pouze několik chybějících hodnot, aby měl datový soubor monotónní vzor chybění v imputovaném souboru dat. V porovnání s úplnou imputací dat, která imputuje všechny chybějící hodnoty, metoda MCMC pro dosažení monotónního vzoru neúplných dat imputuje v každé iteraci méně chybějících hodnot a dosahuje přibližné stacionarity v datovém souboru velmi rychle (Schafer, 1997, s. 227).

## PRAKTICKÁ UKÁZKA STATISTICKÉ ANALÝZY NEÚPLNÝCH DAT VYUŽITÍM MI

V rámci praktické ukázky statistické analýzy neúplných dat budou na praktickém příkladu realizovány:

- ukázka používaných technik zpracování chybějících údajů se zaměřením na metodu MI,
- problémy, které mohou nastat při použití těchto technik,

*[5] V anglicky psané literatuře se toto rozšíření dat o simulované hodnoty nazývané termínem data augmentation. Data augmentation vylepšuje odhady parametrů opakovaným nahrazováním (simulací) chybějících údajů - poznámka autora.*

- implementace procedury SAS pro MI za předpokladu:
  - mnohorozměrného normálního rozdělení a
  - plně podmíněně specifikovaného rozdělení (FCS) pro vybrané proměnné
- diagnostika imputace.

**Cílem statistické analýzy** s chybějícími údaji bude:

- minimalizace zkreslení odhadů parametrů,
- maximalizace využití dostupných informací,
- získání nejlepších odhadů nejistoty.

Pro účely praktické ukázky statistické analýzy neúplných dat byla vybrána data longitudinálního šetření High School and Beyond realizovaného v rámci Národního programu longitudinálních studií USA[6]. Tento národní výzkumný program byl založen za účelem studia vzdělávacího, profesního a osobního vývoje mladých lidí. Začíná na základní nebo střední škole a sleduje je v průběhu času, kdy přebírají role a povinnosti dospělých. Šetření zahrnovalo dvě středoškolské kohorty – kohortu maturantů (maturitní ročník 1980) a kohortu žáků druhého ročníku (ročník 1980). Údaje pro studii poskytli studenti, školní administrátoři, učitelé, rodiče a administrativní registry.

**Použitý datový soubor** obsahuje 200 pozorování z celého zkoumaného vzorku středoškolských studentů s demografickými informacemi o studentech, jako je jejich pohlaví (FEMALE), socioekonomický status (SES), typ navštěvované školy (proměnná SCHTYP) a etnický původ (RACE). Obsahuje také řadu výsledků standardizovaných testů, včetně testů čtení (READ), psaní (WRITE), matematiky (MATH), vědy (SCIENCE) a společenských věd (SOCST). Datový soubor obsahuje jak kategorické proměnné, a to proměnné FEMALE, SES, SCTYP a RACE, tak i spojité proměnné. Vybrané popisné charakteristiky úplného datového souboru ilustruje obrázek č. 3.

***Obrázek č. 3: Vybrané popisné statistiky ukázkového souboru úplných dat***

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum | N Miss |
|---|---|---|---|---|---|---|---|
| id | idNumber | 200 | 100.5000000 | 57.8791845 | 1.0000000 | 200.0000000 | 0 |
| female | gender | 200 | 0.5450000 | 0.4992205 | 0 | 1.0000000 | 0 |
| race | race | 200 | 3.4300000 | 1.0394722 | 1.0000000 | 4.0000000 | 0 |
| ses | socioeconomic status | 200 | 2.0550000 | 0.7242914 | 1.0000000 | 3.0000000 | 0 |
| schtyp | type of school | 200 | 1.1600000 | 0.3675260 | 1.0000000 | 2.0000000 | 0 |
| prog | type of program | 200 | 2.0250000 | 0.6904772 | 1.0000000 | 3.0000000 | 0 |
| read | reading score | 200 | 52.2300000 | 10.2529368 | 28.0000000 | 76.0000000 | 0 |
| write | writing score | 200 | 52.7750000 | 9.4785860 | 31.0000000 | 67.0000000 | 0 |
| math | math score | 200 | 52.6450000 | 9.3684478 | 33.0000000 | 75.0000000 | 0 |
| science | science score | 200 | 51.8500000 | 9.9008908 | 26.0000000 | 74.0000000 | 0 |
| socst | social studies score | 200 | 52.4050000 | 10.7357935 | 26.0000000 | 71.0000000 | 0 |

***Zdroj: vlastní zpracování***

Soubor dat, který obsahuje chybějící údaje ve sledovaných proměnných, vychází ze souboru úplných dat. Ačkoli soubor neúplných dat obsahuje také 200 pozorování, šest z proměnných má méně než 200 pozorování. Chybějící informace se pohybují v rozmezí 4,5%. (READ) až 9% (FEMALE a PROG) případů v závislosti na proměnné. Situaci v datech ilustruje následující obrázek č. 4.

[6] *High School and Beyond Longitudinal Study dostupné na http://nces.ed.gov/surveys/hsb/.*

***Obrázek č. 4: Vybrané popisné statistiky ukázkového souboru neúplných dat***

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum | N Miss |
|---|---|---|---|---|---|---|---|
| ID | idNumber | 200 | 100.5000000 | 57.8791845 | 1.0000000 | 200.0000000 | 0 |
| FEMALE | gender | 182 | 0.5549451 | 0.4983428 | 0 | 1.0000000 | 18 |
| RACE | race | 200 | 3.4300000 | 1.0394722 | 1.0000000 | 4.0000000 | 0 |
| SES | socioeconomic status | 200 | 2.0550000 | 0.7242914 | 1.0000000 | 3.0000000 | 0 |
| SCHTYP | type of school | 200 | 1.1600000 | 0.3675260 | 1.0000000 | 2.0000000 | 0 |
| PROG | type of program | 182 | 2.0274725 | 0.6927511 | 1.0000000 | 3.0000000 | 18 |
| READ | reading score | 191 | 52.2879581 | 10.2107174 | 28.0000000 | 76.0000000 | 9 |
| WRITE | writing score | 183 | 52.9508197 | 9.2577729 | 31.0000000 | 67.0000000 | 17 |
| MATH | math score | 185 | 52.8972973 | 9.3608367 | 33.0000000 | 75.0000000 | 15 |
| SCIENCE | science score | 184 | 51.3097826 | 9.8178332 | 26.0000000 | 74.0000000 | 16 |
| SOCST | social studies score | 200 | 52.4050000 | 10.7357935 | 26.0000000 | 71.0000000 | 0 |

***Zdroj: vlastní zpracování***

Neúplnost dat na první pohled nevypadá jako příliš rozsáhlá, takže bychom se mohli přiklonit k pokusu analyzovat pozorované neúplné údaje v nedoplněném stavu. Taková strategie analýzy dat je obvykle označována jako analýza kompletních případů a je implicitním postupem pro většinu statistických softvérů. Například při analýze regresního modelu, kde závislá vysvětlovaná proměnná READ je regresována na vysvětlujících proměnných WRITE, MATH, FEMALE a PROG. Výchozí analytickou strategií v systému SAS (a podobně i v jiných systémech) je analýza kompletních případů realizovaná procedurou GLM. Při pohledu na výstupy procedury GLM je patrné, že v analýze bylo použito pouze 130 případů; tedy více než třetina případů ze souboru neúplných dat (tj. 70 vyloučených z 200 celkem) byla z analýzy vyloučena z důvodu chybějících údajů. Výstupy procedury GLM jsou znázorněny na obrázku č. 5.

***Obrázek č. 5: Odhad parametrů modelu z neúplných dat procedurou GLM***

| | |
|---|---|
| Number of Observations Read | 200 |
| Number of Observations Used | 130 |

Dependent Variable: READ reading score

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 5895.48143 | 1179.09629 | 23.69 | <.0001 |
| Error | 124 | 6172.12627 | 49.77521 | | |
| Corrected Total | 129 | 12067.60769 | | | |

| Parameter | | Estimate | | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | | 13.02649943 | B | 4.12354544 | 3.16 | 0.0020 |
| WRITE | | 0.44108340 | | 0.09264775 | 4.76 | <.0001 |
| FEMALE | female | -2.70633778 | B | 1.36519467 | -1.98 | 0.0496 |
| FEMALE | male | 0.00000000 | B | . | . | . |
| MATH | | 0.32105246 | | 0.09514356 | 3.37 | 0.0010 |
| PROG | academic | 1.81115548 | B | 1.65485900 | 1.09 | 0.2759 |
| PROG | general | 0.51774275 | B | 1.88083319 | 0.28 | 0.7836 |
| PROG | vocation | 0.00000000 | B | . | . | . |

***Zdroj: vlastní zpracování***

## NAHRAZENÍ DAT MNOHONÁSOBNOU IMPUTACÍ

Mnohonásobná imputace je v podstatě iterační forma stochastické imputace. Namísto doplnění jediné hodnoty se však používá rozdělení pozorovaných dat k odhadu více hodnot. Tyto hodnoty se pak použijí v datové analýze, například

v regresním modelu, a výsledky se zkombinují. Každá imputovaná hodnota obsahuje náhodnou složku, jejíž velikost odráží míru, do jaké ostatní proměnné v imputačním modelu nemohou předpovědět její skutečné hodnoty (White et al, 2011).

**Příprava na použití MI**. Nejprve je potřebné zjistit počet a podíl chybějících hodnot v analyzovaných údajích. V programovém systému SAS je možné zjistit počet chybějících hodnot pro všechny proměnné pomocí procedury MEANS s parametrem NMISS. Výstup procedury MEANS je zobrazen na obrázku č. 6.

***Obrázek č. 6: Počet chybějících hodnot vybraných proměnných souboru dat***

| Variable | Label | N Miss |
|---|---|---|
| FEMALE | gender | 18 |
| WRITE | writing score | 17 |
| READ | reading score | 9 |
| MATH | math score | 15 |
| PROG | type of program | 18 |

***Zdroj: vlastní zpracování***

Podíl chybějících informací na zájmových proměnných lze posoudit připojením příznaků chybějících údajů, resp. indikátorových proměnných pro chybějící údaje. Četnost a podíl chybových proměnných vybraných pro datovou analýzu lze určit pomocí procedury FREQ. Výstupní četnosti a podíly chybějících hodnot jsou uvedeny na obrázku č. 7. Proměnné s nejvyšším podílem chybějících informací jsou PROG a FEMALE s 9% Obecně platí, že proměnné s vysokým podílem chybějících informací mají největší vliv na konvergenci zadaného imputačního modelu.

***Obrázek č. 7: Podíly chybějících hodnot vybraných proměnných souboru dat***

| female_flag | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 182 | 91.00 | 182 | 91.00 |
| 1 | 18 | 9.00 | 200 | 100.00 |

| write_flag | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 183 | 91.50 | 183 | 91.50 |
| 1 | 17 | 8.50 | 200 | 100.00 |

| read_flag | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 191 | 95.50 | 191 | 95.50 |
| 1 | 9 | 4.50 | 200 | 100.00 |

| math_flag | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 185 | 92.50 | 185 | 92.50 |
| 1 | 15 | 7.50 | 200 | 100.00 |

| prog_flag | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 182 | 91.00 | 182 | 91.00 |
| 1 | 18 | 9.00 | 200 | 100.00 |

***Zdroj: vlastní zpracování***

Následuje zjištění vzoru chybějících dat neúplných proměnných využitím funkcionality systému SAS procedurou MI. Vzor chybění je ilustrován na obrázku č. 8. Každá skupina (sloupec označený "Group") představuje soubor pozorování v souboru dat, která mají stejný vzor chybějících informací. Například skupina 1 představuje 130 pozorování v datech, která mají úplné informace o všech 5 proměnných, které jsou předmětem analytického zájmu. Procedura rovněž poskytuje průměry pro každou proměnnou pro tuto skupinu. Soubor neúplných dat pro uvedené proměnné obsahuje celkem 12 vzorů.

***Obrázek č. 8: Vzor chybění vybraných proměnných neúplného souboru dat***

Missing Data Patterns

| | | | | | | | | | Group Means | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Group | SOCST | WRITE | READ | FEMALE | MATH | PROG | Freq | Percent | SOCST | WRITE | READ | FEMALE | MATH | PROG |
| 1 | X | X | X | X | X | X | 130 | 65.00 | 53.138462 | 53.200000 | 52.838462 | 0.600000 | 52.600000 | 2.046154 |
| 2 | X | X | X | X | X | . | 15 | 7.50 | 54.666667 | 56.200000 | 52.733333 | 0.466667 | 55.400000 | . |
| 3 | X | X | X | X | . | X | 11 | 5.50 | 46.909091 | 53.090909 | 51.363636 | 0.272727 | . | 2.000000 |
| 4 | X | X | X | X | . | . | 1 | 0.50 | 51.000000 | 59.000000 | 52.000000 | 1.000000 | . | . |
| 5 | X | X | X | . | X | X | 15 | 7.50 | 50.333333 | 49.933333 | 48.600000 | . | 49.866667 | 1.866667 |
| 6 | X | X | X | . | X | . | 1 | 0.50 | 31.000000 | 44.000000 | 44.000000 | . | 40.000000 | . |
| 7 | X | X | X | . | . | X | 1 | 0.50 | 31.000000 | 33.000000 | 44.000000 | . | . | 1.000000 |
| 8 | X | X | . | X | X | X | 9 | 4.50 | 56.000000 | 51.333333 | . | 0.444444 | 53.444444 | 2.222222 |
| 9 | X | . | X | X | X | X | 13 | 6.50 | 52.000000 | . | 54.230769 | 0.461538 | 57.076923 | 1.923077 |
| 10 | X | . | X | X | X | . | 1 | 0.50 | 56.000000 | . | 55.000000 | 1.000000 | 66.000000 | . |
| 11 | X | . | X | X | . | X | 2 | 1.00 | 41.000000 | . | 47.000000 | 0.500000 | . | 2.000000 |
| 12 | X | . | X | . | X | X | 1 | 0.50 | 51.000000 | . | 39.000000 | . | 40.000000 | 3.000000 |

***Zdroj: vlastní zpracování***

Je také vhodné stanovit potenciální pomocné proměnné. Pomocné proměnné jsou proměnné v neúplném souboru dat, a to buď korelované s proměnnou s chybějícími hodnotami (doporučení je $r > 0{,}4$), nebo se předpokládá, že pomocná proměnná souvisí s chybějícími hodnotami. Jedná se o faktory, které nemají zvláštní význam pro analytický model. Přidávají se však do imputačního modelu, aby zvýšily sílu a/nebo pomohly učinit předpoklad MAR věrohodnějším. Tyto pomocné proměnné zlepšují kvalitu imputovaných hodnot generovaných vícenásobnou imputací (Enders, 2010).

***Obrázek č. 9: Matice korelačních koeficientů k zjištění pomocných proměnných***

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

| | SOCST | WRITE | READ | FEMALE | MATH | SCIENCE | progcat1 | progcat2 |
|---|---|---|---|---|---|---|---|---|
| SOCST | 1.00000 | 0.59750 | 0.61604 | 0.08894 | 0.54509 | 0.45125 | -0.07680 | 0.40956 |
| social studies score | | <.0001 | <.0001 | 0.2325 | <.0001 | <.0001 | 0.3028 | <.0001 |
| | 200 | 183 | 191 | 182 | 185 | 184 | 182 | 182 |
| WRITE | 0.59750 | 1.00000 | 0.58719 | 0.25077 | 0.61825 | 0.54977 | -0.06036 | 0.34387 |
| writing score | <.0001 | | <.0001 | 0.0011 | <.0001 | <.0001 | 0.4398 | <.0001 |
| | 183 | 183 | 174 | 166 | 170 | 168 | 166 | 166 |
| READ | 0.61604 | 0.58719 | 1.00000 | -0.01740 | 0.65890 | 0.63288 | -0.10575 | 0.39023 |
| reading score | <.0001 | <.0001 | | 0.8202 | <.0001 | <.0001 | 0.1661 | <.0001 |
| | 191 | 174 | 191 | 173 | 176 | 176 | 173 | 173 |
| FEMALE | 0.08894 | 0.25077 | -0.01740 | 1.00000 | -0.02408 | -0.09176 | -0.03169 | 0.05004 |
| female | 0.2325 | 0.0011 | 0.8202 | | 0.7567 | 0.2397 | 0.6861 | 0.5233 |
| | 182 | 166 | 173 | 182 | 168 | 166 | 165 | 165 |
| MATH | 0.54509 | 0.61825 | 0.65890 | -0.02408 | 1.00000 | 0.62964 | -0.16511 | 0.44566 |
| math score | <.0001 | <.0001 | <.0001 | 0.7567 | | <.0001 | 0.0325 | <.0001 |
| | 185 | 170 | 176 | 168 | 185 | 169 | 168 | 168 |
| SCIENCE | 0.45125 | 0.54977 | 0.63288 | -0.09176 | 0.62964 | 1.00000 | 0.05672 | 0.20379 |
| science score | <.0001 | <.0001 | <.0001 | 0.2397 | <.0001 | | 0.4666 | 0.0083 |
| | 184 | 168 | 176 | 166 | 169 | 184 | 167 | 167 |
| progcat1 | -0.07680 | -0.06036 | -0.10575 | -0.03169 | -0.16511 | 0.05672 | 1.00000 | -0.56349 |
| academic | 0.3028 | 0.4398 | 0.1661 | 0.6861 | 0.0325 | 0.4666 | | <.0001 |
| | 182 | 166 | 173 | 165 | 168 | 167 | 182 | 182 |
| progcat2 | 0.40956 | 0.34387 | 0.39023 | 0.05004 | 0.44566 | 0.20379 | -0.56349 | 1.00000 |
| general | <.0001 | <.0001 | <.0001 | 0.5233 | <.0001 | 0.0083 | <.0001 | |
| | 182 | 166 | 173 | 165 | 168 | 167 | 182 | 182 |

***Zdroj: vlastní zpracování***

Jedním ze způsobů identifikace těchto pomocných proměnných je zkoumání asociací proměnných důležitých pro datovou analýzu (WRITE, READ, FEMALE a MATH) s dalšími proměnnými v souboru dat. Za tímto účelem lze využít proceduru CORR, jejíž výstupem je matice korelačních koeficientů (viz obrázek č. 9). Podle hodnot korelačních koeficientů r>0,4, nejvhodnějšími pomocnými proměnnými mohou být proměnné s hodnotami výsledků testů SCIENCE a SOCST.

Dále je užitečné v analýze neúplných dat sledovat, zda potenciální pomocné proměnné mohou být považovány za prediktory chybění. K tomu se využije procedura TTEST, kterou se testuje, zda se průměrné výsledky SCIENCE nebo SOCST významně liší v pozorováních s úplnou nebo chybějící informací. Jediný významný rozdíl průměrů byl zjištěn v t-testech skóre SOCST a výsledků matematických testů. Průměrné skóre SOCST je významně nižší u respondentů, kteří nemají vyplněné výsledky testu v matematice (MATH).

Výsledek t-testu naznačuje, že potenciálním prediktorem chybění je proměnná SOCST a jeho zahrnutím do imputačního modelu může pomoci splnění předpokladu MAR (Enders, 2010). Obrázek č. 10 zobrazuje výsledek t-testu pro potenciální prediktor.

***Obrázek č. 10: Výsledky t-testu s potenciálním prediktorem chybění***

Variable: SOCST (social studies score)

| math_flag | Method | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| 0 | | 185 | 52.9784 | 10.4600 | 0.7690 | 26.0000 | 71.0000 |
| 1 | | 15 | 45.3333 | 11.9323 | 3.0809 | 26.0000 | 66.0000 |
| Diff (1-2) | Pooled | | 7.6450 | 10.5709 | 2.8379 | | |
| Diff (1-2) | Satterthwaite | | 7.6450 | | 3.1755 | | |

| math_flag | Method | Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|---|---|
| 0 | | 52.9784 | 51.4611 | 54.4956 | 10.4600 | 9.4918 | 11.6501 |
| 1 | | 45.3333 | 38.7254 | 51.9412 | 11.9323 | 8.7360 | 18.8185 |
| Diff (1-2) | Pooled | 7.6450 | 2.0487 | 13.2414 | 10.5709 | 9.6243 | 11.7255 |
| Diff (1-2) | Satterthwaite | 7.6450 | 0.9063 | 14.3838 | | | |

| Method | Variances | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Pooled | Equal | 198 | 2.69 | 0.0077 |
| Satterthwaite | Unequal | 15.794 | 2.41 | 0.0287 |

| Equality of Variances | | | | |
|---|---|---|---|---|
| Method | Num DF | Den DF | F Value | Pr > F |
| Folded F | 14 | 184 | 1.30 | 0.4200 |

***Zdroj: vlastní zpracování***

**Doplnění chybějících hodnot pomocí MI metodou MCMC**. Při rozhodování o imputaci hodnot pro jednu anebo více proměnných musí být jedno z prvních rozhodování stanovení typu rozdělení, podle kterého budou vybrané proměnné doplněny imputovanými hodnotami. Pravděpodobně nejčastějším přístupem k MI je předpoklad, že všechny proměnné pro imputační model podléhají sdruženému mnohorozměrnému normálnímu rozdělení. Výchozím algoritmem pro generování imputovaných hodnot je algoritmus datové augmentace, který náleží do skupiny MCMC procedur a jehož základní vlastnosti byly uvedeny v předešlých částech tohoto příspěvku. Předpoklad mnohorozměrného normálního rozdělení je v praxi postačující, pokud je rozsah imputovaného vzorku dostatečný (Yucel, 2008).

**Imputační fáze** je fáze, ve které analytik vytvoří imputační model a počet imputovaných datových souborů, které se mají vytvořit. K realizaci imputační fáze se

v prostředí systému SAS použije procedura MI. V rámci procedury MI je možné pomocí volby NIMPUTE zadat počet imputací, které se mají provést. Imputované soubory dat budou vytvořeny pomocí volby OUT= a uloženy do jediného datového souboru s názvem MI_MVN. Procedura automaticky vytvoří indikátorovou proměnnou nazvanou _IMPUTATION_, která očísluje každý nový imputovaný soubor dat. Po příkazu VAR se zadají všechny proměnné potřebné pro imputační model, včetně všech proměnných v předpokládaném analytickém modelu i všechny významné pomocné proměnné. Volba SEED není vyžadována, ale protože procedura MI generuje imputované hodnoty jako náhodný proces, nastavení parametru SEED umožní získat pokaždé stejný imputovaný soubor dat. Nastavení procedury MI ilustruje následující sekvence příkazů:

```
proc mi data= ats.hsb_flag nimpute=10 out=mi_mvn seed=54321;
     var science write read female math progcat1 progcat2;
run;
```

**Analytickou fází** procesu imputace se využitím procedury GLM odhadne lineární regresní model (analytický model) pro každý imputovaný soubor samostatně pomocí příkazu BY a proměnné _IMPUTATION_ vytvořené v předešlé fázi. Výstupem analytické procedury GLM je soubor odhadů parametrů pro každý z vygenerovaného imputačního souboru. Odhady parametrů lineárního regresního modelu se příkazem OUTPUT= procedury GLM uloží do datového souboru A_MVN Soubor odhadů se následně využije v závěrečné části procesu imputace. Ukázka příkazů procedury GLM pro odhady parametrů analytického modelu z imputovaných souborů je uvedena níže:

```
proc glm data = mi_mvn;
     model read = write female math progcat1 progcat2;
     by _imputation_;
     ods output ParameterEstimates=a_mvn;
run;

quit;
```

***Obrázek č. 11: Odhady parametrů modelu z výstupu procedury MIANALYZE***

The MIANALYZE Procedure

Model Information

| | |
|---|---|
| PARMS Data Set | WORK.A_MVN |
| Number of Imputations | 10 |

Variance Information (10 Imputations)

| Parameter | Variance Between | Variance Within | Variance Total | DF | Relative Increase in Variance | Fraction Missing Information | Relative Efficiency |
|---|---|---|---|---|---|---|---|
| intercept | 1.736805 | 11.199790 | 13.110276 | 423.82 | 0.170582 | 0.149727 | 0.985248 |
| write | 0.000761 | 0.005756 | 0.006594 | 557.93 | 0.145486 | 0.130121 | 0.987155 |
| female | 0.479748 | 1.208264 | 1.735987 | 97.392 | 0.436761 | 0.317856 | 0.969194 |
| math | 0.000891 | 0.005789 | 0.006769 | 429.13 | 0.169343 | 0.148777 | 0.985340 |
| progcat1 | 0.374931 | 1.916474 | 2.328899 | 286.98 | 0.215199 | 0.182765 | 0.982051 |
| progcat2 | 0.654644 | 2.202830 | 2.922939 | 148.28 | 0.326901 | 0.256328 | 0.975008 |

Parameter Estimates (10 Imputations)

| Parameter | Estimate | Std Error | 95% Confidence Limits | | DF | Minimum | Maximum | Theta0 | t for H0: Parameter=Theta0 | Pr > \|t\| |
|---|---|---|---|---|---|---|---|---|---|---|
| intercept | 9.575232 | 3.620811 | 2.45825 | 16.69222 | 423.82 | 7.661147 | 11.226972 | 0 | 2.64 | 0.0085 |
| write | 0.387371 | 0.081203 | 0.22787 | 0.54687 | 557.93 | 0.337664 | 0.418381 | 0 | 4.77 | <.0001 |
| female | -2.320250 | 1.317569 | -4.93513 | 0.29463 | 97.392 | -3.372193 | -1.208880 | 0 | -1.76 | 0.0814 |
| math | 0.413743 | 0.082276 | 0.25203 | 0.57546 | 429.13 | 0.364244 | 0.457837 | 0 | 5.03 | <.0001 |
| progcat1 | 2.596406 | 1.526073 | -0.40731 | 5.60012 | 286.98 | 2.014590 | 3.855307 | 0 | 1.70 | 0.0900 |
| progcat2 | 0.809808 | 1.709660 | -2.56864 | 4.18825 | 148.28 | -0.190604 | 2.416967 | 0 | 0.47 | 0.6364 |

***Zdroj: vlastní zpracování***

**Slučovací fáze** imputačního procesu slouží ke statistické inferenci odhadů chybějících hodnot. K tomuto účelu se v prostředí systému SAS využívá procedura MIANALYZE, pro kterou je vstupem soubor odhadů A_MVN včetně kovariančních matic k odhadům standardních chyb odhadů. Tato fáze spojuje odhady parametrů do jediné sady statistik, které vhodně odrážejí nejistotu spojenou s imputovanými hodnotami. Výsledné odhady parametrů jsou jednoduše jen aritmetickým průměrem jednotlivých koeficientů odhadnutých pro každý z imputačních regresních modelů, čímž se zvyšuje efektivita a snižuje se variabilita výběru. Odhady standardních chyb odhadů jsou o něco složitější a jsou dány postupem například v (Pavelka, 2024a nebo Rubin, 1987). Příkladem příkazů v rámci procedury MIANALYZE je následující sekvence:

```
proc mianalyze parms=a_mvn;
     modeleffects intercept write female math progcat1 progcat2;
run;
```

Výstupy analýz doplněných hodnot z procedury MIANALYZE ilustruje obrázek č. 11.

Porovnání analýzy regresního modelu úplných dat, z kompletních případů a z dat po imputaci metodou MI za předpokladu mnohorozměrného normálního rozdělení je zobrazeno na obrázku č. 12. Z obrázku č. 12 je patrné, že odhady parametrů regresního modelu z dat po imputaci se blíží odhadům z úplných dat. Nejvíce odlišné jsou odhady modelu z kompletních případů.

***Obrázek č. 12: Porovnání odhadů analytického modelu ze sledovaných dat***

| | kompletní data | | | úplné případy | | | po 10 imputacích MNorm | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | Estimate | Std Error | Pr > \|t\| | Estimate | Std Error | Pr > \|t\| | Estimate | Std Error | Pr > \|t\| |
| intercept | **9,62** | 3,41 | 0,01 | **13,03** | 4,12 | 0,00 | **9,58** | 3,62 | 0,01 |
| write | **0,37** | 0,07 | <.0001 | **0,44** | 0,09 | <.0001 | **0,39** | 0,08 | <.0001 |
| female | **- 2,70** | 1,10 | 0,01 | **- 2,71** | 1,37 | 0,05 | **- 2,32** | 1,32 | 0,08 |
| math | **0,44** | 0,07 | <.0001 | **0,32** | 0,10 | 0,00 | **0,41** | 0,08 | <.0001 |
| progcat1 | **1,88** | 1,42 | 0,19 | **1,81** | 1,65 | 0,28 | **2,60** | 1,53 | 0,09 |
| progcat2 | **0,23** | 1,51 | 0,88 | **0,52** | 1,88 | 0,78 | **0,81** | 1,71 | 0,64 |

***Zdroj: vlastní zpracování***

**Doplnění chybějících hodnot pomocí MI metodou FCS**. V analytickém systému SAS jsou k dispozici tyto metody FCS: diskriminační funkce a logistická regrese pro binární/kategoriální proměnné a lineární regrese a prediktivní porovnávání průměrů pro spojité proměnné. Ve výchozím nastavení použije SAS diskriminační funkce a regrese. Některé zajímavé vlastnosti každé z těchto možností jsou následující:

- Metoda diskriminační funkce umožňuje zadat apriorní pravděpodobnosti příslušnosti ke skupině. U diskriminační funkce mohou ve výchozím nastavení být kovariáty pouze spojité proměnné. Volbou CLASSEFFECTS= se toto dá změnit.
- Metoda logistické regrese předpokládá uspořádání klasifikačních proměnných, pokud jsou více než dvě úrovně.
- Výchozí metodou imputace pro spojité proměnné je regrese. Regresní metoda umožňuje použít rozsahy a zaokrouhlování imputovaných hodnot.
- Metoda prediktivní shody průměrů poskytne imputované hodnoty, které jsou v souladu s pozorovanými hodnotami. Pokud je nutné použít věrohodné hodnoty,

je tato metoda lepší volbou než použití mezí nebo zaokrouhlení hodnot získaných regresí.

**Imputační fáze** je opět realizována pomocí procedury MI. V rámci imputace metodou FCS byla použita pro kategorické proměnné FEMALE a PROG logistická regrese se spojovací funkcí odpovídající zobecněnému logitu. Proměnné FEMALE a PROG budou imputovány pomocí nestejné sady prediktorů. Pro imputace metodou prediktivní shody průměrů byly zvoleny spojité proměnné MATH, READ a WRITE. Dílčí imputované soubory byly procedurou MI uloženy do výstupního souboru „MI_FCS“. Toto zajišťuje následující sekvence příkazů:

```
proc mi data= ats.hsb_mar nimpute=20 out=mi_fcs;
     class female prog;
     var socst write read female math science prog;
     fcs logistic(female= math science / link=glogit);
     fcs logistic(prog = math socst /link=glogit) regpmm(math read
write);
run;
```

**Analytickou fází** procesu imputace se stejně jako u předešlé metody MI odhadnou parametry analytického modelu pro každý imputovaný soubor samostatně pomocí příkazu BY a proměnné _IMPUTATION_. Jelikož je potřebné modelovat klasifikační proměnné FEMALE a PROG imputované pomocí logistické regrese, použije se procedura GENMOD. Výstupem analytické procedury GENMOD je opět soubor odhadů parametrů pro každý z vygenerovaného imputačního souboru. Syntaxe příkazu je jako u předcházející metody:

```
proc genmod data=mi_fcs;
     class female prog;
     model read= write female math prog / dist=normal;
     by _imputation_;
     ods output ParameterEstimates=gm_fcs;
run;
```

**Slučovací fáze** imputačního procesu umožňující statistickou inferenci odhadů chybějících hodnot je podobně jako u předcházejícího způsobu MI zajištěna procedurou MIANALYZE. Vstupem do procedury je opět soubor odhadnutých parametrů imputačních modelů „GM_FCS“. Jelikož jsou do imputačního i analytického modelu zahrnuty klasifikační proměnné FEMALE a PROB, musí procedura MIANALYZE obsahovat parametr CLASSVAR=LEVEL. Pokud již v rámci imputačního modelu dojde k reparametrizaci klasifikační proměnné (například kdy se proměnná PROG rozloží na kategoriální proměnné pro jednotlivé úrovně-jak tomu bylo u metody předcházející), parametr CLASSVAR není potřebný. Příklad příkazu pro proceduru MIANALYZE je:

```
proc mianalyze parms(classvar=level)=gm_fcs;
     class female prog;
     modeleffects INTERCEPT write female math prog;
run;
```

Výstupy analýz doplněných hodnot z procedury MIANALYZE ilustruje obrázek č. 13.

***Obrázek č. 13: Odhady parametrů modelu z výstupu procedury MIANALYZE***

The MIANALYZE Procedure

| Model Information | |
|---|---|
| PARMS Data Set | WORK.GM_FCS |
| Number of Imputations | 20 |

Variance Information (20 Imputations)

| Parameter | female | prog | Variance Between | Variance Within | Variance Total | DF | Relative Increase in Variance | Fraction Missing Information | Relative Efficiency |
|---|---|---|---|---|---|---|---|---|---|
| INTERCEPT | | | 0.741973 | 11.331351 | 12.110423 | 4591.1 | 0.068754 | 0.064738 | 0.996774 |
| write | | | 0.000864 | 0.005375 | 0.006283 | 911.14 | 0.168779 | 0.146278 | 0.992739 |
| female | female | | 0.276206 | 1.161464 | 1.451480 | 475.92 | 0.249699 | 0.203149 | 0.989945 |
| female | male | | 0 | 0 | 0 | . | . | . | . |
| math | | | 0.000987 | 0.005395 | 0.006431 | 731.67 | 0.192103 | 0.163430 | 0.991895 |
| prog | | academic | 0.322298 | 1.886644 | 2.225057 | 821.38 | 0.179373 | 0.154149 | 0.992352 |
| prog | | general | 0.505042 | 2.206921 | 2.737215 | 506.22 | 0.240287 | 0.196901 | 0.990251 |
| prog | | vocation | 0 | 0 | 0 | . | . | . | . |

Parameter Estimates (20 Imputations)

| Parameter | female | prog | Estimate | Std Error | 95% Confidence Limits | | DF | Minimum | Maximum | Theta0 | t for H0: Parameter=Theta0 | Pr > \|t\| |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| INTERCEPT | | | 9.930907 | 3.480003 | 3.10843 | 16.75339 | 4591.1 | 8.084232 | 12.009803 | 0 | 2.85 | 0.0043 |
| write | | | 0.374015 | 0.079263 | 0.21846 | 0.52958 | 911.14 | 0.317881 | 0.450816 | 0 | 4.72 | <.0001 |
| female | female | | -2.485762 | 1.204774 | -4.85310 | -0.11843 | 475.92 | -3.434171 | -1.405988 | 0 | -2.06 | 0.0396 |
| female | male | | 0 | 0 | . | . | . | 0 | 0 | 0 | . | . |
| math | | | 0.426402 | 0.080196 | 0.26896 | 0.58384 | 731.67 | 0.363046 | 0.479586 | 0 | 5.32 | <.0001 |
| prog | | academic | 2.648451 | 1.491662 | -0.27947 | 5.57637 | 821.38 | 1.383416 | 4.132862 | 0 | 1.78 | 0.0762 |
| prog | | general | 0.665257 | 1.654453 | -2.58518 | 3.91570 | 506.22 | -0.856897 | 1.814812 | 0 | 0.40 | 0.6878 |
| prog | | vocation | 0 | 0 | . | . | . | 0 | 0 | 0 | . | . |

***Zdroj: vlastní zpracování***

Podobně jako u předcházející metody MI tato imputační metoda dosahuje výsledků, které se velmi blíží hodnotám odhadovaným z kompletních údajů. Srovnání odhadů parametrů lineárního regresního modelu z dat imputovaných metodou FCS, z analýzy kompletních případů a úplných dat je ilustrováno na obrázku č. 14.

***Obrázek č. 14: Porovnání odhadů analytického modelu ze sledovaných dat***

| | kompletní data | | | úplné případy | | | po 20 imputacích FCS | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | Estimate | Std Error | Pr > \|t\| | Estimate | Std Error | Pr > \|t\| | Estimate | Std Error | Pr > \|t\| |
| intercept | **9,62** | 3,41 | 0,01 | **13,03** | 4,12 | 0,00 | **9,93** | 3,48 | 0,00 |
| write | **0,37** | 0,07 | <.0001 | **0,44** | 0,09 | <.0001 | **0,37** | 0,08 | <.0001 |
| female | **- 2,70** | 1,10 | 0,01 | **- 2,71** | 1,37 | 0,05 | **- 2,49** | 1,20 | 0,04 |
| math | **0,44** | 0,07 | <.0001 | **0,32** | 0,10 | 0,00 | **0,43** | 0,08 | <.0001 |
| progcat1 | **1,88** | 1,42 | 0,19 | **1,81** | 1,65 | 0,28 | **2,65** | 1,49 | 0,08 |
| progcat2 | **0,23** | 1,51 | 0,88 | **0,52** | 1,88 | 0,78 | **0,67** | 1,65 | 0,69 |

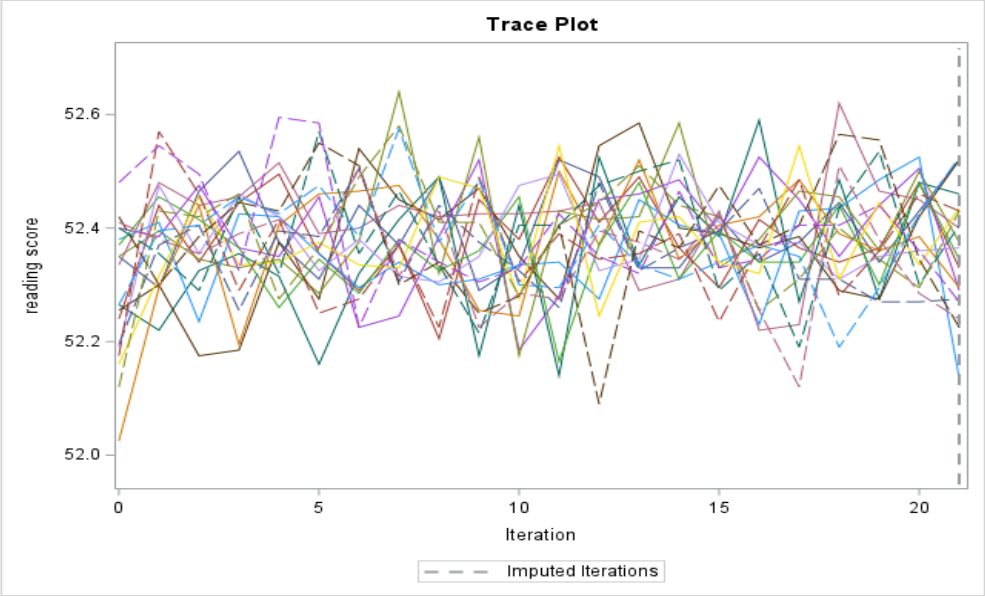
***Zdroj: vlastní zpracování***

## DIAGNOSTIKA IMPUTACÍ

Mimo běžného porovnání dosažených odhadů po provedené imputaci by datový analytik měl posoudit také konvergenci použitého imputačního modelu. Toto posouzení by mělo být provedeno pro různé imputované proměnné, ale zejména pro proměnné s vysokým podílem chybějících hodnot. Konvergence postupu procedury MI znamená, že generující algoritmus dosáhl vhodného stacionárního aposteriorního rozdělení. Konvergenci pro každou imputovanou proměnnou lze posoudit pomocí TRACE grafů. Tyto grafy lze vyžádat na řádku příkazem v proceduře MI. Dlouhodobé trendy v TRACE grafech a vysoká sériová závislost z jedné generované hodnoty na druhou svědčí o pomalé konvergenci ke stacionaritě[7]. Ve výchozím nastavení SAS poskytují TRACE grafy odhadů pro střední hodnoty pro každou proměnnou, ale je možné získat grafy i pro směrodatné odchylky. Bod dosažení stacionarity je zachycen čárkovanou kolmicí k ose iterací. Ukázku TRACE grafu představuje obrázek č. 15.

[7] *Stacionární proces má střední hodnotu a rozptyl, které se v čase nemění.*

***Obrázek č. 15: Graf konvergence imputovaných hodnot proměnné READ***




***Zdroj: vlastní zpracování***

## 4. ZÁVĚR

Chybějící data při analýzách dat představují problém, který je vhodné v zájmu získání nezkreslených odhadů a síly statistických testů optimálně řešit. Mnohonásobná imputace neúplných dat představuje moderní a spolehlivé postupy při imputaci chybějících hodnot, a to nejen jejich bodovými odhady, ale také i odpovídajícími nejistotami odhadů. Programové prostředí SAS se vyznačuje funkcionalitou mnohonásobné imputace již od verze 9. Z tohoto důvodu je systém SAS vybaven procedurami MI a MIANALYZE, které tuto funkcionalitu zabezpečují, a tak společně dotvářejí efektivní možnosti imputování neúplných dat.

## LITERATURA

Brand, J. P. L. (1999). *Development, Implementation, and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets.* Ph.D. thesis, Erasmus University

Bryk, A. S. & Raudenbush, S. W. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd Edition. Thousand Oaks, CA: SAGE Publications, Inc.

Enders, C., K. (2010). *Applied Missing Data Analysis*. New York: The Guilford Press.

Gelman, A. & Carlin, J. B. & Stern, H. S. & Dunson, B. D. & Vehtari, A. & Rubin, D. B. (2021). *Bayesian Data Analysis*, 3rd Edition. New York: Chapman & Hall/CRC.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology, 64(5), 402 – 406.*

Ling, A. & Montez-Rath, M. & Mathur, M. & Kapphahn, K. & Desai, M. (2020). How to Apply Multiple Imputation in Propensity Score Matching with Partially Observed Confounders: A Simulation Study and Practical Recommendations. *Journal of Modern Applied Statistical Methods, 19(1), 2 – 64.*

Newsom, J. T. (2023). Missing Data and Missing Data Estimation in SEM. Course of Psy 523/623 Structural Equation Modeling, Spring 2023. Portland State University. https://web.pdx.edu/~newsomj/semclass/ho_missing.pdf

Pavelka, R. (2024). Imputace chybějících dat pomocí Bayesovského modelování. *Slovenská štatistika a demografia, 34(4), 21 – 41.*

https://ssad.statistics.sk/SSaD/index.php/imputace-chybejicich-dat-pomoci-bayesovskeho-modelovani/
Pavelka, R. (2024b). Statistická analýza chybějících dat. *Slovenská štatistika a demografia, 34(2), 3 – 25.* https://ssad.statistics.sk/SSaD/index.php/statisticka-analyza-chybejicich-dat/
Raghunathan, T. E. & Lepkowski, J. M. & Van Hoewyk, J. & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology, 27(1), 85 – 95.*
Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63(3), 581 – 592.* http://dx.doi.org/10.1093/biomet/63.3.581
Rubin, D. B. (1978). Multiple imputations in sample surveys-A phenomenological Bayesian approach to nonresponse. In: Proceedings of the Survey Research Methods Section. Alexandria, VA: American Statistical Association. 20 – 34
Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.
Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association. Vol. 90, s. 822 – 828.*
Russell, D. W., & Sinharay, S. (2001). The Use of Multiple Imputation for the Analysis of Missing Data. *Psychological Methods, 6(4), 317 – 329.*
SAS Institute Inc. (2023). SAS/STAT® 15.3 User's Guide. Cary, NC: SAS Institute Inc.

**RESUMÉ**
Zejména pro analýzy metodami mnohorozměrných statistik představují chybějící údaje problém. Ačkoliv neúplná data ve výběrovém souboru mohou být zastoupena v relativně malém procentu, může tato situace v zjištěných datech vyústit v relativně velmi malý soubor s kompletními údaji; zejména v případě, kdy u různých jednotek chybí hodnoty různých veličin. V běžné praxi výběrových zjišťování jednotky, u kterých byly zaznamenány nevyplněné hodnoty zjišťovaných ukazatelů, jsou převážně z dalších analýz vyloučeny. Vynechání jednotek z analýz může mít značné negativní dopady – snížení přesnosti odhadů a síly vykonávaných statistických testů a může vést až ke zkresleným výsledkům nevhodných k zobecňování na cílovou populaci.

Důvody, proč je mnohonásobná imputace efektivnější než kterákoli z metod jednoduché imputace, jsou následující:

- nikdy se nepoužívá jediná imputovaná hodnota,
- odhady rozptylu odrážejí odpovídající míru nejistoty, která obklopuje odhady parametrů,
- před provedením vícenásobné imputace je třeba učinit několik rozhodnutí, včetně rozhodnutí o rozdělení, pomocných proměnných a počtu imputací, které mohou ovlivnit kvalitu imputace,
- i když mnohonásobná imputace není zázračná, a i když může napomoci zvýšit sílu výsledků statistických testů a zvýšit přesnost odhadů, nemělo by se od ní očekávat, že poskytne „významné" výsledky, když jiné techniky, jako je například metoda dostupných případů, nedokážou najít významné asociace a
- mnohonásobná imputace je jedním z nástrojů, kterým mohou datoví analytici řešit velmi častý problém chybějících údajů.

**RESUME**
Missing data is a problem especially for analyses using multivariate statistical methods. Although incomplete data in the random sample may be represented by a relatively

small percentage, this situation may result in a relatively very small set of complete data in the observed data; especially when the values of different quantities are missing for different units. In the normal practice of sample surveys of these units, for which unfilled values of the surveyed indicators were recorded, they are mostly excluded from further analyses. Omitting units from analyses can have significant negative impacts - reducing the accuracy of estimates and the power of the statistical tests performed - and can lead to distorted results unsuitable for generalization to the target population.

The reasons why multiple imputation is more efficient than any of the single imputation methods are as follows:

- A single imputed value is never used,
- variance estimates reflect an appropriate level of uncertainty surrounding the parameter estimates,
- Several decisions need to be made before performing multiple imputation, including decisions on the distribution, covariables, and number of imputations what can affect the quality of the imputation,
- While multiple imputation is not miraculous, and while it can help increase the power of statistical test results and improve the precision of estimates, it should not be expected to provide "significant" results when other techniques, such as the available item method, fail to find significant associations and
- Multiple imputation is one of the tools that data analysts can use to address the very common problem of missing data.

***PROFESIJNÝ ŽIVOTOPIS***

***Ing. Roman Pavelka, PhD.**, v rokoch 1995 – 2010 pracoval v poradenskej spoločnosti Trexima, s. r. o. Na pozícii štatistik analytik sa zaoberal najmä analýzami mzdových a personálnych údajov. Podieľal sa na tvorbe pravidelných štatistických prehľadov a správ. Spolupracoval s akademickými pracoviskami, agentúrami i súkromnými subjektami na realizácii a vyhodnocovaní ad hoc štatistických výskumov. Oblasť jeho vedeckého záujmu predstavujú výberové zisťovania, odhady a štatistické modely. V rokoch 2012 až 2013 sa zúčastnil na zahraničnej stáži v Spojenom kráľovstve. Od roku 2013 pôsobil v Národnom ústave certifikovaných meraní vzdelávania (NÚCEM), kde zaisťoval štatistické vyhodnocovanie výsledkov testovania žiakov a študentov. Od roku 2015 pracuje v odbore metód štatistických zisťovaní Štatistického úradu SR.*

**KONTAKT**
roman.pavelka@statistics.sk