



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MŠMT
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

Chybějící pozorování v datech a problematické aspekty jednoduchých metod řešení

Ivan Petrůšek



2021

Informace o autorech:

Mgr. Ivan Petrušek

Sociologický ústav AV ČR, v. v. i.

ivan.petrusek@soc.cas.cz



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

OBSAH

ÚVOD	0
1 KLÍČOVÉ DEFINICE	1
1.1 MECHANISMY CHYBĚJÍCÍCH HODNOT	2
2 PŘEDSTAVENÍ DAT	6
3 ZÁKLADNÍ METODY ŘEŠENÍ PROBLÉMU CHYBĚJÍCÍCH DAT	10
3.1 ANALÝZA KOMPLETNÍCH PŘÍPADŮ	10
3.2 NAHRAZOVÁNÍ CHYBĚJÍCÍCH HODNOT ARITMETICKÝM PRŮMĚREM	13
3.3 NAHRAZOVÁNÍ CHYBĚJÍCÍCH HODNOT LINEÁRNÍ REGRESÍ	14
3.4 NAHRAZOVÁNÍ CHYBĚJÍCÍCH HODNOT STOCHASTICKOU LINEÁRNÍ REGRESÍ	17
3.5 PROBLEMATICKÝ ASPEKT VŠECH METOD NAHRAZUJÍCÍCH JEDINOU HODNOTOU	18
SHRNUTÍ	20
SEZNAM POUŽITÉ LITERATURY	21
SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK	22
PŘÍLOHA P I: R SKRIPT PRO REPLIKOVÁNÍ GRAFŮ A ANALÝZ	23



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MŠMT
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

ÚVOD

Tento studijní materiál slouží jako podklad pro jednodenní workshop „*Práce s chybějícími pozorováními v společenskovedních datech*“. Tento workshop pro Univerzitu Tomáše Bati se uskutečnil vzdáleně (přes platformu *Google Meet*) 26. ledna 2021.

Cílem tohoto materiálu je představit mechanismy chybějících hodnot jako klíčový koncept analýzy dat s chybějícími pozorováními. Tyto mechanismy totiž představují klíčovou proměnnou, která determinuje, jak významný problém chybějící data představují a do jaké míry budou úspěšné základní metody řešení tohoto problému. Z hlediska metod řešení problému chybějících pozorování tento studijní materiál diskutuje *analýzu kompletních případů* (jako defaultní nastavení práce s chybějícími hodnotami ve většině počítačových programů pro analýzu kvantitativních dat), *nahrazování chybějících hodnot aritmetickým průměrem*, *nahrazování chybějících hodnot prostřednictvím lineární regrese a stochastické lineární regrese*. S výjimkou stochastické lineární imputace se jedná o nevhodné způsoby vypořádání se s chybějícími hodnotami v datech. Studijní materiál prezentuje hlavní důvody a vysvětlení, proč tyto metody ve většině situací neobstojí.

Studijní materiál prezentuje jak teoretický tak empirický argument proti využívání těchto metod v analytické praxi. Nepřímo tak odkazuje na základní motivace pro používání mnohonásobných imputací chybějících hodnot jako v současnosti nejvhodnější dostupné metodě řešení problému chybějících pozorování v datech. Mnohonásobné imputace chybějících hodnot jsou zmíněny v závěru tohoto studijního materiálu. Potenciální zájemce o tuto klíčovou metodu řešení problému lze odkázat na příslušnou literaturu (Buuren, 2018; Enders, 2010; Rubin, 1987; Schafer, 1999).

Mechanismy chybějících hodnot a jednoduché metody řešení jsou v textu vysvětlovány s využitím grafů. Kompletní analýza je plně replikovatelná prostřednictvím přiloženého R skriptu (viz *Příloha P I*). Skript s výjimkou grafu 1. nevyžaduje instalaci žádných speciálních balíčků. Mimo jiné tak skript ukazuje, jak lze diskutované metody práce s chybějícími hodnotami implementovat v základním prostředí R (tzn. bez využívání speciálních funkcí určených k nahrazování chybějících hodnot).



1 KLÍČOVÉ DEFINICE

Chybějící pozorování představují v drtivě většině společenskovedních dat všudypřítomný fenomén. Chybějící pozorování jsou pro výzkumníky a datové analytiky obtíží, se kterou se musí nějakým způsobem vypořádat. Často se přitom stává, že tuto obtíž ignorují a spokojí se s defaultním nastavením datového softwaru. Ve většině užívaných softwarů je defaultním nastavením analýza kompletních případů (*listwise deletion*), která případy s chybějícími pozorováními vyřadí z počítané věcné analýzy. Jak ukazuje tento text, takovýto ignorující přístup je často nevhodný.

Předtím, než se pustíme do diskutování výhod a nevýhod metod řešení problému chybějících pozorování v datech, představíme definici chybějících hodnot a definice mechanismů chybějících hodnot. Mechanismy určují různé druhy vztahů mezi chybějícími hodnotami a pozorovanými hodnotami v datech. Právě tyto mechanismy předurčují, jak úspěšná bude aplikovaná metoda práce s chybějícími hodnotami. Přítomný mechanismus chybějících hodnot je mnohem důležitější než podíl chybějících hodnot v datové matici. Pokud totiž data nechybí náhodně, tak už i relativně malý podíl chybějících hodnot může způsobit vážné problémy z hlediska validity výsledků věcné analýzy. Jinými slovy, neexistuje žádné pravidlo, které by určovalo, do jakého podílu chybějících hodnot lze tento problém ignorovat aplikováním analýzy kompletních případů (Buuren, 2018, s. 11).

Chybějící hodnoty představují nepozorované hodnoty, které by byly smysluplnou součástí realizované věcné analýzy, pokud by byly pozorovány (Little & Rubin, 2020, s. 4). I když je tato definice poměrně jednoznačná, tak v praxi může být někdy komplikované rozhodnout, jestli konkrétní hodnota představuje chybějící pozorování. Například v datech z dotazníkových šetření se používají popisky typu „neví“, „odmítl odpovědět“, „bez odpovědi“, „netýká se“ a několik dalších analogických popisků. Pokud skutečná hodnota existuje a jen nebyla pozorována, tak se jedná o chybějící pozorování. Mezi příklady takovýchto proměnných patří věk nebo příjem plnoletého respondenta. Každý jednotlivec musí mít pro obě proměnné nějakou hodnotu, protože se narodil/a v nějakém roce (a tudíž lze věk odvodit) a s velkou



pravděpodobností disponuje nějakým zdrojem příjmů (pokud nemá žádné příjmy, tak se nejedná o chybějící hodnotu, ale o platnou hodnotu). Jinými slovy, skutečná hodnota obou proměnných určitě existuje, i když ji respondent odmítl při dotazování uvést.

Pokud se nějaký koncept/proměnná respondenta netýká, tak se o chybějící hodnoty určitě nejedná. Jako příklad mohou posloužit otázky na partnera (např. ohledně dosaženého vzdělání, zaměstnání, věku atd.), pokud respondent aktuálně žádného partnera nemá. U části respondentů, kteří nemají partnera, bude u všech příslušných otázek na partnera uvedeno, že nemají partnera, nebo bude v datové matici „prázdné místo“ (potenciálně popisek „netýká se“). Tato „prázdná místa“ v datech nemá smysl nahrazovat, protože příslušné hodnoty neexistují (a tudíž nepředstavují chybějící hodnoty).

1.1 Mechanismy chybějících hodnot

Mechanismy chybějících hodnot definují vztahy mezi chybějícími hodnotami a pozorovanými hodnotami v datové matici. Zjednodušeně řečeno, tyto mechanismy definují důvody, proč pozorování v datech chybí. Teorie chybějících hodnot definuje tři různé mechanismy chybějících hodnot. Tyto mechanismy a příslušnou terminologii poprvé definoval americký statistik Donald Rubin ve vlivném článku z druhé poloviny sedmdesátých let minulého století (Rubin, 1976), přičemž terminologii dál rozvíjel v dalších svých publikacích (např. Little & Rubin, 2002). Pro mechanismy chybějících hodnot se v literatuře prosadili zkratky zavedené právě Rubinem. Jelikož jsou tyto zkratky v příslušné literatuře všudypřítomné, je potřeba si jejich význam pamatovat.

Prvním mechanismem je tzv. mechanismus MCAR (*missing completely at random*). Podle tohoto mechanismu chybí hodnoty zcela náhodně. To znamená, že výskyt chybějících pozorování v datové matici nezávisí od pozorovaných hodnot, a zároveň nezávisí ani od hodnot, které v datech chybí. Jinými slovy, výskyt chybějících hodnot v datech je absolutně náhodný. Pro ilustraci si představme data, která obsahují pouze dvě proměnné, dosažené vzdělání a postoj k očkování proti onemocnění covid-19, přičemž data chybí pouze u postoje k očkování. Tato data by chyběla zcela náhodně, pokud by pravděpodobnost výskytu chybějících hodnot u postoje byla stejná ve všech skupinách dosaženého vzdělání, a zároveň by



neexistovala žádná souvislost mezi zastávaným postojem k očkování a odmítáním uvádět tento postoj v dotazníku.

Druhým mechanismem je mechanismus MAR (*missing at random*). Překladem tohoto názvu do češtiny získáme částečně zavádějící název, že data *chybí náhodně*. Jedná se o zavádějící název, protože u tohoto mechanismu závisí výskyt chybějících hodnot na pozorovaných hodnotách, ale nezávisí na samotných chybějících hodnotách. Zjednodušeně řečeno, hodnoty pozorovaných proměnných způsobují výskyt chybějících hodnot u dalších proměnných v datech. V našem hypotetickém příkladu se dvěma proměnnými by hodnoty postoje k očkování chyběly podle mechanismu MAR, pokud by v jednotlivých skupinách dosaženého vzdělání existovala rozdílná pravděpodobnost odmítat uvést postoj k očkování, ale odmítání uvádět postoj by nezáviselo od samotného zastávaného postoje.

Posledním mechanismem je mechanismus MNAR (*missing not at random*).¹ Jedná se o nejproblematictější mechanismus, protože v této situaci výskyt chybějících hodnot závisí od samotných hodnot, které chybí. Zjednodušeně řečeno, samotné chybějící hodnoty u konkrétní proměnné způsobují výskyt chybějících hodnot u této proměnné. Pro náš hypotetický příklad se dvěma proměnnými by hodnoty postoje k očkování chyběly podle mechanismu MNAR, pokud by lidé s rozdílnými postoji k očkování v různé míře odmítali odpovídat na otázku ohledně svého postoje k očkování. Například bychom si mohli představit situaci, že lidé s negativním postojem k očkování by měli mnohem větší pravděpodobnost odmítat vyjádřit tento svůj postoj při dotazování než lidé s pozitivním postojem k očkování.

V praxi není příliš pravděpodobné, aby data chyběla zcela náhodně. Mechanismus MCAR tedy není v analýze dat s chybějícími pozorováními příliš častý. Mnohem častěji se setkáváme s mechanismy MAR nebo MNAR. Z hlediska praktické analýzy představuje velkou komplikaci skutečnost, že analytik nikdy s jistotou neví, podle kterého mechanismu jeho

¹ Pro tento mechanismus používali oba autoři v starších edicích jejich vlivné knihy označení NMAR (*not missing at random*) (Little & Rubin, 2002, s. 12-19). Označení NMAR a MNAR tedy označují ten samý mechanismus, přičemž v literatuře o chybějících pozorováních se lze setkat s oběma označeními.

data chybí. Tato nejistota je způsobena samotnou povahou chybějících dat: jelikož tyto hodnoty neznáme, tak nemůžeme srovnávat chybějící hodnoty u konkrétní proměnné s pozorovanými hodnotami této proměnné.

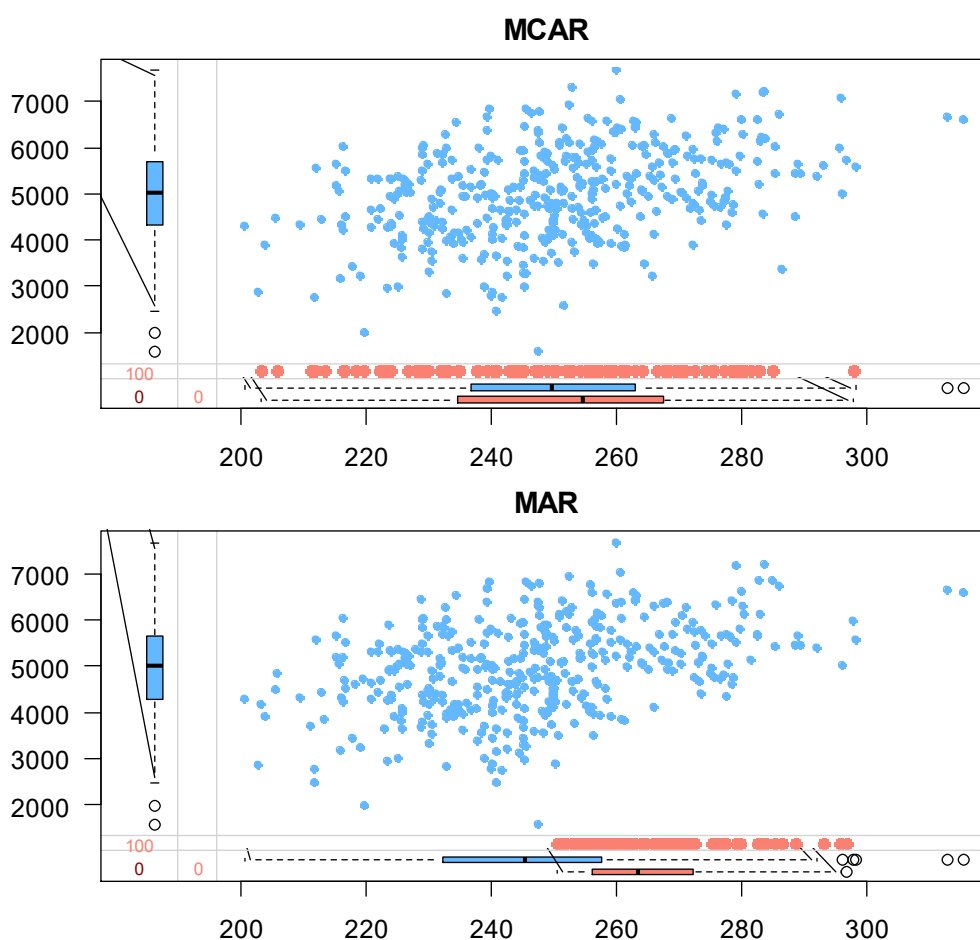
Při práci s daty, která obsahují chybějící pozorování, je před vlastní analýzou potřeba prozkoumat vztahy mezi chybějícími a pozorovanými hodnotami. V softwarech na analýzu dat je k tomu implementovaných hned několik procedur. Asi nejznámější je Littleův MCAR test (Little, 1988), který pro kardinální proměnné testuje globální nulovou hypotézu, že u nich data chybí zcela náhodně. Pro kardinální proměnné lze použít také t-testy pro dva nezávislé výběry, přičemž srovnávané skupiny jsou určeny tím, jestli jsou hodnoty u nějaké další proměnné pozorovány, nebo chybí. U kategoriálních proměnných lze srovnávat relativní četnosti platných odpovědí v skupinách, podle toho, jestli chybí hodnoty u nějaké další proměnné. Pokud se tyto četnosti výrazně liší, tak to indikuje, že data nechybí podle mechanismu MCAR, ale spíše podle mechanismu MAR.

Mechanismy chybějících hodnot lze zkoumat také graficky. K tomuto účelu se u kardinálních proměnných často využívá marginální graf (*marginsplot*).² U takového bodového grafu jsou chybějící hodnoty znázorněny po stranách červenou barvou, přičemž případy, které mají pozorované hodnoty u obou proměnných, jsou znázorněny modrou barvou ve vlastním bodovém grafu. Příklady dvou marginálních grafů na simulovaných datech vytvořených pro potřeby tohoto studijního materiálu znázorňuje graf 1. Z obou marginálních grafů lze vyčíst, že data chybí pouze u proměnné na vertikální ose grafu. Pozorované hodnoty proměnné na horizontální ose pro 100 případů, u nichž chybí hodnoty proměnné na vertikální ose, jsou znázorněny jako červené kroužky ve spodní části grafů. Z vrchního marginálního grafu je patrné, že jsou tyto červené body rozloženy po celé šířce horizontální osy. Tato skutečnost naznačuje, že data ve vrchním marginálním grafu by mohla chybět podle mechanismu MCAR (další možnost, kterou nelze vyloučit, je mechanismus MNAR). Spodní

² Existuje několik dalších specifických grafických nástrojů (např. *spineplot* a *spinogram*), které byly vyvinuty s cílem prozkoumávat mechanismy chybějících hodnot v datech. Pro omezený rozsah tohoto studijního materiálu tyto nástroje nediskutujeme.

marginální graf ukazuje, že se chybějící hodnoty proměnné zobrazené na vertikální ose vyskytují pouze u nadprůměrných hodnot proměnné zobrazené na vertikální ose. Tato skutečnost vylučuje možnost, že by data chyběla podle mechanismu MCAR. Jinými slovy se zdá, že výskyt chybějících hodnot u proměnné y je asociován s vyššími hodnotami proměnné x , a tudíž data ve spodním marginálním grafu chybí podle mechanismu MAR. I tu je však nutno dodat, že ve spodním marginálním grafu by mohla data potenciálně chybět podle mechanismu MNAR.

Graf 1. Marginální graf při dvou mechanismech chybějících hodnot

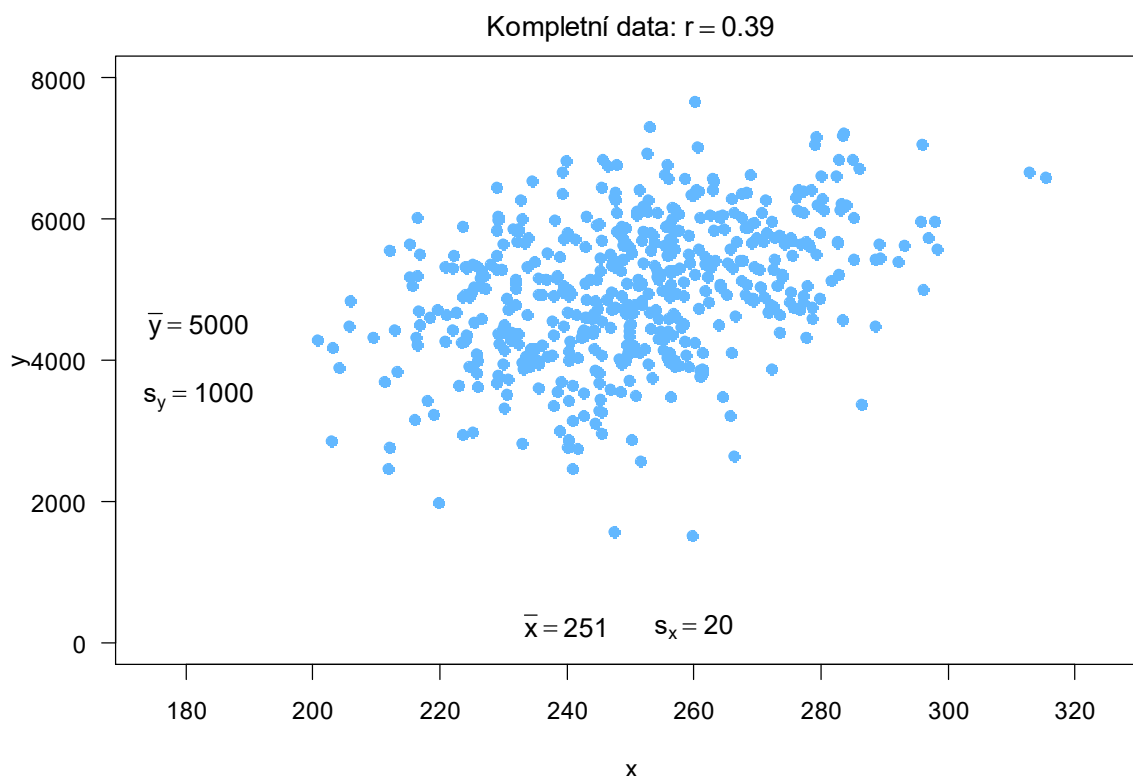


Poznámka: Oba marginální grafy byly vytvořeny s využitím funkce *marginplot* z balíčku **VIM**.

2 PŘEDSTAVENÍ DAT

Abychom si dokázali lépe představit, co jednotlivé metody práce s chybějícími hodnotami dělají, a jaké jsou implikace jejich aplikování v praxi, vytvořili jsme pro potřeby tohoto studijního materiálu v programu R simulovaný datový soubor. Tato simulovaná data jsou tvořena dvěma proměnnými (s technickým označením x a y). Konkrétně bylo vytvořeno 500 případů, které jsou znázorněny na grafu 2. Případy byly generovány z dvourozměrného normálního rozdělení, přičemž hodnota Pearsonova korelačního koeficientu mezi proměnnými je 0,39. Proměnná na horizontální ose má průměr 251 a směrodatnou odchylku 20 ($\bar{x} = 251, s_x = 20$). Proměnná na vertikální ose má průměr 5000 a směrodatnou odchylku 1000 ($\bar{y} = 5000, s_y = 1000$). Celá analýza prezentovaná v tomto studijním materiálu je plně replikovatelná (kompletní R skript je součástí *Přílohy P I*).

Graf 2. Kompletní simulovaná data

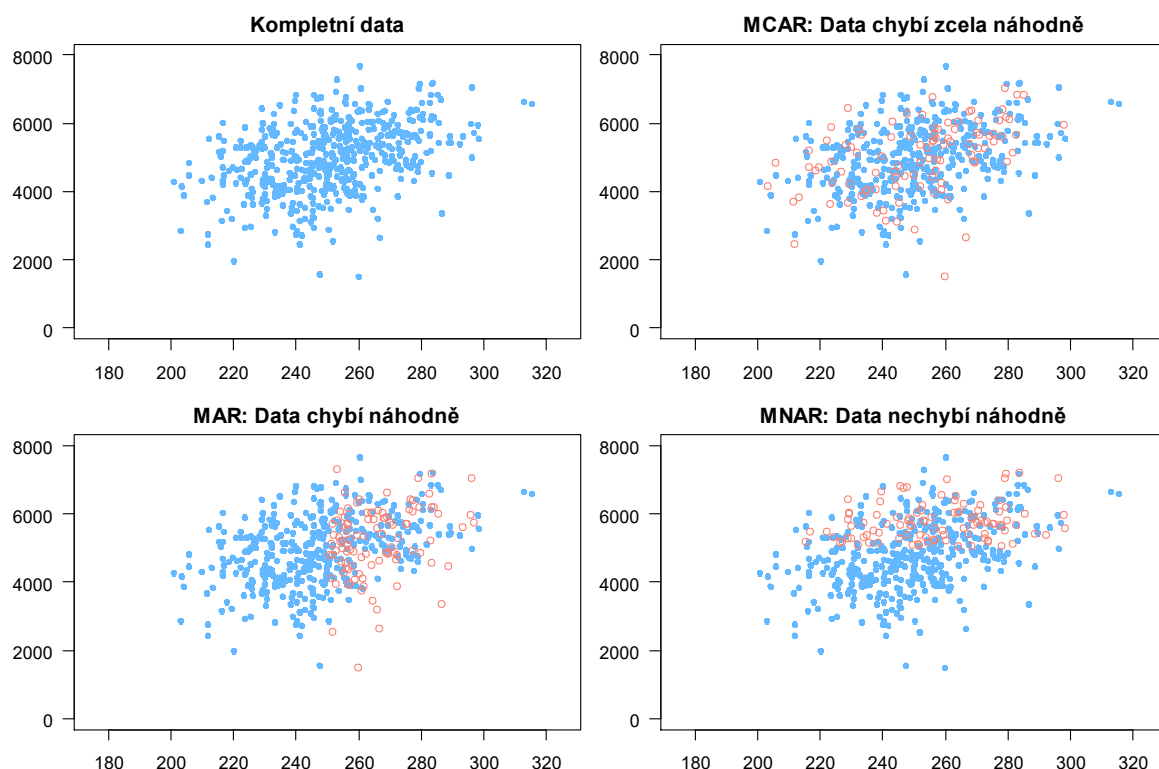


Po vytvoření kompletních dat ze simulovaného datového souboru vymažeme pětinu případů podle jednotlivých mechanismů chybějících hodnot. Pro zjednodušení analýzy a interpretace výsledků vymazáváme pouze hodnoty proměnné y . Graf 3 znázorňuje, u kterých případů byly smazány hodnoty proměnné y . U mechanismu MCAR bylo náhodně vybráno 100 případů, u nichž byla smazána původní hodnota proměnné y . Na grafu 3 jsou tyto smazané hodnoty (reprezentované červenými prázdnými kroužky) rozmístěny po celé ploše bodového grafu vpravo nahoře. Tato skutečnost ilustruje, že data chybí zcela náhodně.

Jelikož u mechanismu MAR závisí výskyt chybějících hodnot od pozorovaných hodnot, tak vymazáváme hodnoty proměnné y na základě pozorovaných hodnot proměnné x . Pro potřeby tohoto studijního materiálu bylo smazáno 100 hodnot proměnné y u případů, které mají nadprůměrné hodnoty proměnné x . Jinými slovy, nejdříve byly vyfiltrovány všechny případy s nadprůměrnou hodnotou proměnné x (celkem se jednalo o 247 případů) a následně bylo v této skupině náhodně smazáno 100 hodnot proměnné y . Chybějící data podle mechanismu MAR jsou znázorněna v levé dolní části grafu 3. Příslušný bodový graf ilustruje, že smazané hodnoty se skutečně nachází výhradně v pravé části grafu (tzn. vpravo od průměru proměnné na horizontální ose grafu).

Graf 3. Kompletní data a chybějící pozorování podle tří mechanismů chybějících hodnot





Poznámka: Případy s pozorovanými hodnotami u obou proměnných jsou znázorněny modrou barvou. Případy, u kterých byla smazána hodnota proměnné na vertikální ose, jsou znázorněny jako červené kroužky.

Nakonec představíme data chybějící podle mechanismu MNAR. Při tomto mechanismu závisí výskyt chybějících hodnot od samotných hodnot, které chybí. V našem případě byly smazány hodnoty proměnné y u případů, které měly hodnotu proměnné y větší než medián této proměnné. Konkrétně tedy byly nejdříve vyfiltrovány všechny případy s hodnotou proměnné y větší než medián (celkem se jednalo o 250 případů) a následně bylo v této skupině náhodně smazáno 100 hodnot proměnné y . Chybějící data podle mechanismu MNAR jsou znázorněna v pravé dolní části grafu 3. Příslušný bodový graf ukazuje, že se smazané hodnoty nachází výhradně v horní polovině příslušného grafu (což odpovídá mechanismu MNAR).

V tomto studijním materiálu byly chybějící pozorování u mechanismů MAR a MNAR vytvořeny s relativně striktními formálními omezeními. V reálných datech lze očekávat, že se budou chybějící hodnoty vyskytovat po celé ploše příslušných bodových grafů, přičemž

pouze v některých částech grafů bude větší hustota chybějících pozorování. Klíčovým parametrem, který ovlivňuje problematičnost těchto dvou „nenáhodných“ mechanismů při aplikování různých metod řešení problémů chybějících dat, je *síla vztahu mezi hodnotami pozorovaných (u mechanismu MAR) a neznámých hodnot (u mechanismu MNAR), a výskytem chybějících hodnot v analyzovaných datech*. Například při mechanismu MAR by byla významně problematičtější situace, pokud by chybělo 20 % dat proměnné y pro 20 % případů s nejvyššími hodnotami proměnné x . Situace by byla naopak méně problematická (ve srovnání se situací znázorněnou na grafu 3), pokud by se chybějící hodnoty vyskytovaly po celé šířce grafu a pravděpodobnost výskytu chybějících hodnot by se jen mírně zvyšovala s rostoucími hodnotami proměnné x .



3 ZÁKLADNÍ METODY ŘEŠENÍ PROBLÉMU CHYBĚJÍCÍCH DAT

V následující stěžejní části studijního materiálu postupně představujeme čtyři základní metody pro práci s chybějícími hodnotami. U těchto čtyř metod pak na vytvořených simulovaných datech ukazujeme, jak tyto metody ovlivňují odhad průměru a rozptylu proměnné s chybějícími hodnotami (v našem případě se jedná o proměnnou y) a odhad korelačního koeficientu mezi oběma proměnnými. Grafické ilustrace přitom umožňují lépe pochopit, jak tyto metody fungují a proč ve většině situací nepředstavují vhodné metody řešení problému chybějících pozorování. V příslušných grafech 4, 5, a 6 jsou případy s pozorovanými hodnotami u obou proměnných znázorněny modrou barvou a případy s nahrazenými hodnotami proměnné y znázorněny červenou barvou. I když v případě tohoto studijního materiálu pracujeme vždy pouze s jedinou aplikací konkrétní metody u každého ze tří mechanismů chybějících hodnot, tak prezentované závěry byly odvozeny jak analyticky (Little & Rubin, 2020), tak demonstrovány řadou simulačních studií (Enders, 2010; Petrušek, 2015; Schafer & Graham, 2002).

3.1 Analýza kompletních případů

Analýza kompletních případů, která bývá v softwarech pro analýzu dat většinou označována jako *listwise deletion*, před samotným spuštěním výpočtů požadované věcné analýzy vynechá z výpočtu všechny případy, u kterých chybí hodnota alespoň jedné proměnné vstupující do analýzy. To znamená, že do věcné analýzy vstupuje podmnožina původní datové matice, u které jsou pozorovány všechny hodnoty všech analyzovaných proměnných. Jedná se o jednu z nejstarších metod řešení problému chybějících hodnot, která je pro svoji jednoduchost v praxi velmi často využívána (v mnoha statistických softwarech se jedná o defaultní metodu nakládání s chybějícími pozorováními).³

³ Například SPSS, STATA a SAS mají jako defaultní metodu nastavenou právě *listwise deletion* (Buren, 2018, s. 4).

Obecně lze konstatovat, že analýza kompletních případů vede k nevychýleným odhadům všech relevantních parametrů pouze u mechanismu MCAR. Tuto skutečnost potvrzuje také graf 4, když odhady všech zkoumaných parametrů jsou velmi podobné (některé jsou dokonce stejné) jako odhady těchto parametrů na kompletních datech (graf 2). U mechanismu MCAR totiž případy vstupující do věcné analýzy při aplikaci analýzy kompletních případů představují náhodný výběr ze všech případů v datové matici. Jinými slovy, 400 případů zobrazených v grafu 4 představuje náhodný výběr z 500 případů v kompletních datech.

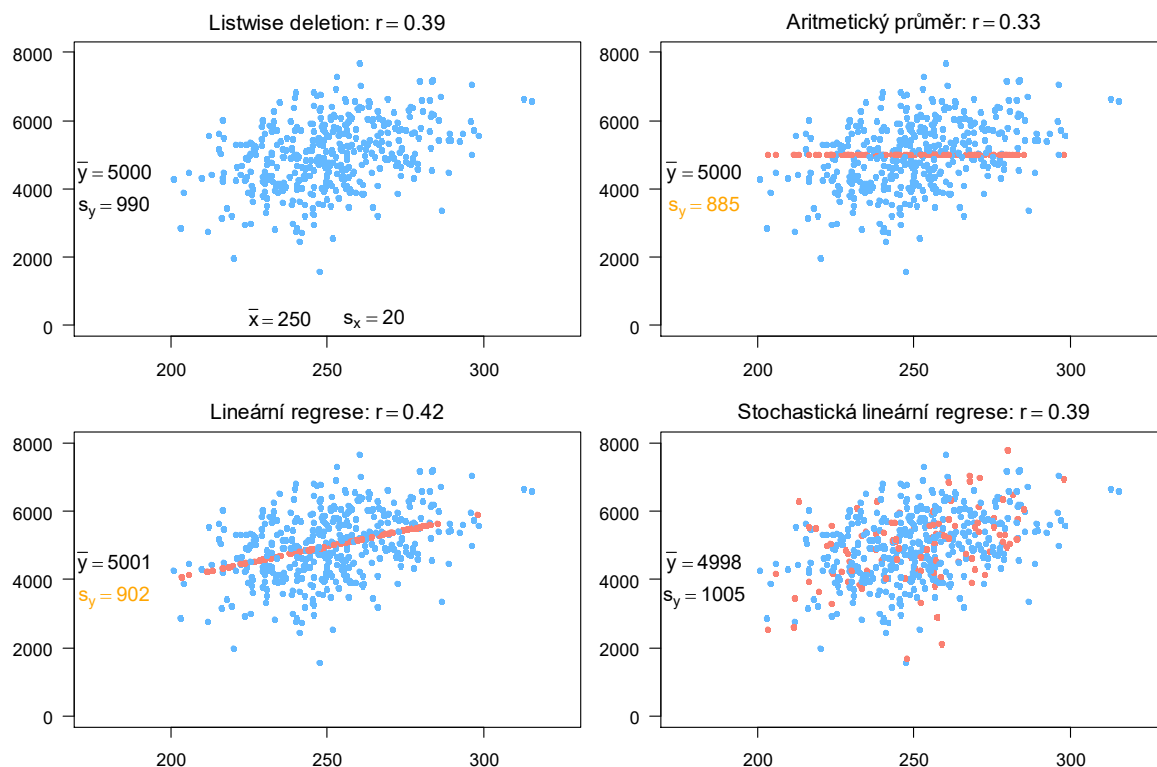
U mechanismů MAR a MNAR už metoda vede k vychýleným odhadům (některých) parametrů. Konkrétní směr vychýlení závisí od příčiny vzniku chybějících dat a velikosti příslušných parametrů způsobujících výskyt chybějících pozorování v datech. Například u mechanismu MNAR (graf 6) vidíme, že pokud chybí hodnoty pouze u případů s vyššími hodnotami proměnné y , tak logicky odhad průměru proměnné y bude významně podhodnocen ($\bar{y} = 4808$ ve srovnání s průměrem $\bar{y} = 5000$ na kompletních datech).

Analýza kompletních případů může být velmi užitečná při odhadování regresních modelů. Pokud výskyt chybějících hodnot v datech závisí jen na hodnotách nezávisle proměnných (a nezávisí na hodnotách závisle proměnné), vede analýza kompletních případů k nevychýleným odhadům regresních koeficientů při všech třech mechanismech chybějících hodnot. Tato výhodná vlastnost analýzy kompletních případů platí v regresní analýze, když se chybějící hodnoty vyskytují jak u závisle tak u nezávisle proměnných (Buuren, 2018, s. 57–58).

Podstatnou nevýhodu analýzy kompletních případů představuje skutečnost, že se dobrovolně zbavujeme informací obsažených v datech. U každého vyřazeného případu z analýzy totiž ignorujeme hodnoty všech proměnných, které jsou u daného případu pozorovány. Pokud věcná analýza pracuje s větším počtem proměnných, tak může dojít k výraznému snížení počtu případů, na kterých se věcná analýza spočítá. Toto může být problematické například při budování výše zmíněných regresních modelů, když postupně přidávání nezávisle pro-

měnných s chybějícími pozorováními do modelu vede ke snižování počtu případů, na kterých se model spočítá. Každý z modelů je tak odhadnut na jiné podmnožině datového souboru.⁴

Graf 4. Fungování metod při mechanismu MCAR



Poznámka: Případy s pozorovanými hodnotami u obou proměnných jsou znázorněny modrou barvou. Případy s nahrazenými hodnotami proměnné y jsou znázorněny červenou barvou. U listwise deletion k nahrazování chybějících hodnot nedochází, tudíž graf vlevo nahoře neobsahuje žádné červené kroužky.

Další nevýhodou analýzy kompletních případů je snižování počtu případů při statistickém usuzování/inferenci. Počet případů totiž figuruje ve jmenovateli pro výpočet směrodatných

⁴ Tomuto problému je možné se vyhnout tak, že si výzkumník před realizací jakýchkoli analýz vytvoří podmnožinu datového souboru s požadovanými proměnnými, která bude obsahovat pouze případy s pozorovanými hodnotami u všech proměnných.

chyb odhadovaných parametrů. S menším počtem případů jsou tedy nezbytně asociovány větší směrodatné chyby odhadů (tento vztah není lineární). Větší směrodatné chyby se pak odrazí v širších intervalech spolehlivosti pro zkoumané parametry a vyšších p -hodnotách u statistických testů. Jinými slovy, kvůli menšímu počtu případů se může stát, že nezamítáme nulovou hypotézu testu, kterou bychom v případě nepřítomnosti chybějících pozorování v datové matici zamítali.

3.2 Nahrazování chybějících hodnot aritmetickým průměrem

Při nahrazování chybějících hodnot aritmetickým průměrem se nejdříve na kompletních případech nahrazované kardinální proměnné vypočítá průměr a tímto průměrem jsou následně nahrazeny všechny chybějící hodnoty příslušné kardinální proměnné. Pro nominální a ordinální proměnné existují analogie nahrazování chybějících hodnot aritmetickým průměrem. U nominálních proměnných jsou chybějící hodnoty nahrazovány modem a u ordinálních proměnných jsou chybějící pozorování nahrazovány mediánem. Výsledkem aplikace těchto tří analogických metod je tedy „kompletní“ datová matice, která neobsahuje žádné chybějící hodnoty.

Nahrazování chybějících hodnot aritmetickým průměrem představuje jednoznačně nejhorší metodu řešení problému chybějících pozorování v datech. Tato metoda totiž vede k vychýleným odhadům důležitých parametrů už při mechanismu MCAR.⁵ Jak ukazují grafy 4, 5 a 6, tak všechny nahrazené hodnoty proměnné y leží na přímce, která je rovnoběžná s horizontální osou grafu.⁶ Červené kroužky ležící na této přímce tudíž mají nulový rozptyl (jelikož byly všechny chybějící hodnoty nahrazeny konstantou – příslušným aritmetickým průměrem). Pokud tedy část hodnot nahrazované proměnné disponuje pouze konstantní hodnotou, tak příslušná metoda práce s chybějícími hodnotami významným způsobem podhodnocuje

⁵ Jedinou výjimku v tomto případě představuje odhad aritmetického průměru nahrazované proměnné, který je nevychýlený při mechanismu MCAR.

⁶ Pokud by chyběly hodnoty také u proměnné x , tak nahrazené hodnoty pro tuto druhou proměnnou by ležely na vertikální přímce, která by byla rovnoběžná s vertikální osou grafu.

odhad rozptylu nahrazované proměnné. To se následně projeví také v odhadnuté směrodatné chybě průměru nahrazované proměnné, která bude významně podhodnocená. Podhodnocená směrodatná chyba se následně projeví v užších intervalech spolehlivosti a menších p -hodnotách.

Jelikož nahrazené hodnoty leží na přímce rovnoběžné s osou grafu, tak tyto hodnoty mají nulovou korelaci s proměnnou zobrazenou na příslušné ose. Přesněji řečeno, korelaci mezi oběma proměnnými ve skupině 100 případů není možné spočítat, protože směrodatná odchylka proměnné y se rovná nule, a tedy součin směrodatných odchylek obou proměnných ve jmenovateli pro výpočet korelačního koeficientu se také rovná nule. Pokud podmnožina dat v bodovém grafu má nulovou korelaci, tak tato skutečnost bude nevyhnutně podhodnocovat odhad celkové korelace mezi proměnnými. Tuto velkou slabinu nahrazování chybějících hodnot aritmetickým průměrem dobře ilustrují příslušné grafy 4, 5 a 6, když u všech tří mechanismů chybějících hodnot je odhadnutá korelace výrazně nižší (0,33; 0,36 a 0,34) než korelace vypočtená na původním kompletním datovém souboru (0,39 v grafu 2).

Z výše uvedených důvodů nelze nahrazování chybějících hodnot aritmetickým průměrem (a ani mediánem nebo modem) vůbec doporučit (Little & Rubin, 2020, s. 70). Tato jednoduchá metoda představuje spíše historický relikt (jedná se o jednu z nejstarších metod řešení problému). Využívání této metody při existenci mnohem efektivnějších metod tedy znamená ignorování poznatků a inovací, které byly v oblasti chybějících pozorování uskutečněny za posledních přibližně padesát let.

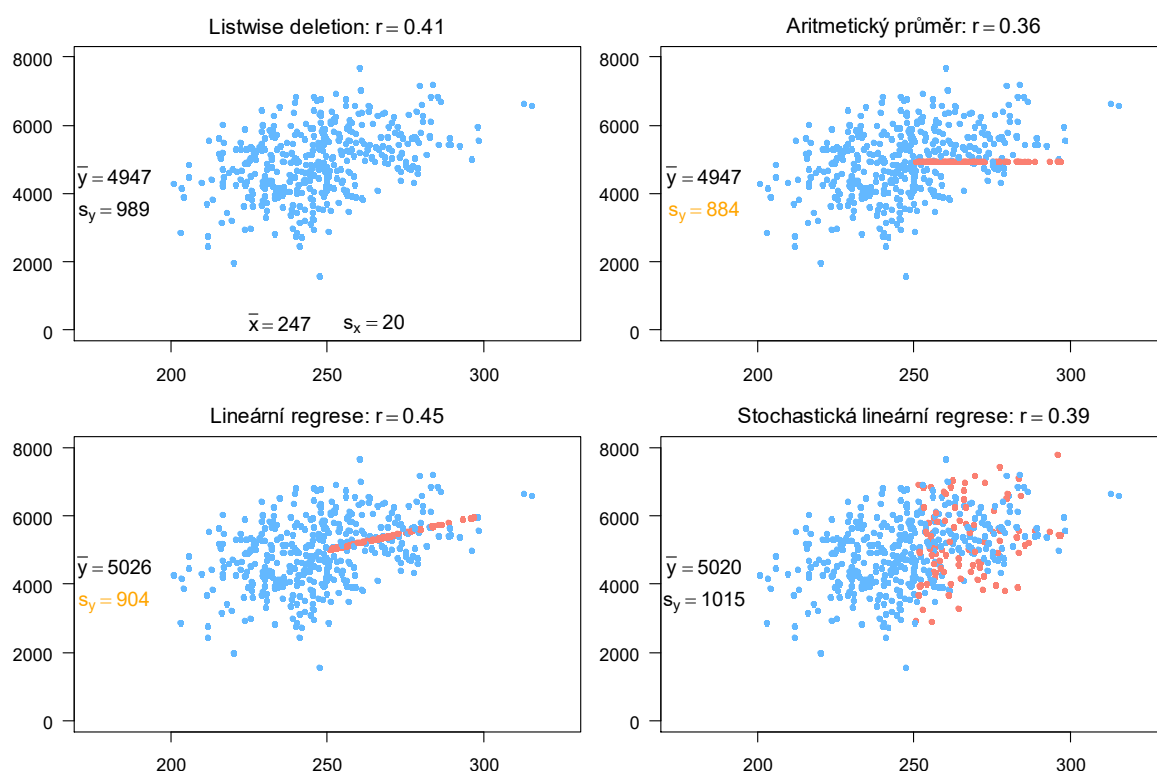
3.3 Nahrazování chybějících hodnot lineární regresí

Nahrazování chybějících hodnot prostřednictvím lineární regrese ve srovnání s nahrazováním chybějících hodnot aritmetickým průměrem využívá také informace o hodnotách dalších proměnných v datech. U této metody se nejdříve na kompletních případech odhadne regresní rovnice, která popisuje vztahy mezi vybranými proměnnými v datové matici. Proměnná s chybějícími hodnotami v této rovnici představuje závisle proměnnou a proměnné s pozorovanými hodnotami představují nezávisle proměnné. V následujícím kroku se pou-



žijí odhadnuté regresní koeficienty k nahrazování chybějících hodnot, přičemž hodnoty pozorovaných proměnných se pro každý případ s chybějící hodnotou dosadí do odhadnuté regresní rovnice. V případě našich simulovaných dat na grafu 4 se tedy nejdříve na 400 kompletních případech odhadla regresní rovnice ($y = 214,1 + 19,1x$). Následně u 100 případů s chybějícími hodnotami proměnné y byla do regresní rovnice dosazena hodnota proměnné x a odhadnuta nahrazená hodnota proměnné y (pro 100 případů).

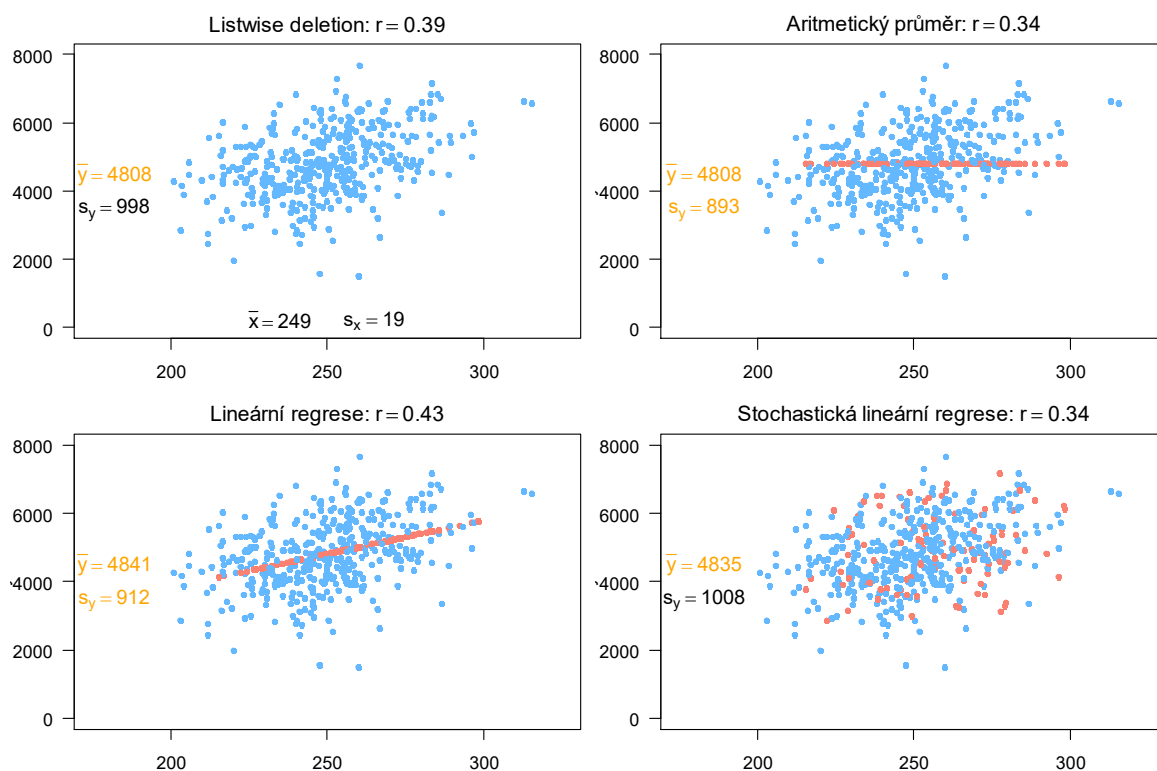
Graf 5. Fungování metod při mechanismu MAR



Jak ukazují grafy pro všechny tři mechanismy, po nahrazení chybějících hodnot lineární regrese leží všechny nahrazené hodnoty na přímce s nenulovou směrnici. V našem případě má přímka kladnou směrnici, protože mezi oběma proměnnými existuje kladná korelace. Ze skutečnosti, že má přímka nenulovou směrnici vyplývá, že nahrazené hodnoty nemají nulový rozptyl. Z grafů je ale rovněž patrné, že nahrazené hodnoty nemají takový rozptyl, jaký měla původně pozorovaná data. Nahrazování chybějících hodnot lineární regrese tedy podhodnocuje rozptyl nahrazované proměnné při všech mechanismech chybějících hodnot.

Jelikož chybějící hodnoty jsou u této metody nahrazovány deterministicky, tak nahrazené hodnoty mají s hodnotami dalších proměnných perfektní závislost. Pro naše simulovaná data to znamená, že u 100 případů s nahrazenými hodnotami je korelace mezi proměnnými x a y rovná 1. Protože u pětiny případů pozorujeme perfektní lineární závislost, tak při počítání korelace na všech 500 případech vede tato metoda k nadhodnocování celkové korelace. Ve srovnání s korelací na úrovni 0,39 (grafy 2) získáme u jednotlivých mechanismů výrazně vyšší korelace (0,42; 0,45 a 0,43). Zjednodušeně řečeno, tento způsob nahrazování chybějících hodnot uměle nadhodnocuje závislosti mezi proměnnými.

Graf 6. Fungování metod při mechanismu MNAR



Nahrazování chybějících hodnot lineární regresí představuje ve srovnání s nahrazováním aritmetickým průměrem jen marginální zlepšení. Metoda vede k nevychýleným odhadům pouze u průměru nahrazované proměnné u mechanismů MCAR a MAR (to znamená, že pokud výskyt chybějících hodnot u proměnné nezávisí od samotných hodnot, které chybí).

U ostatních parametrů (rozptyly, korelační a regresní koeficienty) ale tato metoda vede k vychýleným odhadům. Tuto metodu tudíž nelze doporučit pro praktické používání. Stejně tak nelze doporučit ani její analogie pro nominální proměnné (binární/multinomiální logistickou regresi) a ordinální proměnné (ordinální logistickou regresi).

3.4 Nahrazování chybějících hodnot stochastickou lineární regresí

Poslední metodou diskutovanou v rámci tohoto studijního materiálu je nahrazování chybějících hodnot stochastickou lineární regresí. Jedná se o rozšíření nahrazování chybějících hodnot lineární regresí. Ve srovnání s lineární regresí přidává jeden nový krok: ke každé nahrazené hodnotě prostřednictvím lineární regrese se ještě připočítá odchylka náhodně vygenerovaná z normálního rozdělení. Cílem tohoto speciálního kroku je zrestaurovat variabilitu nahrazených dat takovým způsobem, aby tato variabilita co nejblíže připomínala variabilitu původních dat. Příslušné normální rozdělení, z kterého jsou odchylky generovány, má průměr nula a rozptyl odpovídající reziduálnímu rozptylu regresního modelu. Pokud je například reziduální rozptyl lineárního modelu pro nahrazování chybějících hodnot malý, tak budou příslušné červené body kolem pomyslné regresní přímky rozmístěny relativně blízko. Naopak, pokud je reziduální rozptyl modelu pro nahrazování chybějících hodnot velký, tak budou červené body rozmístěny kolem pomyslné regresní přímky mnohem výrazněji.

Stochastická lineární imputace představuje nejlepší dostupnou jednoduchou metodu pro práci s chybějícími daty. Tato metoda vede k nevychýleným odhadům klíčových parametrů při mechanismech MCAR a MAR. Grafy 4 a 5 demonstrují, že připočítání odchylek náhodně vygenerovaných z normálního rozdělení restauruje rozptyl nahrazené proměnné y . Je patrné, že po tomto kroku jsou červené body rozptýleny po grafu způsobem, který bychom očekávali, pokud by naše data neobsahovala žádné chybějící pozorování. Na těchto grafech variuje odhadnutý průměr a rozptyl proměnné y , korelace mezi proměnnými x a y jen v rámci výběrové chyby. Jinými slovy, nejedná se o systematické vychýlení, které bychom dostali při velkém počtu opakování simulace, ale o variování v rámci výběrové chyby.

Problematicčnost mechanismu MNAR podtrhuje skutečnost, že ani stochastická lineární regrese (a její analogie pro nominální a ordinální proměnné) nevede u tohoto mechanismu k nevychýleným odhadům zkoumaných parametrů. I když na grafu 6 vypočítaný rozptyl



proměnné y odpovídá skutečnému rozptylu, tak tato vlastnost stochastické lineární regrese není garantována. Jinými slovy, odhad tohoto rozptylu je důsledkem generování dat podle mechanismu MNAR v tomto konkrétním případě, a nelze ho zobecnit do konstatování o vhodnosti metody při odhadování rozptylu nahrazované proměnné. Pokud by byl způsob generování chybějících pozorování podle mechanismu MNAR jiný, tak je možné, že by byl vychýlený i odhad tohoto parametru (pro srovnání viz Petrůšek, 2015, Kapitoly 5–6).

3.5 Problematický aspekt všech metod nahrazujících jedinou hodnotou

Dalším problémem všech tří diskutovaných metod, které nahrazují každou chybějící hodnotu pouze jednou hodnotou, je, že v následné věcné analýze se už dále nerozlišuje, jestli byla konkrétní hodnota původně pozorována, nebo původně chybějící. Jinými slovy, v následné věcné analýze se s nahrazenými hodnotami pracuje takovým způsobem, jako by byly původně pozorovány. Věcná analýza tak vůbec nebere v potaz nejistotu spojenou s procesem nahrazování chybějících hodnot. Důsledkem tohoto aspektu jsou podhodnocené směrodatné chyby odhadovaných parametrů (a tudíž užší intervaly spolehlivosti pro příslušné parametry).

Právě tuto nejistotu spojenou s nahrazováním chybějících pozorování řeší mnohonásobné imputace chybějících hodnot. Tento přístup, který je v současnosti považován za nejvhodnější způsob vypořádání se s chybějícími daty, nahrazuje každé chybějící pozorování několika hodnotami současně. Rozptýlenost nahrazených hodnot u jednoho chybějícího pozorování má zohlednit skutečnost, že jednotlivé nahrazené hodnoty jsou pouze jedny z mnoha možných nahrazených hodnot. Tyto nahrazené hodnoty by v žádném případě neměly být považovány za původně pozorované hodnoty (tak jak to dělají v tomto studijním materiálu diskutované jednoduché metody). Původcem revoluční myšlenky mnohonásobných imputací je už zmiňovaný Donald Rubin (Rubin, 1987).

Mnohonásobné imputace sice představují nejvhodnější dostupnou metodu řešení chybějících pozorování v datech, ale i tento přístup má své nevýhody. Nevhodná implementace procedury negarantuje, že se nám podaří vypořádat s nejkompexnějším mechanismem chybějících hodnot MNAR. Celá procedura je relativně komplikovaná (proces se skládá ze tří fází:





imputace, analýza a slučování výsledků), což v praxi mnohé výzkumníky odrazuje. Nastavování procedury (její první fázi), výpočetní náročnost fáze imputace a také následný data management představují další potenciálně odrazující faktory.



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MŠMT
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

SHRNUTÍ

Chybějící hodnoty představují všudypřítomný fenomén téměř každých dat. Do jaké míry jsou chybějící pozorování problematické, záleží od mechanismu chybějících hodnot v analyzovaných datech. Mechanismy chybějících hodnot popisují vztahy mezi pozorovanými a chybějícími hodnotami v datech. Nejproblematičtějším mechanismem je mechanismus MNAR, při kterém závisí výskyt chybějících hodnot od samotných pozorování, které chybí.

Mezi základní přístupy řešení problému chybějících hodnot patří metody založené na vynechávání případů s chybějícími hodnotami z analýzy a metody založené na nahrazování chybějících hodnot. Základní metody trpí celou řadou nedostatků. Jednoznačně nejhorší z těchto jednoduchých technik je nahrazování chybějících hodnot aritmetickým průměrem (případně mediánem nebo modem). Naopak nejlepší vlastnosti vykazuje nahrazování chybějících hodnot prostřednictvím stochastické lineární regrese. Pokud je podíl chybějících pozorování v datech relativně nízký, tak má dobré vlastnosti i analýza kompletních případů (zejména při regresní analýze). Žádná z diskutovaných metod si ale nedokáže poradit s daty chybějícími podle mechanismu MNAR.



SEZNAM POUŽITÉ LITERATURY

- Buuren, S. van. (2018). *Flexible imputation of missing data* (Second edition). CRC Press, Taylor and Francis Group.
- Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.
- 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(404), 1198–1202. <https://doi.org/10.2307/2290157>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed). Wiley.
- Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (Third edition). Wiley.
- Petrúšek, I. (2015). *Analýza chybějících hodnot: Srovnání metod při zkoumání determinantů politické znalosti a příjmu*. Sociologický ústav AV ČR.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.
- Schafer, J. L. (1999). Multiple Imputation: A Primer. *Statistical methods in medical research*, 8(1), 3–15.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037//1082-989X.7.2.147>





SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

MAR Missing at Random

MCAR Missing Completely at Random

MNAR Missing not at Random



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MŠMT
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

PŘÍLOHA P I: R SKRIPT PRO REPLIKOVÁNÍ GRAFŮ A ANALÝZ

```
set.seed(68434)

# Vytvoření kardinální proměnné x (generováním 500 hodnot z normálního roz-
# dělení)
x <- rnorm(n = 500, mean = 250, sd = 20)

# Průměr proměnné x byl nastaven na 250
mean(x)

# Směrodatná odchylka proměnné y byla nastavena na 20
sd(x)

# Vytvoření kardinální proměnné y
y <- 20*x + 50 + rnorm(n = 500, mean = 0, sd = 930)

mean(y)

sd(y)

cor(x, y)

### Vytvoření datové matice data_sim (do této matice se budou nahrávat
všechny další odvozené proměnné)

data_sim <- data.frame(x, y)
```





Vytváření chybějících hodnot podle tří mechanismů chybějících hodnot

1. MCAR (Data chybí zcela náhodně)

Náhodně vymažeme hodnoty proměnné y u 100 případů

```
set.seed(36)
```

```
data_sim$MCAR <- as.logical(sample(x = c(0, 1), size = 500, replace = TRUE,  
prob = c(0.8, 0.2)))
```

```
data_sim$y_MCAR <- y
```

```
data_sim$y_MCAR <- ifelse(data_sim$MCAR == TRUE, data_sim$y_MCAR == NA,  
data_sim$y_MCAR)
```

2. MAR (Data chybí náhodně)

U proměnné y bude u případů s nadprůměrnými hodnotami u proměnné x vymazáno 100 hodnot proměnné y

```
indikator_mar <- data_sim$x > mean(data_sim$x)
```

```
set.seed(18)
```

```
data_sim$MAR <- ifelse(indikator_mar == TRUE, as.logical(sample(x = c(0, 1),  
size = 247, replace = TRUE, prob = c(0.5952, 0.4048))), FALSE)
```

```
data_sim$y_MAR <- y
```

```
data_sim$y_MAR <- ifelse(data_sim$MAR == TRUE, data_sim$y_MAR == NA,  
data_sim$y_MAR)
```





3. MNAR (Data nechybí náhodně)

U proměnné y bude u případů s hodnotami vyššími jako medián této proměnné vymazáno 100 hodnot proměnné y

```
indikator_mnar <- data_sim$y > median(data_sim$y)
```

```
set.seed(9)
```

```
data_sim$MNAR <- ifelse(indikator_mnar == TRUE, as.logical(sample(x = c(0, 1),  
size = 250, replace = TRUE, prob = c(0.6, 0.4))), FALSE)
```

```
data_sim$y_MNAR <- y
```

```
data_sim$y_MNAR <- ifelse(data_sim$MNAR == TRUE, data_sim$y_MNAR == NA,  
data_sim$y_MNAR)
```





Nahrazování chybějících hodnot podle tří nahrazovacích metod u jednotlivých mechanismů

A.) Nahrazování u mechanismu MCAR

```
colors_MCAR <- ifelse(data_sim$MCAR == TRUE, yes = "salmon", no = "steel-blue1")
```

Nahrazování chybějících hodnot aritmetickým průměrem

```
data_sim$mean_MCAR <- data_sim$y_MCAR
```

```
data_sim$mean_MCAR[data_sim$MCAR == TRUE] <- mean(data_sim$y_MCAR, na.rm = TRUE)
```

Nahrazování chybějících hodnot lineární regresí

```
data_sim$reg_MCAR <- data_sim$y_MCAR
```

```
fit_reg_MCAR <- lm(y_MCAR ~ x, data = data_sim, na.action = na.omit)
```

```
data_sim$reg_MCAR[data_sim$MCAR == TRUE] <- (fit_reg_MCAR$coefficients[1] + fit_reg_MCAR$coefficients[2]*data_sim$x[data_sim$MCAR == TRUE])
```

Nahrazování chybějících hodnot stochastickou lineární regresí

```
data_sim$stoch_reg_MCAR <- data_sim$y_MCAR
```

```
fit_stoch_reg_MCAR <- lm(y_MCAR ~ x, data = data_sim, na.action = na.omit)
```

```
set.seed(17)
```

```
data_sim$stoch_reg_MCAR[data_sim$MCAR == TRUE] <-
```

```
(fit_stoch_reg_MCAR$coefficients[1] +
```

```
fit_reg_MCAR$coefficients[2]*data_sim$x[data_sim$MCAR == TRUE] +
```

```
rnorm(100, mean = 0, sd = sd(fit_reg_MCAR$residuals)))
```





B.) Nahrazování u mechanismu MAR

```
colors_MAR <- ifelse(data_sim$MAR == TRUE, yes = "salmon", no = "steelblue1")
```

Nahrazování chybějících hodnot aritmetickým průměrem

```
data_sim$mean_MAR <- data_sim$y_MAR
```

```
data_sim$mean_MAR[data_sim$MAR == TRUE] <- mean(data_sim$y_MAR, na.rm  
= TRUE)
```

Nahrazování chybějících hodnot lineární regresí

```
data_sim$reg_MAR <- data_sim$y_MAR
```

```
fit_reg_MAR <- lm(y_MAR ~ x, data = data_sim, na.action = na.omit)
```

```
data_sim$reg_MAR[data_sim$MAR == TRUE] <- (fit_reg_MAR$coefficients[1] +  
fit_reg_MAR$coefficients[2]*data_sim$x[data_sim$MAR == TRUE])
```

Nahrazování chybějících hodnot stochastickou lineární regresí

```
data_sim$stoch_reg_MAR <- data_sim$y_MAR
```

```
fit_stoch_reg_MAR <- lm(y_MAR ~ x, data = data_sim, na.action = na.omit)
```

```
set.seed(2)
```

```
data_sim$stoch_reg_MAR[data_sim$MAR == TRUE] <-  
(fit_stoch_reg_MAR$coefficients[1] +  
fit_reg_MAR$coefficients[2]*data_sim$x[data_sim$MAR == TRUE] + rnorm(100,  
mean = 0, sd = sd(fit_reg_MAR$residuals)))
```





C.) Nahrazování u mechanismu MNAR

```
colors_MNAR <- ifelse(data_sim$MNAR == TRUE, yes = "salmon", no = "steel-blue1")
```

Nahrazování chybějících hodnot aritmetickým průměrem

```
data_sim$mean_MNAR <- data_sim$y_MNAR
```

```
data_sim$mean_MNAR[data_sim$MNAR == TRUE] <- mean(data_sim$y_MNAR,  
na.rm = TRUE)
```

Nahrazování chybějících hodnot lineární regresí

```
data_sim$reg_MNAR <- data_sim$y_MNAR
```

```
fit_reg_MNAR <- lm(y_MNAR ~ x, data = data_sim, na.action = na.omit)
```

```
data_sim$reg_MNAR[data_sim$MNAR == TRUE] <- (fit_reg_MNAR$coefficients[1]  
+ fit_reg_MNAR$coefficients[2]*data_sim$x[data_sim$MNAR == TRUE])
```

Nahrazování chybějících hodnot stochastickou lineární regresí

```
data_sim$stoch_reg_MNAR <- data_sim$y_MNAR
```

```
fit_stoch_reg_MNAR <- lm(y_MNAR ~ x, data = data_sim, na.action = na.omit)
```

```
set.seed(2)
```

```
data_sim$stoch_reg_MNAR[data_sim$MNAR == TRUE] <-  
(fit_stoch_reg_MNAR$coefficients[1] +  
fit_reg_MNAR$coefficients[2]*data_sim$x[data_sim$MNAR == TRUE] +  
rnorm(100, mean = 0, sd = sd(fit_reg_MNAR$residuals)))
```





Konstrukce grafů

Graf 1: Marginální graf při dvou mechanismech chybějících hodnot

```
library(VIM)
```

```
par(mfrow = c(2, 1), mar = (c(2.5, 3.25, 2, 1)), las = 1)
```

```
marginplot(data_sim[, c("x", "y_MCAR")], pch = 20, xlab = "", ylab = "", main =  
"MCAR",
```

```
col = c("steelblue1", "salmon", "red4", "orange", "orange4"), cex.numbers  
= 0.7)
```

```
marginplot(data_sim[, c("x", "y_MAR")], pch = 20, xlab = "", ylab = "", main =  
"MAR",
```

```
col = c("steelblue1", "salmon", "red4", "orange", "orange4"), cex.numbers  
= 0.7)
```

Graf 2. Kompletní simulovaná data

```
par(mfrow = c(1, 1), mar = (c(4, 4, 2, 1)), las = 1)
```

```
plot(data_sim$x, data_sim$y, xlim = c(175, 325), ylim = c(0, 8000), xaxp =  
c(180, 320, 7), xlab = "x", ylab = "y", main = bquote("Kompletní data: r" ==  
.(round(cor(data_sim$y, data_sim$x), 2))), col = "steelblue1", pch = 19)
```

```
text(x = 182, y = 4500, bquote(bar(y) == .(round(mean(data_sim$y)))), cex = 1.1,  
col = "black")
```

```
text(x = 182, y = 3500, bquote(s[y] == .(round(sd(data_sim$y)))), cex = 1.1, col =  
"black")
```

```
text(x = 240, y = 250, bquote(bar(x) == .(round(mean(data_sim$x)))), cex = 1.1,  
col = "black")
```

```
text(x = 260, y = 250, bquote(s[x] == .(round(sd(data_sim$x)))), cex = 1.1, col =  
"black")
```





Graf 3. Kompletní data a chybějící pozorování podle tří mechanismů chybějících hodnot

```
par(mfrow = c(2, 2), mar = (c(2.5, 3, 2, 0.75)), las = 1)
```

```
plot(data_sim$x, data_sim$y, xlim = c(175, 325), ylim = c(0, 8000), xaxp =  
c(180, 320, 7), xlab = "", ylab = "", main = "Kompletní data", col = "steelblue1",  
pch = 20)
```

```
plot(data_sim$x, data_sim$y_MCAR, xlim = c(175, 325), ylim = c(0, 8000), xaxp =  
c(180, 320, 7), xlab = "", ylab = "", main = "MCAR: Data chybí zcela náhodně",  
col = "steelblue1", pch = 20)
```

```
points(data_sim$x[data_sim$MCAR == TRUE], data_sim$y[data_sim$MCAR ==  
TRUE], col = "salmon", pch = 1)
```

```
plot(data_sim$x, data_sim$y_MAR, xlim = c(175, 325), ylim = c(0, 8000), xaxp =  
c(180, 320, 7), xlab = "", ylab = "", main = "MAR: Data chybí náhodně", col = "ste-  
elblue1", pch = 20)
```

```
points(data_sim$x[data_sim$MAR == TRUE], data_sim$y[data_sim$MAR ==  
TRUE], col = "salmon", pch = 1)
```

```
plot(data_sim$x, data_sim$y_MNAR, xlim = c(175, 325), ylim = c(0, 8000), xaxp =  
c(180, 320, 7), xlab = "", ylab = "", main = "MNAR: Data nechybí náhodně", col =  
"steelblue1", pch = 20)
```

```
points(data_sim$x[data_sim$MNAR == TRUE], data_sim$y[data_sim$MNAR ==  
TRUE], col = "salmon", pch = 1)
```





Graf 4. Fungování metod při mechanismu MCAR

```
par(mfrow = c(2, 2), mar = (c(2.5, 3, 2, 0.75)), las = 1)
```

```
plot(data_sim$x, data_sim$y_MCAR, xlim = c(175, 325), ylim = c(0, 8000), col =  
colors_MCAR, xlab = "", ylab = "", main = bquote("Listwise deletion: r" ==  
.(round(cor(data_sim$y, data_sim$x), 2))), pch = 20)
```

```
text(x = 182, y = 4500, bquote(bar(y) == .(round(mean(data_sim$y_MCAR, na.rm  
= TRUE)))), cex = 1.1, col = "black")
```

```
text(x = 182, y = 3500, bquote(s[y] == .(round(sd(data_sim$y_MCAR, na.rm =  
TRUE)))), cex = 1.1, col = "black")
```

```
text(x = 235, y = 250, bquote(bar(x) ==  
.(round(mean(data_sim$x[data_sim$MCAR == FALSE])))), cex = 1.1, col = "black")
```

```
text(x = 265, y = 250, bquote(s[x] == .(round(sd(data_sim$x[data_sim$MCAR ==  
FALSE])))), cex = 1.1, col = "black")
```

```
plot(data_sim$x, data_sim$mean_MCAR, xlim = c(175, 325), ylim = c(0, 8000),  
col = colors_MCAR, xlab = "", ylab = "", main = bquote("Aritmetický průměr: r" ==  
.(round(cor(data_sim$mean_MCAR, data_sim$x), 2))), pch = 20)
```

```
text(x = 183, y = 4500, bquote(bar(y) ==  
.(round(mean(data_sim$mean_MCAR)))), cex = 1.1, col = "black")
```

```
text(x = 183, y = 3500, bquote(s[y] == .(round(sd(data_sim$mean_MCAR)))), cex  
= 1.1, col = "orange")
```

```
plot(data_sim$x, data_sim$reg_MCAR, xlim = c(175, 325), ylim = c(0, 8000), col  
= colors_MCAR, xlab = "", ylab = "", main = bquote("Lineární regrese: r" ==  
.(round(cor(data_sim$reg_MCAR, data_sim$x), 2))), pch = 20)
```





```
text(x = 182, y = 4500, bquote(bar(y) == .(round(mean(data_sim$reg_MCAR)))),  
cex = 1.1, col = "black")
```

```
text(x = 182, y = 3500, bquote(s[y] == .(round(sd(data_sim$reg_MCAR)))), cex =  
1.1, col = "orange")
```

```
plot(data_sim$x, data_sim$stoch_reg_MCAR, xlim = c(175, 325), ylim = c(0,  
8000), col = colors_MCAR, xlab = "", ylab = "", main = bquote("Stochastická line-  
ární regrese: r" == .(round(cor(data_sim$stoch_reg_MCAR, data_sim$x), 2))),  
pch = 20)
```

```
text(x = 183, y = 4500, bquote(bar(y) ==  
.(round(mean(data_sim$stoch_reg_MCAR)))), cex = 1.1, col = "black")
```

```
text(x = 183, y = 3500, bquote(s[y] == .(round(sd(data_sim$stoch_reg_MCAR)))),  
cex = 1.1, col = "black")
```

Graf 5. Fungování metod při mechanismu MAR

```
par(mfrow = c(2, 2), mar = (c(2.5, 3, 2, 0.75)), las = 1)
```

```
plot(data_sim$x, data_sim$y_MAR, xlim = c(175, 325), ylim = c(0, 8000), col =  
colors_MAR, xlab = "", ylab = "", main = bquote("Listwise deletion: r" ==  
.(round(cor(data_sim$y_MAR, data_sim$x, use = "complete.obs"), 2))), pch = 20)
```

```
text(x = 182, y = 4500, bquote(bar(y) == .(round(mean(data_sim$y_MAR, na.rm  
= TRUE)))), cex = 1.1, col = "black")
```

```
text(x = 182, y = 3500, bquote(s[y] == .(round(sd(data_sim$y_MAR, na.rm =  
TRUE)))), cex = 1.1, col = "black")
```

```
text(x = 235, y = 250, bquote(bar(x) == .(round(mean(data_sim$x[data_sim$MAR  
== FALSE])))), cex = 1.1, col = "black")
```





```
text(x = 265, y = 250, bquote(s[x] == .(round(sd(data_sim$x[data_sim$MAR == FALSE])))), cex = 1.1, col = "black")
```

```
plot(data_sim$x, data_sim$mean_MAR, xlim = c(175, 325), ylim = c(0, 8000), col = colors_MAR, xlab = "", ylab = "", main = bquote("Aritmetický průměr: r" == .(round(cor(data_sim$mean_MAR, data_sim$x), 2))), pch = 20)
```

```
text(x = 183, y = 4500, bquote(bar(y) == .(round(mean(data_sim$mean_MAR)))), cex = 1.1, col = "black")
```

```
text(x = 183, y = 3500, bquote(s[y] == .(round(sd(data_sim$mean_MAR)))), cex = 1.1, col = "orange")
```

```
plot(data_sim$x, data_sim$reg_MAR, xlim = c(175, 325), ylim = c(0, 8000), col = colors_MAR, xlab = "", ylab = "", main = bquote("Lineární regrese: r" == .(round(cor(data_sim$reg_MAR, data_sim$x), 2))), pch = 20)
```

```
text(x = 182, y = 4500, bquote(bar(y) == .(round(mean(data_sim$reg_MAR)))), cex = 1.1, col = "black")
```

```
text(x = 182, y = 3500, bquote(s[y] == .(round(sd(data_sim$reg_MAR)))), cex = 1.1, col = "orange")
```

```
plot(data_sim$x, data_sim$stoch_reg_MAR, xlim = c(175, 325), ylim = c(0, 8000), col = colors_MAR, xlab = "", ylab = "", main = bquote("Stochastická lineární regrese: r" == .(round(cor(data_sim$stoch_reg_MAR, data_sim$x), 2))), pch = 20)
```

```
text(x = 183, y = 4500, bquote(bar(y) == .(round(mean(data_sim$stoch_reg_MAR)))), cex = 1.1, col = "black")
```

```
text(x = 183, y = 3500, bquote(s[y] == .(round(sd(data_sim$stoch_reg_MAR)))), cex = 1.1, col = "black")
```





Graf 6. Fungování metod při mechanismu MNAR

```
par(mfrow = c(2, 2), mar = (c(2.5, 3, 2, 0.75)), las = 1)
```

```
plot(data_sim$x, data_sim$y_MNAR, xlim = c(175, 325), ylim = c(0, 8000), col =  
colors_MNAR, xlab = "", ylab = "", main = bquote("Listwise deletion: r" ==  
.(round(cor(data_sim$y_MNAR, data_sim$x, use = "complete.obs"), 2))), pch =  
20)
```

```
text(x = 182, y = 4500, bquote(bar(y) == .(round(mean(data_sim$y_MNAR, na.rm  
= TRUE)))), cex = 1.1, col = "orange")
```

```
text(x = 182, y = 3500, bquote(s[y] == .(round(sd(data_sim$y_MNAR, na.rm =  
TRUE)))), cex = 1.1, col = "black")
```

```
text(x = 235, y = 250, bquote(bar(x) ==  
.(round(mean(data_sim$x[data_sim$MNAR == FALSE])))), cex = 1.1, col = "black")
```

```
text(x = 265, y = 250, bquote(s[x] == .(round(sd(data_sim$x[data_sim$MNAR ==  
FALSE])))), cex = 1.1, col = "black")
```

```
plot(data_sim$x, data_sim$mean_MNAR, xlim = c(175, 325), ylim = c(0, 8000),  
col = colors_MNAR, xlab = "x", ylab = "y", main = bquote("Aritmetický průměr: r"  
== .(round(cor(data_sim$mean_MNAR, data_sim$x), 2))), pch = 20)
```

```
text(x = 183, y = 4500, bquote(bar(y) ==  
.(round(mean(data_sim$mean_MNAR)))), cex = 1.1, col = "orange")
```

```
text(x = 183, y = 3500, bquote(s[y] == .(round(sd(data_sim$mean_MNAR)))), cex  
= 1.1, col = "orange")
```

```
plot(data_sim$x, data_sim$reg_MNAR, xlim = c(175, 325), ylim = c(0, 8000), col  
= colors_MNAR, xlab = "x", ylab = "y", main = bquote("Lineární regrese: r" ==  
.(round(cor(data_sim$reg_MNAR, data_sim$x), 2))), pch = 20)
```





```
text(x = 182, y = 4500, bquote(bar(y) == .(round(mean(data_sim$reg_MNAR)))),  
cex = 1.1, col = "orange")
```

```
text(x = 182, y = 3500, bquote(s[y] == .(round(sd(data_sim$reg_MNAR)))), cex =  
1.1, col = "orange")
```

```
plot(data_sim$x, data_sim$stoch_reg_MNAR, xlim = c(175, 325), ylim = c(0,  
8000), col = colors_MNAR, xlab = "x", ylab = "y", main = bquote("Stochastická li-  
neární regrese: r" == .(round(cor(data_sim$stoch_reg_MNAR, data_sim$x), 2))),  
pch = 20)
```

```
text(x = 183, y = 4500, bquote(bar(y) ==  
.(round(mean(data_sim$stoch_reg_MNAR)))), cex = 1.1, col = "orange")
```

```
text(x = 183, y = 3500, bquote(s[y] == .(round(sd(data_sim$stoch_reg_MNAR)))),  
cex = 1.1, col = "black")
```

