APLIKACE IMPUTAČNÍCH METOD NA DATECH Z ŠETŘENÍ SHARE

Vasilii Marushev

Artem Vitkov

ABSTRAKT

Tato práce se zabývá aplikací imputačních metod na chybějící hodnoty v celoevropském výběrovém šetření SHARE (Survey of Health, Ageing and Retirement in Europe). Zaměřuje se na ošetření non-response pozorování pomocí modelových technik, jako regresní imputace, stochastická imputace a vícenásobná imputace a na aplikaci pokročilých technik, jako vícerozměrných statistických metod. V analýze rozlišujeme mezi typy chybějících dat: MCAR (Missing Completely at Random), MAR (Missing at Random) a MNAR (Missing Not at Random). V práci je provedena komparace pokročilých imputačních metod s tradičními technikami a je posuzována jejich efektivita při nahrazování chybějících hodnot v kontextu sociálně-ekonomického šetření. Výsledky studie poskytují srovnání účinnosti jednotlivých imputačních metod a jejich vlivu na kvalitu dat a následné analýzy. Na základě získaných poznatků je navržen rozhodovací rámec pro výběr optimální imputační metody pro datový soubor easySHARE.

JEL KÓDY

C81, C83, C14

KLÍČOVÁ SLOVA

Imputace, easySHARE, MICE, kNN, MAR, MNAR, MCAR, chybějící hodnoty

ÚVOD

V posledních desetiletích dochází k dynamickému nárůstu dostupných dat, a to nejen v oblasti informačních technologií, ale také v sociálních a ekonomických výzkumech. Jedním z nejrozsáhlejších zdrojů informací o stárnoucí populaci v Evropě je databáze SHARE (Survey of Health, Ageing and Retirement in Europe). Přestože tato databáze poskytuje cenná data o zdravotním stavu, socioekonomických charakteristikách a rodinném zázemí respondentů, stejně jako u mnoha jiných výběrových šetření se i zde setkáváme s problémem **chybějících dat**. Pro následné analýzy je proto nezbytné tyto chybějící hodnoty vhodným způsobem ošetřit.

Imputace, tedy doplnění chybějících hodnot na základě pozorovaných, tvoří klíčový krok v procesu datového zpracování. Rozsah a mechanismus chybějících hodnot (MCAR, MAR, MNAR) přitom významně ovlivňují přesnost a vypovídací schopnost výstupů.

Zatímco tradiční imputační techniky, jako je nahrazení průměrem, mediánem či modem, jsou relativně jednoduché, mohou významně zkreslovat rozdělení dat a vést ke koncentraci hodnot kolem zvoleného parametru a podhodnocení skutečné variability. Použití složitějších metod, například regresní imputace, vícenásobné imputace nebo metod vícerozměrné statistiky, dokáže lépe zachytit vztahy mezi proměnnými a do jisté míry se vypořádat s nejistotou spojenou s náhradou chybějících hodnot.

Cílem naší analýzy je imputace hodnot proměnné ‘thinc_m’ (čisté roční příjmy domácnosti v eurech) pomocí regresní a vícenásobné imputace MICE a pokus o imputaci hodnot pomocí metody k-NN.

EXPLORATORNÍ ANALÝZA DAT

Nejdřív provedeme exploratorní analýzu datového souboru, který obsahoval data za sedmou vlnu celoevropského šetření SHARE. Vybereme pro názornost data jenom za Českou republiku za rok 2017. Důvodem pro výběr tohoto období byl právě fakt, že příjmy domácností v sedmé vlně nebyly imputovány a zároveň toto období obsahovalo nejaktuálnější a nejkompletnější záznam možných nechybějících vysvětlujících proměnných. Jelikož původní datový soubor easySHARE (Easy-SHARE, 2024) celkem obsahoval 108 proměnných různého typu, rozhodli jsme se, kromě proměnné ‘thinc_m’, vybrat dalších 14 relevantních proměnných na základě co nejmenšího podílu chybějících hodnot a přiměřené síly statistické závislosti mezi vysvětlujícími proměnnými a ‘thinc_m’ podle vlastního uvažování autorů.

Seznam analyzovaných proměnných je následovný:

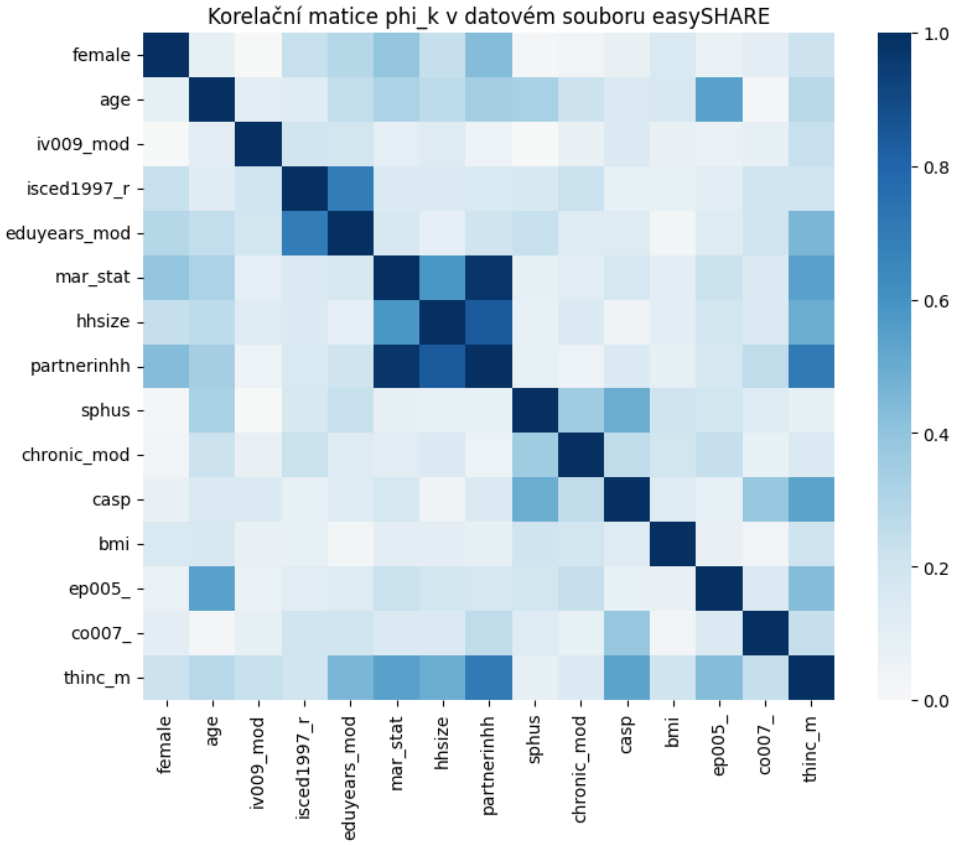
1. 'female' (binární) – pohlaví respondenta
2. 'age' (kvantitativní) – věk při pohovoru (v letech)
3. ‘iv009_mod‘ (nominální) – typ obce podle velikosti
4. ‘isced1997_r’ (ordinální) – úroveň nejvyššího dosaženého vzdělání podle ISCED-97
5. ‘eduyears_mod’ (kvantitativní) – počet let vzdělání
6. ‘mar_stat’ (nominální) – rodinný stav
7. ‘hhsize’ (kvantitativní) – velikost domácnosti

8. ‘partnerinhh’ (binární) – bydlí-li respondent společně s manželem/partnerem
9. ‘sphus’ (ordinální) – vnímání vlastního zdravotního stavu
10. ‘chronic_mod’ (kvantitativní) – počet chronických onemocnění respondenta
11. ‘casp’ (kvantitativní) – index kvality života a blahobytu
12. ‘bmi’ (kvantitativní) – index tělesné hmotnosti respondenta
13. ‘ep005_‘ (nominální) – současná pracovní situace
14. ‘co007_‘ (ordinální) – je domácnost schopna vyjít s penězi
15. ‘thinc_m’ (kvantitativní) – čisté roční příjmy domácnosti v eurech

Provedeme totiž celou analýzu pomocí programovacího jazyka Python. Jelikož soubor obsahuje chybějící hodnoty zakódované různými značeními (malé detaily, jako např. “-12” znamená, že respondent neví/nechce uvádět údaje, kdyžto “-15” znamená absence informace v konkrétním šetření). Pro jednoduchost práce jsme sjednotili této chybějící hodnoty do společného typu NaN.

Podívejme se na statistické vztahy mezi jednotlivými proměnnými v našem datovém souboru bez uvažování o chybějících hodnotách (tzn. pro výpočet míry závislosti analyzujeme jen ty pozorování, kde jsou všechny proměnné uvedeny). Jelikož máme datový soubor, který je zastoupen proměnnými různého typu, není na místě použít klasický Pearsonův korelační koeficient ρ. Použijeme proto jako alternativu pro měření vztahů mezi proměnnými nově vyvinutou míru závislosti **phi_k (*ϕk*)**. Tento koeficient je užitečný pro naši analýzu, jelikož umožnuje měření závislosti mezi proměnnými různého typu, například mezi kvantitativními a nominálními proměnnými.

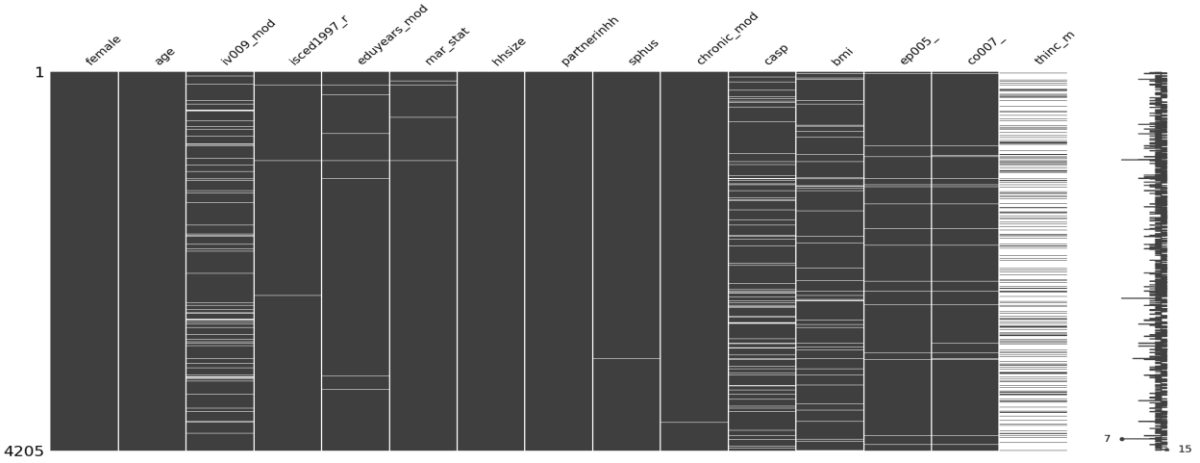
Tato míra závislosti je založena na algoritmickém přístupu, který využívá $\chi 2$ statistiku, přičemž po binování kvantitativních proměnných se použijí pouze četnosti v jednotlivých kategoriích, čímž ***ϕk*** je považuje za kategoriální proměnné. Pro ***ϕk*** neexistuje žádný analytický vzorec, jeho interpretace ale spočívá ve vyhodnocení transformované $\chi 2$ statistiky mezi dvěma binovanými proměnnými jako hodnoty pocházející ze dvourozměrného normálního rozdělení. Jedná se o modifikaci Cramérova ϕ pro případ více kategorií (pro podrobnější definici, prosím vizte (M. Baak, 2019, pp. 9-10).



Obr. 1 Korelační matice phi_k v datovém souboru easySHARE

Pozorujeme, že mezi všemi proměnnými v datovém souboru a ‘thinc_m’ existuje statisticky významný vztah. Tuto skutečnost pak můžeme využít u imputačních metod založených na regresi.

Vizualizujeme chybějící hodnoty pomocí matice (chybějící hodnoty jsou označeny jako bílé):



Obr. 2 Matice chybějících hodnot

Můžeme vidět, že se jedná o poměrné neřídkou datovou matici, uvedeme zde tabulku procentuálního zastoupení chybějících hodnot v jednotlivých proměnných:

thinc_m	78.22
casp	8.04
iv009_mod	7.63
bmi	4.78
co007_	1.88
ep005_	1.64
eduyears_mod	1.43
mar_stat	0.40
isced1997_r	0.36
chronic_mod	0.31
sphus	0.19
female	0.00
age	0.00
hhsize	0.00
partnerinhh	0.00

Obr. 3 Procentuální podíl chybějících hodnoty v jednotlivých proměnných

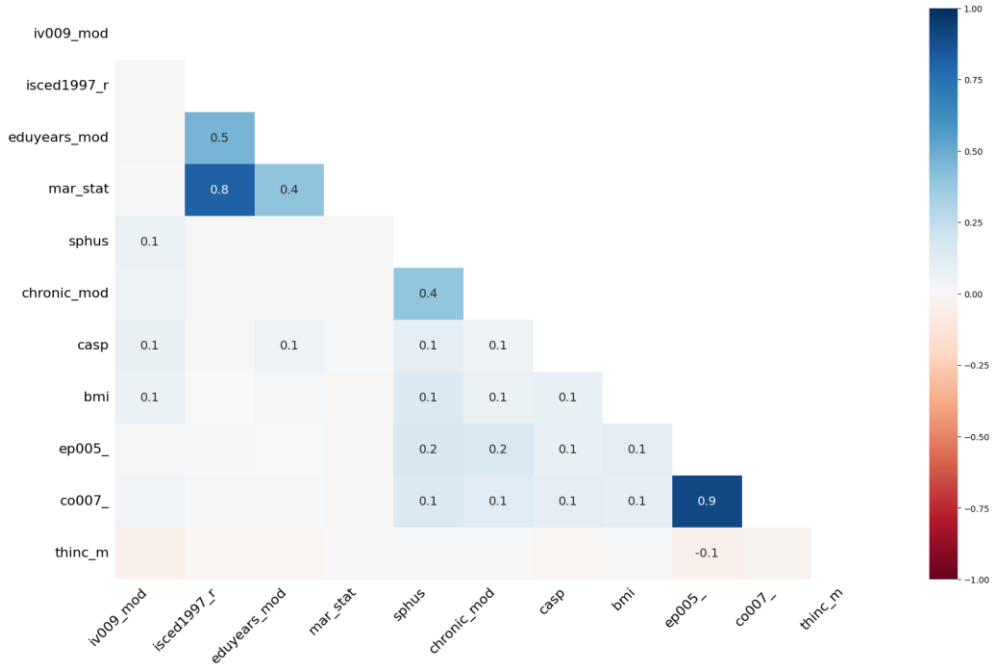
Přičemž, jenom 700 respondentů z 4205 má všechny proměnné vyplněné. Proto by metoda ‚listwise deletion' (metoda, kdy celý záznam je vyloučen z analýzy, pokud chybí hodnota jakékoliv proměnné) by byla velmi nevhodná k použití, jelikož bychom ztratili 83,35% pozorování v datovém souboru. Tím pádem aplikace různých imputačních metod v případě datového souboru easySHARE je velice opodstatněna.

Dále se podívejme na korelaci chybějících hodnot u jednotlivých kombinací proměnných.
Měří jak silně přítomnost nebo nepřítomnost jedné proměnné ovlivňuje přítomnost druhé.
Korelace chybování se pohybuje v rozmezí od -1 do 1, popíšeme hraniční případy:

1 – když je hodnota jedné proměnná přítomna, pak hodnota druhé proměnné chybí **vždy pro každou odpovídající dvojici proměnných**

0 – Přítomné nebo nepřítomné hodnoty proměnných na sebe nemají **žádný vliv**

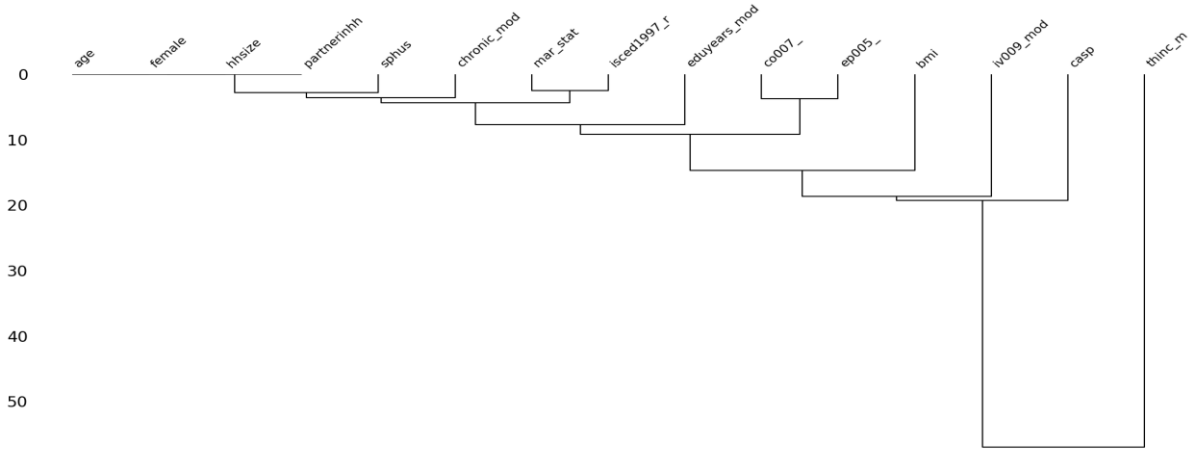
1 - pokud je hodnota jedné proměnná přítomna, pak je hodnota druhé proměnné **vždy pro každou odpovídající dvojici proměnných**



Obr. 4 Korelační matice „chybování“ mezi jednotlivými proměnnými

Můžeme pozorovat, že není patrný žádný vzor mezi chybováním ‘thinc_m’ a ostatními proměnnými v datovém souboru (prázdná políčka znamenají totiž statisticky nevýznamnou korelaci měřenou Cramérovým ϕ).

Dendrogram vystihuje širší pohled na chybějící hodnoty než je viditelný v párové mapě korelace. Využívá totiž algoritmus hierarchického shlukování k tomu, aby proměnné navzájem rozdělil do skupin podle jejich nulové korelace (měřené pomocí binární vzdálenosti).



Obr. 5 Dendrogram podobnosti „chybování“ mezi jednotlivými proměnnými

Můžeme pozorovat, že proměnná ‘thinc_m’ je zastoupena ve zcela disjunktní skupině, což odpovídá závěru z předchozího grafu.

TYPY CHYBĚJÍCÍCH POZOROVÁNÍ

První klasifikaci chybějících údajů zavedli Little & Rubin v roce 1987 (Roderick J., 1986).

Rozdělili mechanismy chybějících údajů do tří kategorií:

MCAR: zcela náhodně chybějící údaje

MAR: náhodně chybějící

MNAR: chybějící nikoliv náhodně

MCAR

Úplně náhodně chybějící údaje (MCAR): V tomto případě chybějící pozorování nemají žádný vztah k jiným pozrovnáním nebo proměnným v souboru údajů. Zcela náhodně chybějící data znamenají, že pravděpodobnost chybění jedné nezávislé proměnné nezávisí na ostatních sledovaných nezávislých proměnných. Jinými slovy, pravděpodobnost chybění nezávislé proměnné závisí pouze na některých vnějších faktorech. Například: náhodný účastník neuvedl svůj věk.

K odhalení vzoru MCAR můžeme použít statistické testy. Jedním ze známých statistických testů pro tento účel je $\chi 2$ test nezávislosti. Test začíná stanovením nulové hypotézy a vypočítá míru blízkosti mezi pozorovanými četnostmi a očekávanými četnostmi za platnosti nulové hypotézy.
Nulová a alternativní hypotézy jsou v tomto případě následující:

H0: Neexistuje žádná závislost mezi chybějícími údaji v konkrétním sloupci a ostatními sloupci.
H1: Existuje závislost mezi chybějícími údaji v alespoň jednom ze sloupců a ostatními sloupci.

Pro účely této analýzy použijeme modifikovanou verzi $\chi 2$ testu: Littleův MCAR test.
Zvolíme hladinu 5% významnosti. Pomocí vlastní implementace algoritmu (vizte část „Příloha“) ze článku (Li, 2013) v Pythonu provedeme tento statistický test.

```
✓ def little_mcar_test(df, alpha=0.05): ···

Reject null hypothesis: Data is not MCAR (p-value=0.0000, chi-square=286.3277)
```

Obr. 6 Výsledek Littleova MCAR testu

Na základě vypočítané p-hodnoty, která je rovna numerické nule, na jakékoliv rozumné hladině významnosti zamítáme nulovou hypotézu o nezávislosti chybovosti hodnot.
Máme dostatek důkazů tvrdit, že data v našem souboru n**ejsou MCAR**.
Dále bychom měli statisticky otestovat datový soubor na další typy NA. Bohužel ale neexistuje jediné univerzální a jednoznačné kritérium které by rozhodlo, jestli jsou data buď MAR nebo MNAR. Přesto lze nalézt postupy a testy, kterými se mechanismus chybějících dat dá **přibližně** posoudit.

MAR

Náhodně chybějící data (MAR): Chybějící data souvisejí s jinými pozorovanými proměnnými. Například pokud mladší lidé mají tendenci vynechávat otázku o svém příjmu, pak věk souvisí s chybějícími údaji o příjmu.

Toto v reálním světě často znamená, že např. starší nebo movitější lidé častěji odmítají odpovědět na dotaz o výši příjmů. Pak je chybění příjmů podmíněno věkem a bohatstvím (což je v datovém souboru pozorované), ale není přímo podmíněno „skutečnou" výší příjmů (nepozorovaný faktor).

Ověříme tento předpoklad pomocí modelu logistické regrese typu: **"chybí/nechybí hodnota proměnné 'thinc_m' ~ X"**. Zde pro každou proměnnou s chybějícími hodnotami vytvoříme binární indikátor typu (chybí = 1, nechybí = 0) a zkusíme předpovědět chybění z ostatních pozorovaných proměnných:

- Pokud chybění lze vysvětlit pozorovanými proměnnými (tj. je tam statisticky významný vztah), tak *přinejmenším* víme, že to není MCAR.
- Pokud se chybějící hodnoty jeví závislé jen na pozorovaných proměnných, je přiměřené (byť ne 100% jisté) předpokládat MAR.

Hlavní problém spočívá v tom, že MAR v datech nemůžeme formálně prokázat nebo zamítnout. Fakt, že nám logistický model dobře vysvětlí binární indikátor pomocí pozorovaných proměnných, je náznak MAR, ale pořád existuje šance, že se ve hře uplatňují i **nepozorované** faktory, a data mohou být MNAR.

Provedeme netrénování modelu v Pythonu pomocí jen vybraných proměnných. Nebude nás zajímat hodnota Pseudo-R^2 statistiky nebo ROC-křivka. Spíš se zaměříme na statistickou významnost jednotlivých regresních koeficientů.

Logit Regression Results

Dep. Variable:	missing_thinc_m	No. Observations:	4205
Model:	Logit	Df Residuals:	4190
Method:	MLE	Df Model:	14
Date:	Tue, 14 Jan 2025	Pseudo R-squ.:	0.03163
Time:	04:11:05	Log-Likelihood:	-2134.3
converged:	True	LL-Null:	-2204.1
Covariance Type:	nonrobust	LLR p-value:	9.143e-23

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.8268	0.660	4.281	0.000	1.533	4.121
female	-0.0676	0.083	-0.810	0.418	-0.231	0.096
age	-0.0453	0.005	-9.058	0.000	-0.055	-0.035
iv009_mod	-0.0948	0.031	-3.025	0.002	-0.156	-0.033
isced1997_r	-0.0052	0.004	-1.179	0.239	-0.014	0.003
eduyears_mod	0.0287	0.013	2.202	0.028	0.003	0.054
mar_stat	0.1376	0.037	3.716	0.000	0.065	0.210
hhsize	0.1403	0.055	2.536	0.011	0.032	0.249
partnerinhh	-0.1326	0.090	-1.471	0.141	-0.309	0.044
sphus	0.2187	0.054	4.071	0.000	0.113	0.324
chronic_mod	0.0108	0.033	0.327	0.743	-0.054	0.075
casp	0.0113	0.009	1.301	0.193	-0.006	0.028
bmi	0.0057	0.009	0.643	0.520	-0.012	0.023
ep005_	0.0002	0.006	0.028	0.977	-0.011	0.011
co007_	0.0023	0.047	0.048	0.962	-0.091	0.095

Obr. 7 Výstup logistické regrese

Vidíme, zde, že pomocí testu o věrohodnostním poměru jsme prokázali významnost modelu oproti modelu pouze s konstantním členem. Kromě toho, opravdu vidíme, že regresní koeficienty u poloviny prediktorů jsou statisticky významné, což by mohlo svědčit o tom, že data jsou přinejmenším MAR a že prediktory, které vyšli statisticky významnými, jsou **pozorované faktory** přinášející náhodné „chybování“.

MNAR

Data chybějící nenáhodně (MNAR): Zde nastává další úroveň komplikace tím, že chybění závisí na nepozorovaných údajích. Například účastníci mohou odmítnout uvést svůj příjem, pokud patří do skupiny s vysokými příjmy.

Data jsou MNAR, pokud pravděpodobnost chybění **závisí** na **nepozorovaných** hodnotách samotných. V takovém případě žádné *externí* proměnné (které v datovém souboru máme) nedokážou úplně vysvětlit mechanismus chybění.
Obvykle pro odhalení MNAR se používají **pokročilejší modely**:

1. Heckmanův výběrový model – typicky v ekonomii pro situace, kdy chybí data o příjmech právě pro ty respondenty, kteří nejvíc vydělávají („samoselekce“).
2. „Pattern mixture“ modely – rozdělují data podle “vzoru chybění” a modelují každou skupinu zvlášť.
3. Analýza citlivosti – simulace, kolik by se změnily výsledky, kdyby chybějící data byla systematicky vyšší/nižší než odhad podle MAR.

MNAR v datech lze dokázat formálně pomocí statistických nástrojů jen obtížně, jelikož vyžaduje informaci o něčem *nepozorovaném*. Zkusíme ale provést analýzu citlivosti: simulaci, kolik by se změnily výsledky, kdyby chybějící data byla systematicky vyšší/nižší než odhad podle MAR.

U MAR se běžně předpokládá, že umělé schování části hodnot a jejich imputace (plus validace) je dostatečně robustní metoda, protože replikujeme mechanismus, který se zakládá na ostatních pozorovaných proměnných. U MNAR, na druhou stranu, vždy existuje riziko, že skutečný mechanismus je „jiný“ – v praxi se proto u velké části studií vyskytuje, že aspoň *předpokládají* MAR a provedou citlivostní analýzu, aby si ověřily, kam až může situace zajít, je-li mechanismus chybění zkreslen vůči MAR.

Zkusíme ale provést analýzu citlivosti: simulaci, kolik by se změnily výsledky, kdyby chybějící data byla systematicky vyšší/nižší než odhad podle MAR. Pokud i výraznější (např. ±10–20 %) posun imputovaných hodnot nemění hlavní závěry, datový soubor (resp. analýza) je relativně robustní vůči MNAR a můžeme (s jistou opatrností) setrvat u MAR. Pokud se závěry prudce

mění už při malém posunu, je datový soubor citlivý na MNAR a je potřeba s interpretací zacházet opatrně. Citlivostní analýza není 100% důkaz proti MNAR, spíše testuje, nakolik by MNAR ovlivnilo výsledky. Zavedeme sadu faktorů ***shifts***, o které systematicky "posuneme" (zvýšíme nebo snížíme) pouze imputované hodnoty ‚thinc_m'. Tím budeme simulovat, že ve skutečnosti jsou chybějící hodnoty "jiné" než MAR odhad.

	shift_%	mean_thinc_m	r2	coef_age
0	-20.0	12109.607454	0.36968	-76.063567
1	-15.0	12866.457920	0.36968	-80.817540
2	-10.0	13623.308386	0.36968	-85.571513
3	-5.0	14380.158852	0.36968	-90.325486
4	0.0	15137.009318	0.36968	-95.079459
5	5.0	15893.859784	0.36968	-99.833432
6	10.0	16650.710249	0.36968	-104.587405
7	15.0	17407.560715	0.36968	-109.341378
8	20.0	18164.411181	0.36968	-114.095351

Obr. 8 Citlivostní analýza změny imputovaných hodnot ‚thinc_m'

Můžeme pozorovat, že i ±5 % odchylka imputovaných hodnot **způsobí změny ve výsledcích** (v odhadech regresních koeficientů nebo v průměru klíčové doménové hodnoty), je datový soubor **citlivý na** potenciální MNAR.

Tím pádem, stanovíme předpoklad, že **data jsou MNAR, nikoliv MAR.** Prakticky toto znamená, že lidé s **velmi vysokým příjmem** častěji odmítnou na otázku o příjmu odpovědět, zatímco průměrné příjmy reportují normálně. Drtivá většina běžných imputačních technik ale (včetně MICE) předpokládá alespoň MAR. Pokud je mechanismus chybění MNAR, mohou výsledky trpět zkreslením, jelikož data o příjmech chybějí systematicky, a pozorované hodnoty proměnných tento vzor chybění nevysvětlí. Prakticky ověříme tento předpoklad v další části analýzy.

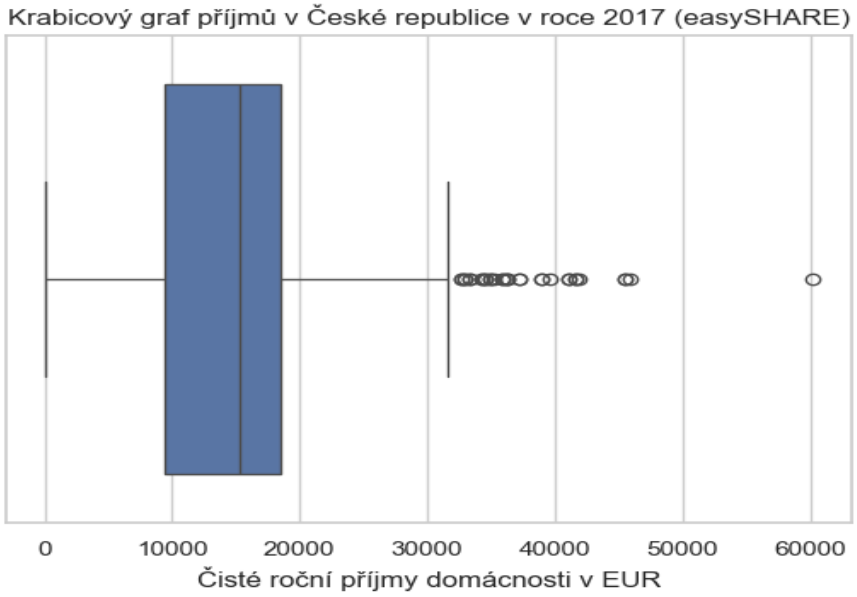
VÍCENÁSOBNÁ IMPUTACE A NÁSLEDNÝ REGRESNÍ ODHAD

Jak vyplývá ze závěrů modelu logistické regresi a citlivostní analýzy v předchozí kapitole, použijeme pro modelování non-response model MNAR. Toto značně zkomplikuje použití dalších metod, ale protože se jedná o populační mechanizmus generování NA. Avšak, nebudeme uměle generovat pozorování, protože by tento přístup nebyl odůvodnitelný kvůli těmto okolnostem:

1. Statistický: není známý data-generující proces „chybování“
2. Etický: nemůžeme uměle vytvořit data o příjmech domácností bez ohledu na socioekonomické faktory sociálního zabezpečení staršího obyvatelstva v Česku a bez informací o regionálním členění

Napříč tomuto zkusíme nahradit non-response v proměnné 'thinc_m' pomocí jednoduchého regresního odhadu. Je potřeba mít na paměti, že hlavním problémem v induktivních úsudcích pomocí regresní analýzy je homogenita analyzovaného datového souboru. Tento spíše epistemologický předpoklad je sice splněn tím, že analyzujeme šetření českých respondentů ve věku starším, než 50 let, ale zároveň sotva můžeme prokázat, že celorepublikové příjmy domácností jsou stejné.

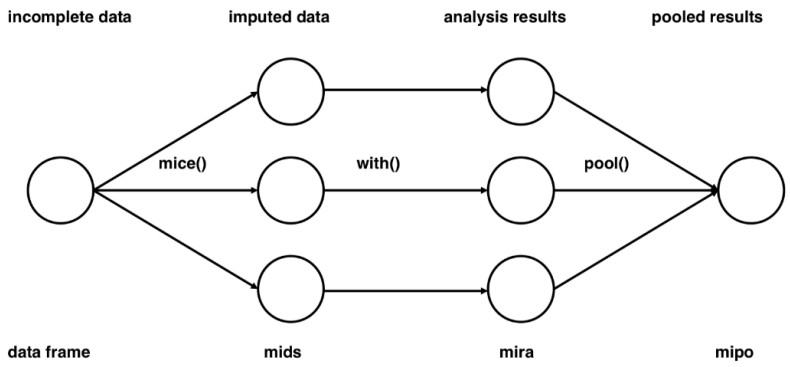
Náznakem této skutečnosti by mohl soužit krabicový graf všech příjmů, který vykazuje dostatečné množství odlehlých hodnot (pokus o logaritmování příjmů vedl k podobným výsledkům co se týká zešikmenosti rozdělení a přítomnosti odlehlých hodnot).



Obr. 9 Krabicový graf čistých ročních příjmů domácností v EUR v roce 2017

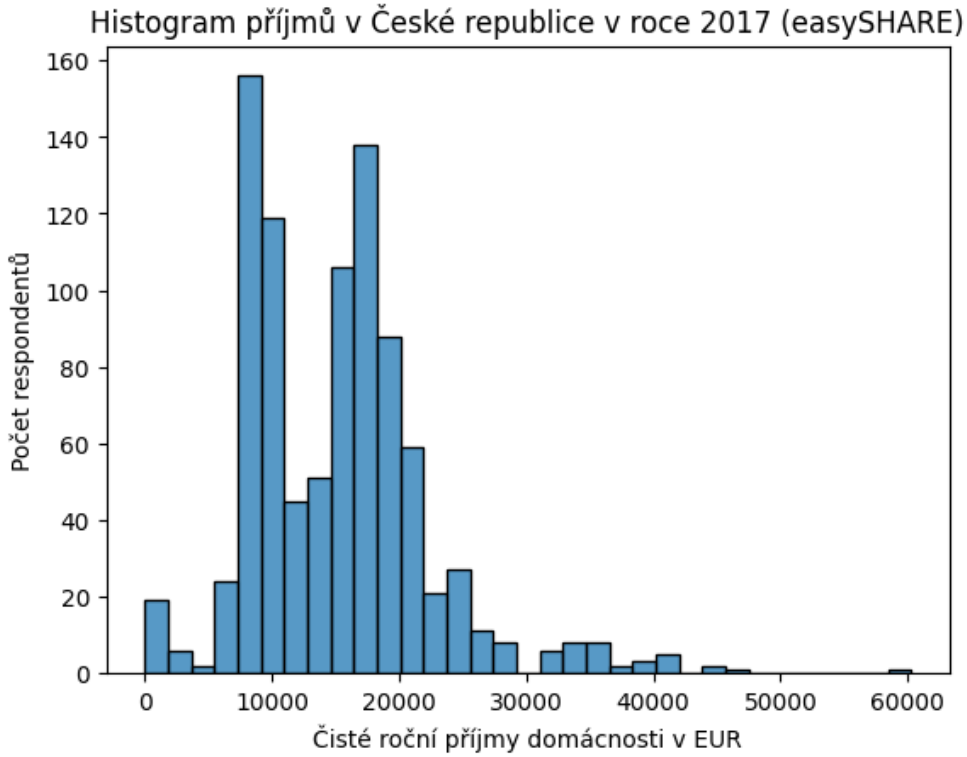
Další problém s regresní imputací by teoreticky mohl spočívat v tom, že máme řídkou datovou matici. I výběrem jenom relevantních prediktorů bychom nevyřešili tento problém, protože z matice chybějících hodnot vyplývá, že datový soubor obsahuje jen malý počet řádků, kde by zároveň byly všechny hodnoty proměnných vyplněné.

Budeme postupovat dále pomocí algoritmu MICE – *Multiple Imputation by Chained Equations*. Zdrojový článek s podrobným popisem této metody lze nalézt v (Stef van Buuren, 2011).
Vícenásobná imputace (MICE) dokáže iterativně doplňovat hodnoty tak, že se v každé iteraci imputuje jedna proměnná z ostatních (Scikit-learn, 2011). Učiníme takto právě pro všechny vysvětlující proměnné, abychom po iterativní imputaci mohli provést regresní imputaci proměnné 'thinc_m'.



Obr. 10 Princip fungování vícenásobné imputace MICE, Zdroj: https://stefvanbuuren.name/publication/2011-01-01_vanbuuren2011a/

Validace kvality imputace je obecně **nejproblematičtější část** práce s chybějícími daty. Důvodem je, že u reálných (zejména MNAR) dat neznáme *skutečné* hodnoty tam, kde je nahradíme odhadem. Abychom otestovali, zda imputační technika dobře funguje, použijeme **heuristiku**, která je známá jako **umělé maskování** (‚poškození') části hodnot, abychom mohli porovnat nahrané (imputované) hodnoty s původní realitou.
Musíme *zdůraznit*, že jelikož neznáme mechanizmus chybění dat v populaci, budeme maskovat známá data náhodnou permutací. Můžeme pozorovat na histogramu příjmů bimodální výběrové rozdělení, které **podle úvahy autorů** naznačuje chybění pozorování nejen u vysokopříjmových skupin obyvatelstva, ale i respondentů s menšími příjmy. Mimo jiné, z povahy věci příjmy domácností lze obvykle modelovat v populaci pomocí jednorozměrného logaritmicko-normálních nebo modelu konečných směsí rozdělení (Malá, 2015). Proto musíme v našem výběru zohlednit strukturu příjmů v populaci.



Obr. 11 Histogram čistých ročních příjmů domácností v EUR v roce 2017

Takto bude fungovat algoritmus této metody:

1. Vybereme část dat, kde je ‘thinc_m’ reálně známé.
2. Odhadneme parametry logaritmicko-normálního rozdělení ze stávajících hodnot proměnné 'thinc_m' pomocí metody maximální věrohodnosti. (vizte)
3. Vygenerujeme pravděpodobností výběru pro každý index na základě odhadnutého logaritmicko-normálního rozdělení a vybereme indexy na základě těchto pravděpodobností. Tedy simulujeme “chybějící” příjmy domácnosti na základě pravděpodobnostního modelu. (vlastní implementaci vizte v části „Příloha“)
4. Skryjeme ‚thinc_m‘ u určitého podílu těchto záznamů (například 20 %).
5. Provedeme odhad ‚thinc_m’ pomocí metody nejmenších čtverců a následně imputujeme
6. Porovnáváme pak imputované hodnoty vs. skutečné původní hodnoty.
7. Použijeme pak metriky pro přesnost imputace jako:
    - RMSE (střední kvadratická odchylka)
    - MAE (střední absolutní odchylka)
    - R^2 (koeficient determinace)
    - grafické porovnání pravděpodobnostních rozdělení

Po iterativní imputaci provedeme regresní odhad chybějících pozorování pomocí metody nejmenších čtverců a vyhodnotíme výsledky, které jsme obdrželi (čím lepší model, tím menší RMSE/MAE a tím větší R^2).

OLS Regression Results

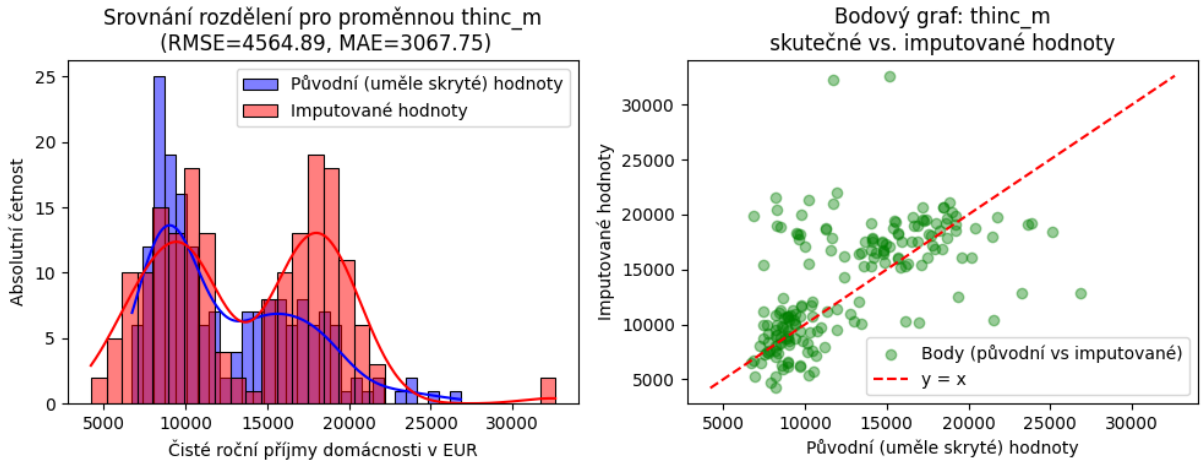
Dep. Variable:	thinc_m	R-squared:	0.370
Model:	OLS	Adj. R-squared:	0.360
Method:	Least Squares	F-statistic:	45.39
Date:	Tue, 14 Jan 2025	Prob (F-statistic):	2.37e-94
Time:	04:12:39	Log-Likelihood:	-9223.1
No. Observations:	916	AIC:	1.848e+04
Df Residuals:	901	BIC:	1.855e+04
Df Model:	14		
Covariance Type:	HC3		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.004e+04	3501.139	5.722	0.000	1.32e+04	2.69e+04
female	-506.6317	431.080	-1.175	0.240	-1352.670	339.406
age	-95.0795	30.570	-3.110	0.002	-155.077	-35.082
iv009_mod	-312.7320	154.567	-2.023	0.043	-616.086	-9.378
isced1997_r	1.6498	6.799	0.243	0.808	-11.693	14.993
eduyears_mod	123.4808	87.780	1.407	0.160	-48.797	295.758
mar_stat	682.1994	318.285	2.143	0.032	57.532	1306.866
hhsize	635.2108	421.766	1.506	0.132	-192.548	1462.970
partnerinhh	-4749.4561	745.472	-6.371	0.000	-6212.520	-3286.392
sphus	-253.0075	263.419	-0.960	0.337	-769.993	263.979
chronic_mod	47.7892	173.145	0.276	0.783	-292.026	387.605
casp	72.8280	44.627	1.632	0.103	-14.757	160.413
bmi	56.5252	50.473	1.120	0.263	-42.533	155.583
ep005_	45.5578	63.964	0.712	0.476	-79.978	171.093
co007_	1129.1060	241.078	4.684	0.000	655.966	1602.246

Obr. 12 Vystup regresní analýzy metodou MNČ

Všechny regresory vzhledem k počtu pozorování vyšli staticky významné, nicméně vidíme relativně malý podíl vysvětlené variability (upravený koeficient determinace R^2= 0.36) i tak velkým počtem vysvětlujících proměnných (14 prediktorů). Vzhledem k multikolinearitě regresorů, pozorujeme, že většina regresních koeficientů vyšla statisticky nevýznamná, nicméně multikolinearita nezkresluje odhad ‘thinc_m‘ (jsou stálé BLUE) a zásadně neporušuje předpoklady pro použití lineárního modelu.

VÝSLEDKY PRO REGRESNÍ ANALÝZU



Obr. 13 Porovnání rozdělení a bodový graf a skutečných a imputovaných příjmů pomocí lineární regrese

Pozorujeme zde, že kombinace metod MICE imputace a regresní imputace vyšla adekvátní, ale ne zcela přesná. Dokázali jsme pomocí naších prediktorů odhadnout jen průměrné příjmy, bez ohledu na vysokopříjmové skupiny respondentů. Potom odhad hodnot u vysokopříjmových skupin bývá problematický právě kvůli absenci informací z chvostu rozdělení. Tímto se regresní analýza jeví, podle našeho názoru, jako metoda imputace příjmů nevhodná, protože v podstatě odhaduje podmíněnou střední hodnotu bez ohledu na asymetrii rozdělení platů v populaci.

Proto v další části práce se zaměříme na pokročilejší metody neparametrické statistiky a strojového učení, abychom obešli možná omezení a předpoklady potřebné k imputaci hodnot.

K-NEJBLIŽŠÍCH SOUSEDŮ (K-NN)

Tentokrát použijeme neparametrickou statistickou metodu, jelikož budeme preferovat přesnost výsledku vůči interpretačním výhodám.

Nejdříve zkusíme provést imputaci a validaci pomocí algoritmu k-nejbližších sousedů (k-NN) přičemž doporučované hodnota pro parametr k pro datové soubory s velkým počtem proměnných jsou v intervalu [10; 25]. Dospěli jsme ke konkrétní hodnotě postupnou minimalizací střední čtvercové chyby při nastavení parametru **k=23** se ukázal jako optimální řešení, kdy jsme získali dost podobných pozorování, ale přitom příliš nepotlačili lokální variabilitu dat než se spoléhat na jeden globální vztahy mezi proměnnými (Marimont & Shapiro, Nearest Neighbour Searches and the Curse of Dimensionality, 1979). Také by tato skutečnost mohla naznačovat určitou nelinearitu ve vztazích mezi analyzovanými proměnnými.

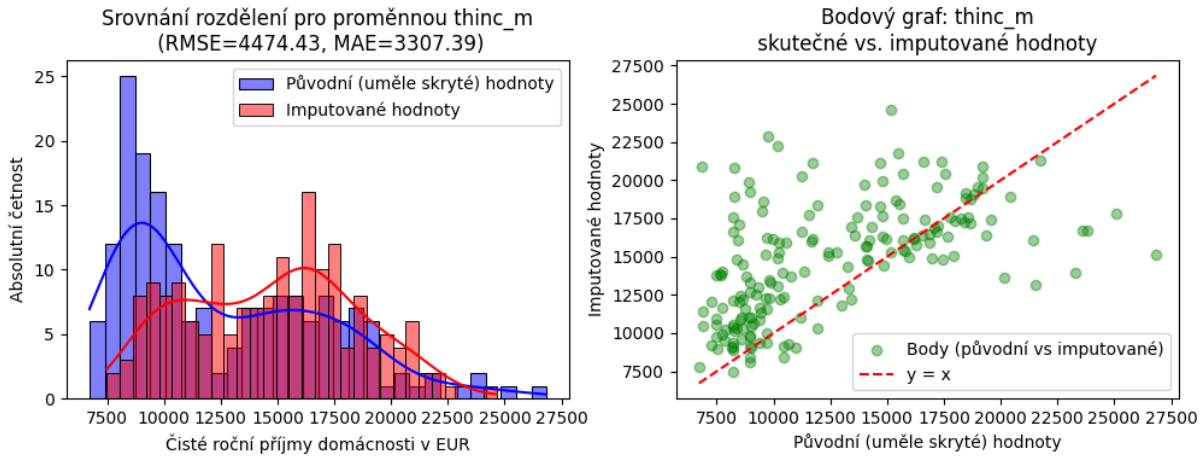
Metoda k-NN je jednou z jednodušších a zároveň poměrně efektivních technik pro doplňování (imputaci) chybějících hodnot v datovém souboru, používáme ji tak že:

1. **Definujeme „sousedství“**: Pro každý řádek (respondenta, záznam), u něhož chybí hodnota v konkrétní proměnné, hledáme **k** „nejbližších sousedů“ – tedy pozorování, které mají podobné hodnoty v ostatních (vyplněných) proměnných.
2. **Výběr vzdálenosti**: K měření „podobnosti“ používáme Gowerovu vzdálenost (vizte (Gower, 1971))
3. **Způsob doplnění**: Jakmile najdeme **k** sousedů, kteří mají tuto proměnnou vyplněnou, dopočítáme chybějící hodnotu jako (například) průměr nebo mód (u kategorických dat) z těchto sousedů.

Tento přístup dokáže zachytit vztah mezi proměnnými, protože imputovaná hodnota vychází z několika **podobných** záznamů, nikoli jen z podmíněné střední hodnoty celé proměnné.

VÝSLEDKY PRO METODU K-NEJBLIŽŠÍCH SOUSEDŮ (K-NN)

Podívejme se na výsledky imputace další metodou. Můžeme pozorovat, že, kromě toho, že nám vyšla o trochu lepší hodnota RMSE. Musíme ale, přiznat, že model není zdaleka optimální, protože také nedokáže podchytit vysokopříjmové skupiny, byť jsme lépe dokázali lépe modelovat mediánové platy. Pokud se podíváme na bodový graf, budeme pozorovat tendenci podhodnocovat nadhodnocovat maskované průměrné příjmy respondentů a také větší rozptýlenost hodnot.



Obr. 14 Porovnání rozdělení a bodový graf a skutečných a imputovaných příjmů pomocí metody k-NN

ZÁVĚR

Hlavní motivací této analýzy bylo zjistit typ chybějících pozorování, znázornit vzory „chybování“ mezi proměnnými a poskytnout rozhodovací rámec pro porovnání imputačních metod v kontextu výběrového šetření SHARE.

Po následném porovnání imputačních technik jsme dospěli k výsledku, že imputace pomocí vícerozměrné statistické metody k-NN funguje trochu lépe na základě metriky RMSE, ale nemá výhodu v interoperabilitě, kterou má deskriptivní interpretace regresní analýzy. Avšak, nás zajímá hlavně přesnost imputace a, jak můžeme pozorovat na obrazcích 13 a 14, tyto metody stále nejsou schopny imputovat příjmy movitějších domácností, i když spolehlivě odhadnou koncentraci hodnot kolem mediánu. Regresní imputace dokonce napodobila bimodální rozdělení příjmů, které, ale ve skutečnosti neodpovídá populačnímu modelu logaritmicko-normálního rozdělení.

Kvantitativní modely v ekonomii přece jenom dokážou modelovat realitu na základě měřitelných ukazatelů, ale pro účely imputace příjmů domácností by bylo dobré zachycovat informaci i nedatové povahy. Lze se pokusit snížit míru MNAR doplněním informací z administrativních zdrojů (např. daňové záznamy, pojišťovací data) nebo dalším dotazováním.

Nicméně, cíle práce byly splněny, hlavně protože jsme popsali rozhodovací rámec pro všeobecný postup pro práci s chybějícími hodnotami a vyzkoušeli dva odlišné přístupy modelování chybějících hodnot v datovém souboru s velkým počtem relevantních proměnných. Na základě RMSE, a hlavně grafické analýzy můžeme posoudit, že metoda nejbližšího souseda k-NN je vhodnější k použití v datovém souboru easySHARE, jelikož lépe odpovídá struktuře příjmů v populaci a také imputuje data z chvostu rozdělení, kde můžeme předpokládat, že nejvíc příjmů chybí.

REFERENCE

Easy-SHARE. (2024). Načteno z https://share-eric.eu/fileadmin/user_upload/Release_Guides/ea

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 857-871.

Li, C. (2013). *Little's test of missing completely at random.* Načteno z The Stata Journal : https://journals.sagepub.com/doi/pdf/10.1177/153686

M. Baak, R. K. (2019, Březen 19). *A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics.* Retrieved from https://arxiv.org/abs/1811.11440: https://phik.readthedocs.io/en/latest/

Malá, I. (2015). *VÍCEROZMĚRNÝ PRAVDĚPODOBNOSTNÍ MODEL ROZDĚLENÍ* . Načteno z POLITICKÁ EKONOMIE: https://polek.vse.cz/pdfs/pol/2015/07/06.pdf

Marimont, R., & Shapiro, M. (1979). Nearest Neighbour Searches and the Curse of Dimensionality. *IMA Journal of Applied Mathematics*, 59–70.

Marimont, R., & Shapiro, M. (1979). Nearest Neighbour Searches and the Curse of Dimensionality. *IMA Journal of Applied Mathematics*, 59–70.

Roderick J., A. L. (1986). *Statistical Analysis with Missing Data.*

Scikit-learn. (2011). *Dataset transformations.* Načteno z Imputation of missing values: https://scikit-learn.org/1.5/modules/

Stef van Buuren, K. G.-O. (2011). *Multivariate Imputation by Chained.* Načteno z Journal of Statistical Software: https://www.jstatsoft.org/article/view/v045i03

PŘÍLOHA

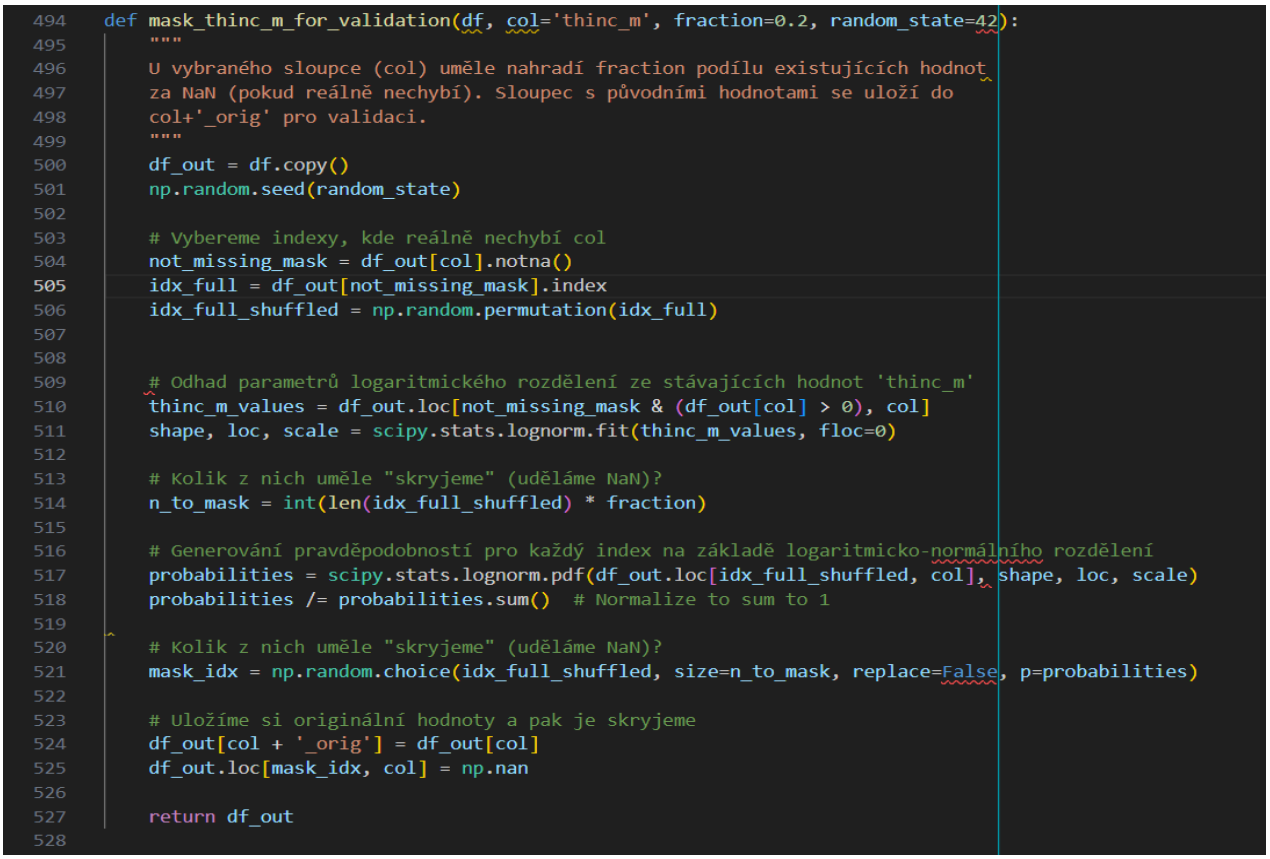
```python
# Vlastní implementace Littleova testu


def little_mcar_test(data, alpha=0.05):
    """
    Performs Little's MCAR (Missing Completely At Random) test on a dataset with missing values.
    """
    data = pd.DataFrame(data)
    data.columns = ['x' + str(i) for i in range(data.shape[1])]
    data['missing'] = np.sum(data.isnull(), axis=1)
    n = data.shape[0]
    k = data.shape[1] - 1
    df = k * (k - 1) / 2
    chi2_crit = chi2.ppf(1 - alpha, df)
    chi2_val = ((n - 1 - (k - 1) / 2) ** 2) / (k - 1) / ((n - k) * np.mean(data['missing']))
    p_val = 1 - chi2.cdf(chi2_val, df)
    if chi2_val > chi2_crit:
        print(
            'Reject null hypothesis: Data is not MCAR (p-value={:.4f}, chi-square={:.4f})'.format(p_val, chi2_val)
        )
    else:
        print(
            'Do not reject null hypothesis: Data is MCAR (p-value={:.4f}, chi-square={:.4f})'.format(p_val, chi2_val)
        )


# Test MCAR
little_mcar_test(df_nan, 0.05)

```

Obr. 15 Implementace funkce pro provedení Litlleova MCAR testu v Pythonu



```python
def mask_thinc_m_for_validation(df, col='thinc_m', fraction=0.2, random_state=42):
    """
    U vybraného sloupce (col) uměle nahradí fraction podílu existujících hodnot
    za NaN (pokud reálně nechybí). Sloupec s původními hodnotami se uloží do
    col+'_orig' pro validaci.
    """
    df_out = df.copy()
    np.random.seed(random_state)

    # Vybereme indexy, kde reálně nechybí col
    not_missing_mask = df_out[col].notna()
    idx_full = df_out[not_missing_mask].index
    idx_full_shuffled = np.random.permutation(idx_full)


    # Odhad parametrů logaritmického rozdělení ze stávajících hodnot 'thinc_m'
    thinc_m_values = df_out.loc[not_missing_mask & (df_out[col] > 0), col]
    shape, loc, scale = scipy.stats.lognorm.fit(thinc_m_values, floc=0)

    # Kolik z nich uměle "skryjeme" (uděláme NaN)?
    n_to_mask = int(len(idx_full_shuffled) * fraction)

    # Generování pravděpodobností pro každý index na základě logaritmicko-normálního rozdělení
    probabilities = scipy.stats.lognorm.pdf(df_out.loc[idx_full_shuffled, col], shape, loc, scale)
    probabilities /= probabilities.sum()  # Normalize to sum to 1

    # Kolik z nich uměle "skryjeme" (uděláme NaN)?
    mask_idx = np.random.choice(idx_full_shuffled, size=n_to_mask, replace=False, p=probabilities)

    # Uložíme si originální hodnoty a pak je skryjeme
    df_out[col + '_orig'] = df_out[col]
    df_out.loc[mask_idx, col] = np.nan

    return df_out

```

Obr. 16 Implementace algoritmu maskování a pravděpodobnostního výběru z log-normálního rozdělení v Pythonu