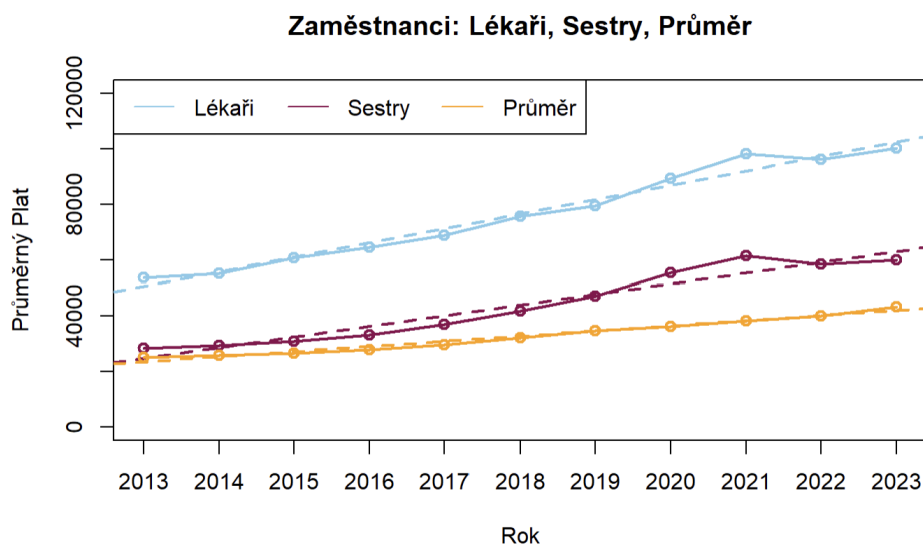


SAN Semestrální práce

Matěj Dynda, Vendula Hrnčířová, Milan Španko, Štěpán Gorichev
Leden 2024

1. Zadání

Zdravotnický systém hraje klíčovou roli v kvalitě života ve společnosti. V České republice, podobně jako v dalších zemích, čelí zdravotnictví mnoha výzvám [1], [2], [3]. Na Obrázku 1 lze pozorovat vývoj platu zdravotníků vůči průměrnému platu obyvatel v České republice. Tento vývoj má poměrně lineární trend všech tří složek. Vidíme, že plat lékařů roste rychleji (strmost 5192), než růst pro zdravotní sestry (strmost 3834) a celorepublikový průměr (strmost 1853) v posledních deseti letech. Avšak, rychlost růstu počtu lékařů (strmost 903) a sester (strmost 639) je oproti růstu počtu obyvatel výrazně pomalejší (strmost 24086). To ukazuje na potenciální nedostatek zdravotnického personálu, což má za následek vyšší pracovní tlak na lékaře a sestry. Ukazuje se, že zdánlivě rychlý růst platů zdravotníků se příliš neodráží na jejich počtu. Ačkoli, vysoké platy mohou být jedním z lákavých důvodů pro výběr tohoto povolání, podle posledních zpráv jsou však tyto trendy v platovém ohodnocení zkreslené tím, že se po zdravotnickém personálu chce, aby pracoval nad rámec standardní pracovní doby, a to v podobě přesčasů nebo služeb.



Obrázek 1: Vývoj průměrného platu zdravotníků a průměrného platu obyvatel ČR ([ČSÚ](#))

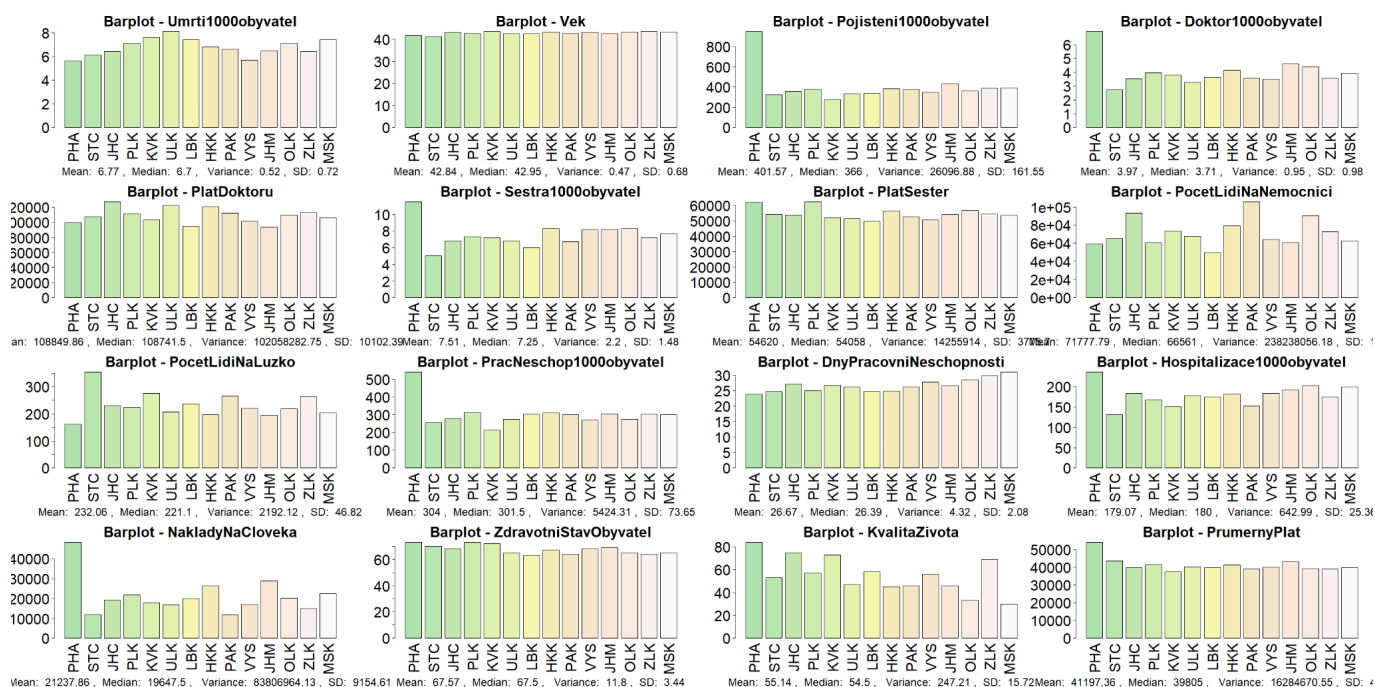
Tyto trendy by měly být zohledněny při plánování rozpočtů pro zdravotnictví, při vzdělávacích a personálních strategiích na další roky, aby došlo k zajištění dostačující počtu zdravotníků a zamezilo se jejich přepracování. Současná data však poukazují na velké regionální rozdíly, co se lékařské péče týče, což může mít zásadní vliv na vývoj společnosti v ČR. Socioekonomické podmínky mohou výrazně ovlivňovat kvalitu a dostupnost zdravotní péče. Tento vliv se následně může odrazit i na úmrtnosti pacientů v nemocnicích. Rádi bychom prozkoumali, jak různé socioekonomické faktory - platové ohodnocení zdravotnického personálu, počet lékařů a lůžek, regionální ukazatele životní úrovně - mohou ovlivňovat úmrtnost v jednotlivých krajích ČR. Pro tuto analýzu jsme si vybrali data pro rok 2022, jelikož byly k dispozici vhodné datasety a pandemie Covid-19 už byla na ústupu.

2. Výzkumná otázka

Mají socioekonomické podmínky zdravotnického personálu vliv na úmrtnost pacientů v nemocnicích v různých krajích ČR?

3. Data

Nenašli jsme analýzu, která by se zabývala stejnou otázkou jako naše práce, a proto jsme se rozhodli prozkoumat faktory, které by mohli mít na naši závislou proměnnou vliv a které mají rozdílné hodnoty napříč kraji viz. Obrázek 2. Vybrané proměnné poskytují komplexní pohled na socioekonomické a zdravotnické podmínky v jednotlivých krajích ČR (Tabulka 1).



Obrázek 2: Porovnání vybraných proměnných napříč kraji

Hlavní město Praha	PHA	Karlovarský kraj	KVK	Pardubický kraj	PAK	Olomoucký kraj	OLK
Středočeský kraj	STC	Liberecký kraj	LBK	Kraj Vysočina	VYS	Zlínský kraj	ZLK
Jihočeský kraj	JHC	Ústecký kraj	ULK	Jihomoravský kraj	JHM	Moravskoslezský kraj	MSK
Plzeňský kraj	PLC	Královéhradecký kraj	HKK				

Tabulka 1: Zkratky jednotlivých krajů

Jedná se o následující proměnné:

- **Umrtí1000obyvatel** – Udává počet úmrtí v nemocnicích na 1000 obyvatel v daném kraji. Tato proměnná je důležitá pro hodnocení efektivity zdravotnického systému. (UZIS: Databáze zemřelých)
- **Vek** – Průměrný věk obyvatel v jednotlivých krajích. Vyšší věk obyvatel může mít vliv na rozvoj nemocí u organismu a tudíž šance na úmrtí/hospitalizaci jedince. ([ČSÚ](#))
- **Pojisteni1000obyvatel** – Počet pojištěných osob na 1000 obyvatel. Zdravotní pojištění motivuje lidi k docházení na preventivní prohlídky a dřívejšího vyhledání lékaře v případě problému. ([ČSÚ](#))
- **Doktor1000obyvatel** – Počet lékařů na 1000 obyvatel. Tato proměnná reflektuje dostupnost lékařské péče v krajích. ([Personální kapacity 2023 UZIS](#))
- **PlatDoktoru** – Průměrný plat lékařů (z veřejného sektoru). Výše platu může ovlivňovat motivaci lékařů k výkonu povolání a jejich setrvání v oboru. Je vhodné ho porovnávat s průměrným platem obyvatel v regionu. ([Personální kapacity 2023 UZIS](#))

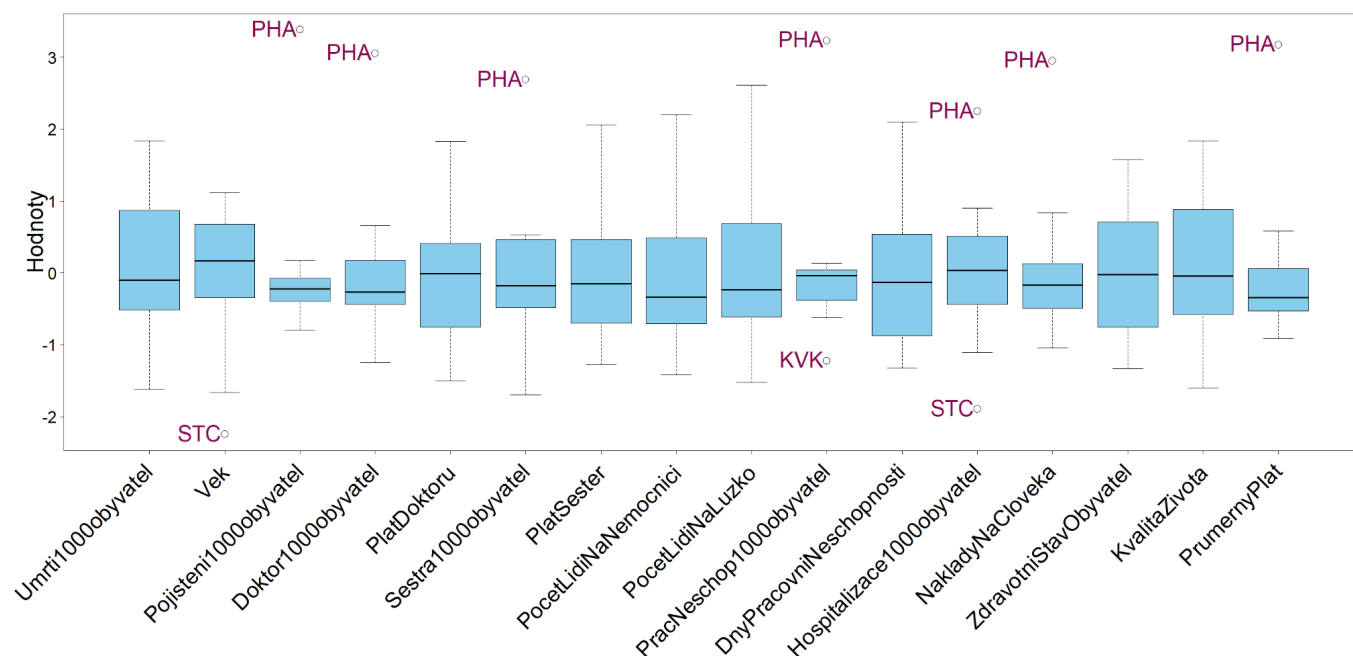
- **Sestra1000obyvatel** – Počet zdravotních sester na 1000 obyvatel. Dostatečný počet sester je klíčový pro poskytování efektivní péče. ([Personální kapacity 2023 UZIS](#))
- **PlatSester** – Průměrný plat zdravotních sester (z veřejného sektoru). Výše platu ovlivňuje atraktivitu povolání a míru spokojenosti sester. Je vhodné ho porovnávat s průměrným platem obyvatel v regionu. ([Personální kapacity 2023 UZIS](#))
- **PocetLidiNaNemocnici** – Průměrný počet pacientů na jednu nemocnici akutní péče. Ukazuje kapacitu a vytížení zdravotnických zařízení v krajích. ([ČSÚ](#))
- **PocetLidiNaLuzko** – Průměrný počet osob na jedno nemocniční lůžko. Odráží dostupnost lůžkových kapacit v nemocnicích a schopnost reagovat na extrémní události (Covid-19). ([ČSÚ](#))
- **PracNeschopnost1000obyvatel** – Počet případů pracovní neschopnosti na 1000 obyvatel. Indikuje obecný zdravotní stav pracující populace. Data k této proměnné nevylučují opakovanou pracovní neschopnost u jednoho obyvatele v daném roce. ([ČSÚ](#))
- **DnyPracovniNeschopnosti** – Průměrný počet dnů strávených v pracovní neschopnosti. Ukazuje závažnost zdravotních problémů v daném regionu. ([ČSÚ](#))
- **Hospitalizace1000obyvatel** – Počet hospitalizací na 1000 obyvatel. Odráží jak výskyt zdravotních problémů v populaci, tak možnost využití nemocniční péče v regionu. (UZIS: Přehled hospitalizačních případů)
- **NakladyNaCloveka** – Průměrné zdravotnické náklady na obyvatele. Ukazuje investice do zdravotní péče na jednotlivce. ([UZIS](#))
- **ZdravotniStavObyvatel** – Kolik procent obyvatel subjektivně hodnotí, že mají *velmi dobrý* či *dobrý zdravotní stav*. Odráží celkovou úroveň zdraví populace. ([ČSÚ](#))
- **KvalitaZivota** – Objektivní hodnocení kvality života v regionu na základě zdravotních, ekonomických a sociálních faktorů. Je důležitým ukazatelem celkového blahobytu. ([Místo pro život](#))
- **PrumernyPlat** – Průměrná mzda v daném regionu. Ekonomická situace obyvatel může ovlivňovat jejich přístup ke zdravotní péči. ([ČSÚ](#))

4. Problematika analýzy malého počtu vzorků

Vzhledem k tomu, že dataset je omezen na počet krajů v České republice (tj. 14), bude velmi obtížné dosáhnout spolehlivých a robustních výsledků. Malá velikost vzorku omezuje schopnost modelů vysvětlit velkou část variability a snižuje jejich statistickou sílu. K ověření výsledků je vhodné používat více metod a navzájem je porovnávat. Například testování normality pomocí *Shapiro-Wilk* testu je velmi citlivé na velikost vzorku a je vhodné ho proto doplnit například vizuální kontrolou *Q-Q* plot. Jedním z hlavních nebezpečí modelování na malém vzorku je *overfitting* (přeučení), které vede ke špatné generalizaci, přestože může vykazovat vysokou přesnost na trénovacích datech. Je proto potřeba tyto modely testovat například pomocí *Cross Validate*. V případech, kdy máme k dispozici pouze malý vzorek, často také bývá vhodnější používat jednodušší modely. Nakonec je důležité zahrnout tyto omezení do interpretace výsledků.

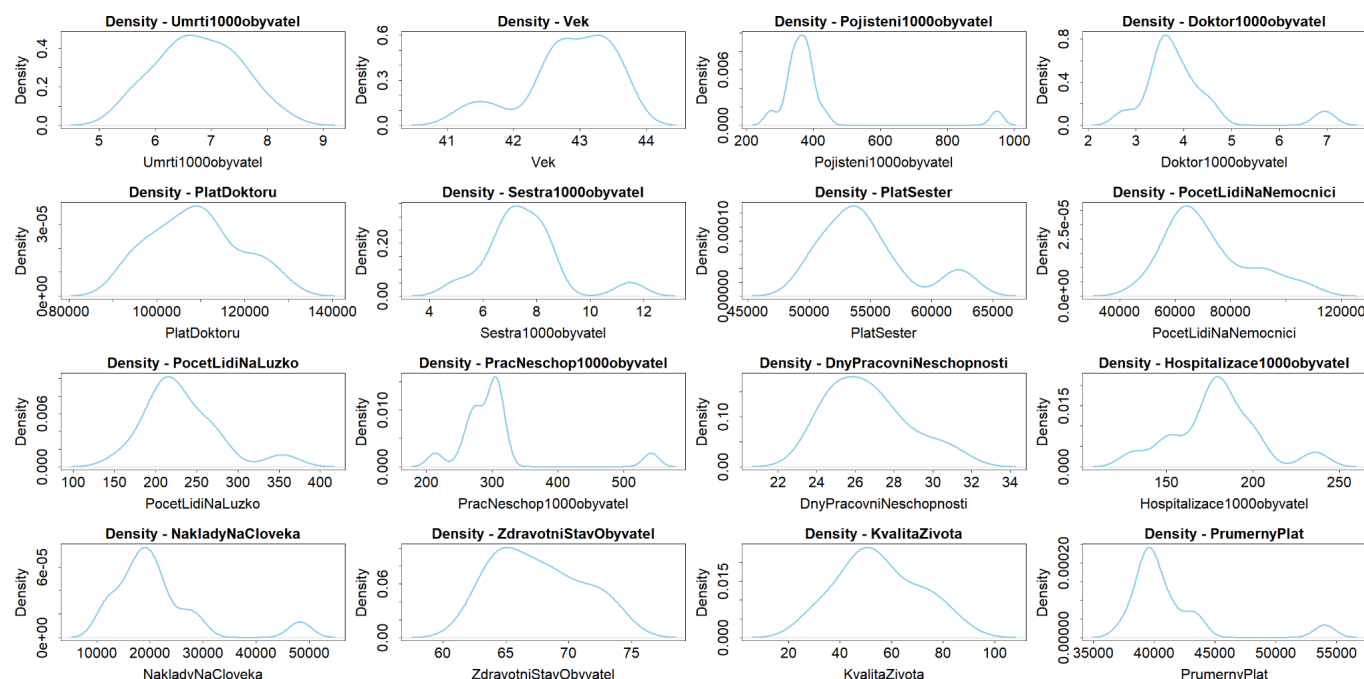
5. Proměnné

Na Obrázku 3 vidíme *boxplot* graf jednotlivých uvažovaných proměnných. Můžeme vidět, že Praha bude často v roli *anomálie*, kdy její hodnoty jsou často výrazně vyšší oproti zbytku krajů. Středočeský kraj a Karlovarský kraj bude také potřeba kontrolovat jakožto možné *outliery*. Můžeme dále říct, že proměnné *Pojisteni1000obyvatel* a *PracNeschop1000obyvatel* mají nízkou variabilitu v datech.



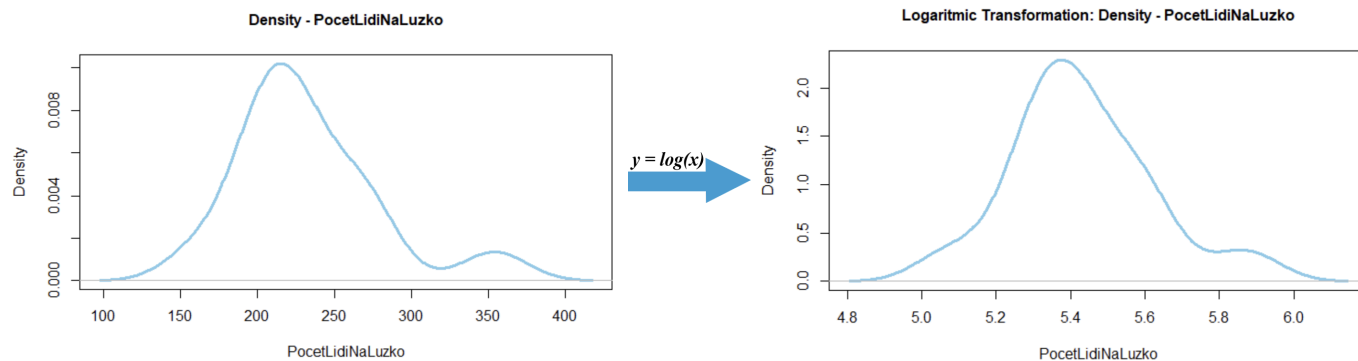
Obrázek 3: Boxplot jednotlivých proměnných

Obrázek 4 zobrazuje hustotní grafy jednotlivých proměnných, na základě, kterých můžeme odhadnout rozdělení dat. Naše závislá proměnná *Umrti1000obyvatel* a většina dalších proměnných odpovídá normální distribuci a proto není třeba ji transformovat (potřeba pomatovat na *anomálie*!). V některých případech však pozorujeme pravostrannou šikmost (*right-skewed*), například u *ZdravotniStavObyvatel*, *DnyPracovniNeschopnosti*, *PocetLidiNaNemocnici* a další.



Obrázek 4: Hustotní grafy jednotlivých proměnných

Ta lze částečně potlačit pomocí logaritmické transformace ($y = \log(x)$), kterou můžeme použít i díky tomu, že všechny hodnoty proměnných jsou kladné. Na Obrázku 5 je ilustrováno, jak logaritmická transformace ovlivňuje pravostrannou šikmost. *Right-skewed* distribuce je často v našich případech dána především extrémními hodnotami, které pocházejí převážně ze vzorku pro Hlavní město Praha. Při porovnání modelů s a bez logaritmické transformace jsme však nepozorovali výrazné vylepšení, a proto zde dále prezentované modely jsou bez transformace dat, což bylo podpořeno i nižšími hodnotami získanými z funkce *skewness*.



Obrázek 5: Vliv logaritmické transformace dat na hustotní distribuci

Dále jsme prověřovali, jaká – a jestli vůbec – existuje korelace mezi jednotlivými proměnnými. Výsledky této analýzy jsou zachyceny v korelační tabulce na Obrázku 6. Křížky uvnitř koleček označují případy, kdy je absolutní hodnota korelačního koeficientu mezi proměnnými vyšší než 0.8. Což naznačuje, že tyto proměnné spolu úzce souvisejí. Můžeme vidět, že zvolená závislá proměnná *Umrți1000obyvatel* nevykazuje silnou korelaci s žádnou jinou proměnnou. To naznačuje, že pro vytvoření úspěšného modelu bude pravděpodobně potřeba použít multivariantní analýzu. Naopak proměnné *Pojisteni1000obyvatel*, *Doktor1000obyvatel*, *Sestra1000obyvatel*, *Hospitalizace1000obyvatel*, *NakladyNaCloveka*, a *PrumernyPlat* vykazují vysokou korelaci s jinými proměnnými.



Obrázek 6: Korelační matice proměnných

6. Model pro úmrtnost v nemocnicích (*Umrsti1000obyvatel*)

6.1. Univariantní analýza

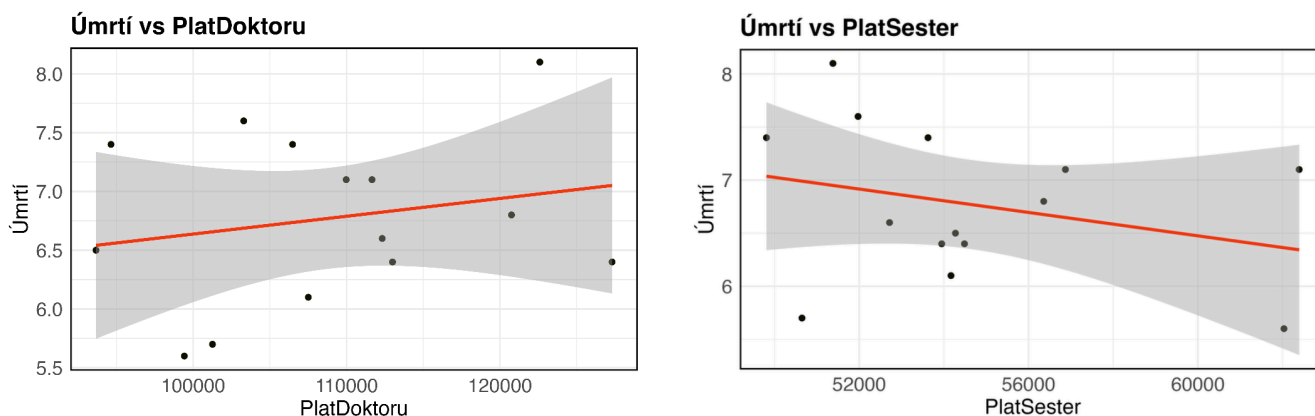
Pro *Umrtnost1000obyvatel* jsme vytvořili jednoduché lineární regresní modely, kde prediktor je vždy jedna proměnná. Ve všech případech byl pozorován lineární trend mezi prediktorem a závislou proměnnou. Pro ověření předpokladů linearity proběhla kontrola normality a homoskedasticity. Normalitu dat jsme ověřili pomocí *Shapiro-Wilk* testu a vizuálně pomocí *Q-Q* plotu. Homoskedasticitu jsme testovali pomocí testu *Breusch-Pagan* a dále *Residuals vs. Fitted Values* plot. Po ověření těchto podmínek byly vytvořeny regresní modely pro jednotlivé prediktory vůči závislé proměnné. V Tabulce 2 jsou zapsané výsledky jednotlivých modelů včetně informací o splnění podmínek pro lineární regresi. Tato tabulka poskytuje přehled o vhodnosti použití lineární regrese pro jednotlivé proměnné a slouží jako základ pro výběr proměnných do případné multivariantní analýzy.

	Modelová statistika		Podmínky lineární regrese	
<i>Umrsti1000obyvatel</i> ~	Multiple R-squared	p-value F-stat	normalita	homoskedasticita
Vek	0.1433	0.1819	✓	✓
Pojisteni1000obyvatel	0.2602	0.06238	✓	✓
Doktor1000obyvatel	0.1172	0.2309	✓	✓
PlatDoktoru	0.04465	0.4683	✓	✓
Sestra1000obyvatel	0.1394	0.1885	✓	✓
PlatSester	0.08264	0.319	✓	✓
PocetLidiNaNemocnici	0.001148	0.9085	✓	✓
PocetLidiNaLuzko	0.000793	0.9239	✓	✓
PracNeschop1000obyvatel	0.2138	0.09596	✓	✓
DnyPracovniNeschopnosti	0.02375	0.5989	✓	✓
Hospitalizace1000obyvatel	0.04077	0.4888	✓	✓
NakladyNaCloveka	0.09828	0.2751	✓	✓
ZdravotniStavObyvatel	0.103	0.2633	✓	✓
KvalitaZivota	0.1577	0.1597	✓	✓
PrumernyPlat	0.2978	0.04354	✓	✓

Tabulka 2: Výsledky univariantní analýzy pro modelování *Umrsti1000obyvatel*

Z analýzy jednoduchých regresních modelů jsme zjistili, že jediným statisticky významným modelem je ten s proměnnou *PrumernyPlat* (označeno červeně). Tento model však vysvětluje pouze 30% variability v datech, což ukazuje, že samostatná proměnná není dostatečná pro přesné vysvětlení úmrtnosti. Nabízí se možnost sestavení přesnějšího modelu pomocí multivariantní analýzy.

Proměnné, které jsou klíčové pro naši hlavní otázku *PlatDoktoru* a *PlatSester*, nedostatečně vysvětlují variabilitu v datech a ani nejsou statisticky významné. Paradoxně vyšlo, že vyšší plat doktorů je spojen s vyšší úmrtností. Tento výsledek odporuje základním předpokladům a může poukazovat na přítomnost skrytých faktorů nebo souvislostí, které nejsou přímo zřejmé. Naopak u zdravotních sester vidíme, že vyšší platy zdravotních sester jsou spojeny s nižší úmrtností, což je v souladu s očekáváním. Regresní křivky těchto modelů jsou zobrazeny na Obrázku 7. Je otázkou, zda se tyto proměnné uplatní v multivariantní analýze.



Obrázek 7: Lineární regrese pro proměnné *PlatDoktoru* a *PlatSester*

6.2. Multivariantní analýza

U všech proměnných byly úspěšně ověřeny předpoklady pro linearitu při univariantní analýze, což nám umožňuje dále pokračovat s multivariantním modelováním pomocí lineární regrese. Předpokladem je, aby počet prediktorů (nezávislých proměnných) byl menší než počet vzorků (tj. musíme mít méně než 14 prediktorů). Pro základní model vyřadíme ty, které v univariantní analýze vykázaly p-value větší než 0.5. Tento model sice přinesl vysvětlení 63% variability, ale byl statisticky nevýznamný. I když test homoskedasticity (*Breusch-Pagan*) a test normality (*Shapiro-Wilk*) ukázaly splnění podmínek pro lineární regresi. Pravděpodobně došlo k *overfittingu* z důvodu malého počtu bodů volnosti. Dalším z problémů může být také multikolinearita mezi prediktory.

Pomocí *Variance Inflation Factor* (VIF) hodnot jsme určili proměnnou, *Pojisteni1000obyvatel*, která měla vysokou multikolinearitou a z modelu jsme je odstranili. Následně jsme aplikovali *stepwise selection* pro výběr menšího počtu prediktorů. Metoda označila za vhodné tyto proměnné: *Sestra1000obyvatel*, *NakladyNaCloveka* a *PrumernyPlat*. *LASSO* metoda ještě doporučila proměnnou *KvalitaZivota*, ale jejím přidáním jsme získali horší výsledky modelu. Výsledný model byl tedy sestaven z proměnných *Sestra1000obyvatel*, *NakladyNaCloveka* a *PrumernyPlat*.

```
lm(formula = Umrti1000obyvatel ~ Sestra1000obyvatel + NakladyNaCloveka + PrumernyPlat, data = promenne)
```

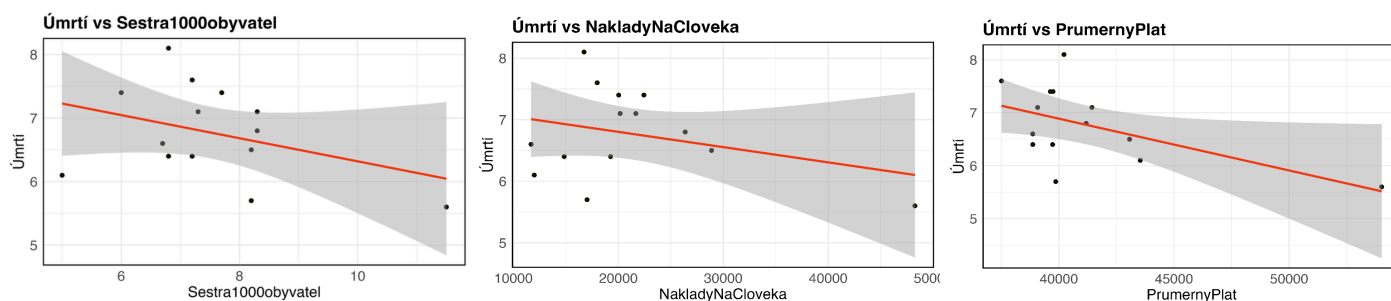
Umrti1000obyvatel	Modelová statistika			Podmínky lineární regrese	
	Multiple R-squared	Adjusted R-squared	p-value F-statistiky	normalita	homoskedasticita
Sestra1000obyvatel, NakladyNaCloveka, PrumernyPlat	0.4987	0.3483	0.06527	✓	✓

Tabulka 3: Výsledky multivariantní analýzy pro modelování *Umrti1000obyvatel*

Model vysvětluje téměř 50% variability dat a je popsán rovnicí ukazující lineární vztah mezi úmrtností v nemocnici na 1000 obyvatel (*Umrtnost1000obyvatel*) a vybranými prediktory:

$$\text{Umrti1000obyvatel} = 15.83 - 0.3722 \cdot \text{Sestra1000obyvatel} + 0.0001039 \cdot \text{NakladyNaCloveka} - 0.0002057 \cdot \text{PrumernyPlat}$$

Model ukazuje, že pokud přidáme o jednu sestru více na 1000 obyvatel, úmrtnost v nemocnicích se sníží o 0.3722, což naznačuje podobný trend jako u univariální analýzy (Obrázek 8). Podezřelé však je, že pokud do zdravotnictví přidáme o 1 korunu na člověka navíc tak na 1000 obyvatel zemře o 0.0001039 lidí více, což neodpovídá modelu z univariální analýzy (Obrázek 8). Dává však smysl, že vyšší úmrtnost může být spojena s vyššími finančními prostředky vynaloženými na léčbu těžce nemocných pacientů, kteří mají vyšší pravděpodobnost úmrtí. Nárůst průměrného platu obyvatelstva o 1 korunu má za následek snížení úmrtnosti v nemocnicích, konkrétně o 0.0002057 na 1000 obyvatel, což naznačuje pozitivní vliv vyššího příjmu na zdravotní stav populace (Obrázek 8). Celkově však model není dostatečně přesný a přibližně polovina variability v úmrtnosti je způsobena jinými faktory, které nebyly zahrnuty. *P-value F-statistiky* je těsně nad 0.05 (0.065), což ukazuje na to, že model není statisticky významný na běžné hladině 0.05



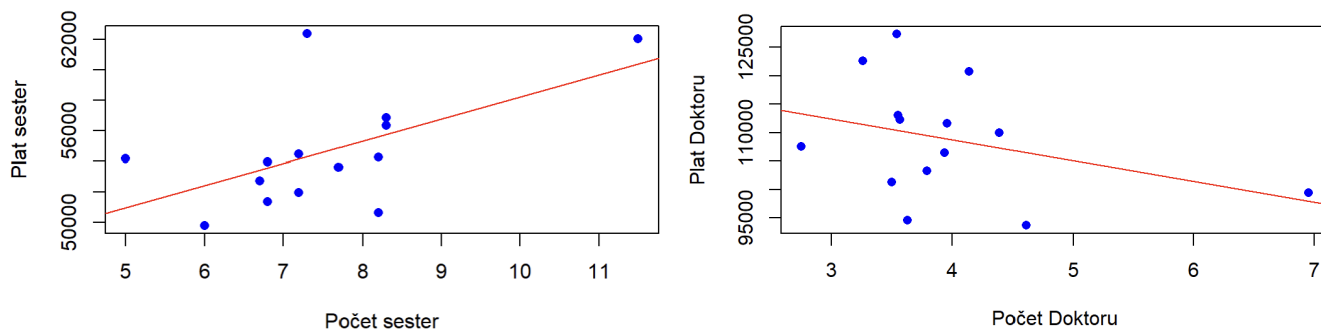
Obrázek 8: Lineární regrese pro proměnné *Sestra1000obyvatel*, *NakladyNaCloveka* a *PrumernyPlat*

6.2.1. Testování multivariantního modelu

Pro ověření modelu jsme použili *Leave-One-Out Cross Validation* (LOO-CV), což je vhodná technika v případě menšího počtu vzorků. Výsledky testování ukázaly, že model není ani tak silný, jak se zprvu zdálo: RMSE = 0.6825, R-squared = 0.2379, MAE = 0.5473. Pokud vytvoříme *predikce* na základě modelu s proměnnými *Sestra1000obyvatel*, *NakladyNaCloveka* a *PrumernyPlat*, tak získáme RMSE = 0.5429 a R-squared = 0.39. Hodnoty ukazují, že model není příliš přesný a vysvětluje malou část variability dat.

Pokoušeli jsme se vytvořit také modely s nelineárními vztahy (například polynom 2 stupně pro proměnnou *PocetLidiNaLuzko*), ale výsledky se ukázaly jako horší než u lineárního modelu. Rovněž se nám nepodařilo identifikovat konkrétní interakce mezi prediktory, které by vedly ke zlepšení modelu. Ani odstranění dat pro Hlavní město Prahu, které bylo v některých případech označeno jako *anomálie*, nevedlo ke zlepšení.

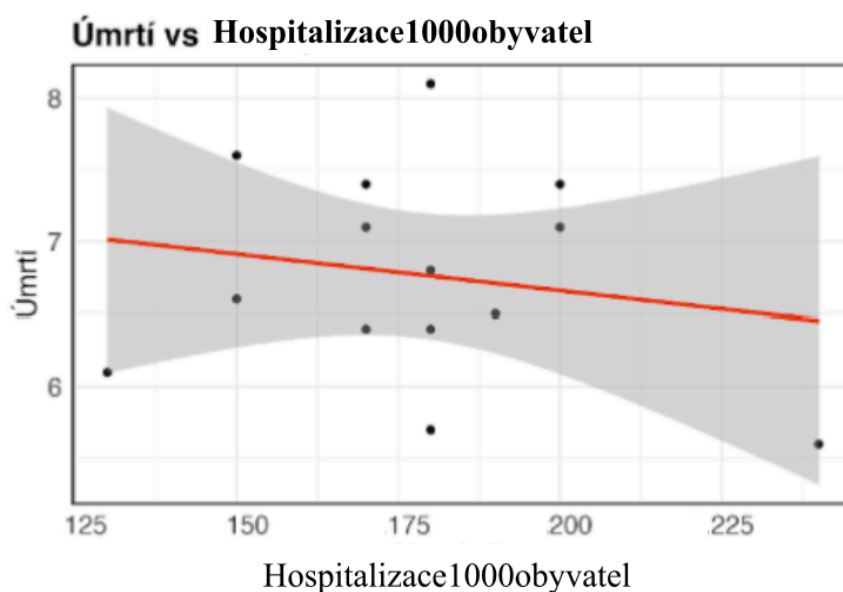
Ve výsledném modelu se opět neobjevily dvě hlavní nezávislé proměnné - *PlatDoktoru* a *PlatSester*. Největší vliv na model *Umrtnost1000obyvatel* měl počet sester, který však koreluje s platovým ohodnocením sester (lze pozorovat na Obrázku 9). Po vykreslení vztahu mezi platem a počtem doktorů se ukázalo, že větší počet doktorů má za následek spíše nižší plat (Obrázek 9). To zkresluje hlavně vzorek pro Prahu, kde je plat nižší než v jiných regionech, a to i přesto, že se jedná o město se zvýšenými životními náklady. To může být důsledkem toho, že oblasti, kde není obecně dostatek lékařů, lákají zdravotníky na vysoké finanční ohodnocení.



Obrázek 9: Lineární regrese pro proměnné *Doktor1000obyvatel* vůči *PlatDoktoru* a *Sestra1000obyvatel* vůči *PlatSester*

7. Model pro míru hospitalizace (*Hospitalizace1000obyvatel*)

Zkoumaný vliv na zdravotnictví by mohl být dobře popsán i proměnnou *Hospitalizace1000obyvatel*, která vyjadřuje počet hospitalizovaných lidí na 1000 obyvatel. Nicméně, regionální rozdíly v počtu hospitalizací (Obrázek 2) naznačují, že tento ukazatel může být ovlivněn komplexním souborem faktorů, jako je zdravotní stav populace, dostupnost a kvalita zdravotní péče a socioekonomické podmínky regionu. Tudíž můžeme očekávat sestavení nového modelu, ve kterém budou hrát roli jiné prediktory. Výsledné hodnoty se dají chápat jako kombinace dvou skutečností: v kraji je více lidí, kteří potřebují hospitalizaci (tedy jsou více nemocní), a kraj si může dovolit hospitalizovat více lidí (tedy má pro ně adekvátní materiální i personální kapacity). Z Obrázku 10 lze s opatrností usoudit (i když se nejedná o statisticky významný model), že se zvyšujícím se počtem hospitalizací dochází ke snížení úmrtí.



Obrázek 10: Lineární regrese pro proměnné *Umrti1000obyvatel* a *Hospitalizace1000obyvatel*

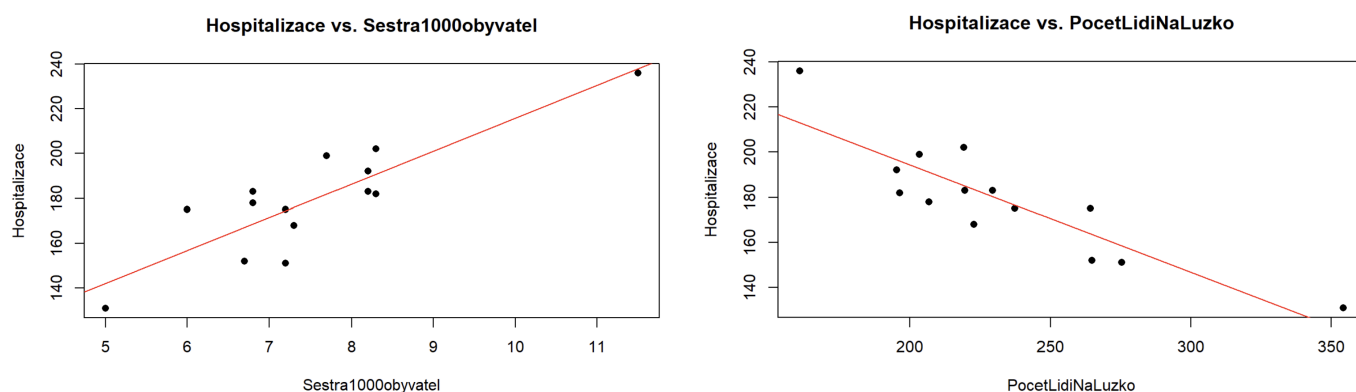
7.1. Univariantní analýza

Univariantní analýza proměnné *Hospitalizace1000obyvatel* ukázala, že získáváme více statisticky významných vztahů než pro *Umrtnost1000obyvatel*. V Tabulce 4 jsou zapsané výsledky pro jednotlivé modely, včetně ověření, zda byly splněny podmínky pro lineární regresi. Z Tabulky 4 vidíme, že proměnné *Pojisteni1000obyvatel*, *Doktor1000obyvatel*, *Sestra1000obyvatel*, *PocetLidiNaLuzko*, *PracNeschop1000obyvatel*, *NakladyNaClloveka* a *PrumernyPlat* dobře popisují závislou proměnnou *Hospitalizaci1000obyvatel*. Tyto proměnné vykazují silné vztahy s mírou hospitalizace, což naznačuje, že faktory jako dostupnost zdravotní péče, počet zdravotnického personálu, ekonomické ukazatele a pracovní neschopnost mohou mít vliv v jednotlivých regionech.

	Modelová statistika		Podmínky lineární regrese	
Hospitalizace1000obyvatel ~	Multiple R-squared	p-value F-statistiky	normalita	homoskedasticita
Umrtil1000obyvatel	0.04077	0.4888	✓	✓
Vek	0.001765	0.8866	✓	
Pojisteni1000obyvatel	0.5228	0.003476	✓	✓
Doktor1000obyvatel	0.6655	0.0003749	✓	✓
PlatDoktoru	0.02779	0.569	✓	✓
Sestra1000obyvatel	0.7439	7.208e-05	✓	✓
PlatSester	0.1754	0.136	✓	✓
PocetLidiNaNemocnici	0.03824	0.5029	✓	✓
PocetLidiNaLuzko	0.7711	3.624e-05	✓	✓
PracNeschop1000obyvatel	0.5155	0.003828	✓	✓
DnyPracovniNeschopnosti	0.01887	0.6395	✓	✓
NakladyNaCloveka	0.6607	0.0004095	✓	✓
ZdravotniStavObyvatel	0.001983	0.8532	✓	
KvalitaZivota	0.004224	0.8253	✓	✓
PrumernyPlat	0.302	0.04179	✓	✓

Tabulka 4: Výsledky univariální analýzy pro modelování *Hospitalizace1000obyvatel*

Nejlepšími prediktory byly *Sestra1000obyvatel* a *PocetLidiNaLuzko* (Obrázek 11). Křivky na Obrázku 11 ukazují, že s větší personální kapacitou (zdravotních sester), je hospitalizováno více lidí, protože je zajištěna kvalitní péče. Naopak při nedostatku nemocničních lůžek (více lidí na jedno lůžko) míra hospitalizace klesá. To může být způsobeno omezením kapacity nemocnic, což ztěžuje přijímání dalších pacientů na lůžka.



Obrázek 11: Lineární regrese pro proměnné *Sestra1000obyvatel* a *PocetLidiNaLuzko* vůči *Hospitalizace1000obyvatel*

7.2. Multivariantní analýza

Multivariantní analýza nám pomůže získat komplexní pohled na to jak více faktorů (personální a materiální kapacita, socioekonomické faktory a regionální rozdíly) souběžně ovlivňuje míru hospitalizace, místo zaměření se pouze na jeden faktor.

Pro základní multivariantní model jsme odstranili prediktory, které měli v univariantní analýze *p-value* vyšší než 0.5 (Tabulka 4). Tento prvotní model sice splňuje požadavky pro normalitu a homoskedasticitu, ale není statisticky významný (*p-value* 0.1177), vykazuje vyšší míru multikolinearity mezi některými prediktory. Po zúžení modelu na prediktory s *p-value* pod 0.05 v univariantní analýze (Tabulka 5) došlo k výraznému zlepšení. Dosáhli jsme statistické významnosti (*p-value* 0.018) při dodržení předpokladů pro lineární regresi. Pomocí tohoto modelu jsme byli schopni vysvětlit 88 % variability v datech. Nicméně některé proměnné byly opět zatíženy multikolinearitou (v Tabulce 5 označeny tučně).

Hospitalizace1000obyvatel	Modelová statistika			Podmínky lineární regrese	
	Multiple R-squared	Adjusted R-squared	p-value F-stat	normalita	homoskedasticita
Pojisteni1000obyvatel, Doktor1000obyvatel, Sestra1000obyvatel, PocetLidiNaLuzko, PracNeschopen1000obyvatel, NakladyNaCloveka, PrumernyPlat	0.884	0.7487	0.018	✓	✓
Doktor1000obyvatel, Sestra1000obyvatel, PocetLidiNaLuzko, PracNeschopen1000obyvatel	0.865	0.805	0.0006	✓	✓

Tabulka 5: Výsledky multivariantní analýzy pro modelování Umrti1000obyvatel

Pro zajištění nejvhodnějších prediktorů pro náš model jsme udělali feature selection pomocí metody *LASSO* a *Stepwise Selection*. Obě metody nám vyloučily *ProcentoPojistenych*, *PrumernyPlat* a *NakladyNaCloveka*. Vylepšený model *Hospitalizace1000obyvatel* byl tedy sestaven z následujících proměnných *Doktor1000obyvatel*, *Sestra1000obyvatel*, *PocetLidiNaLuzko*, *PracNeschopen1000obyvatel* a dokázal vysvětlit 86.5 % variability dat. Navíc došlo ke zvýšení Adjusted R-squared na 0.805 a model získal vysokou statistickou významnost (*p-value* 0.0006) při zachování všech předpokladů pro lineární regresi.

```
lm(formula = Hospitalizace1000obyvatel ~ Doktor1000obyvatel +  
Sestra1000obyvatel + PocetLidiNaLuzko + PracNeschopen1000obyvatel,  
data = promenne)
```

Model pracuje podle rovnice:

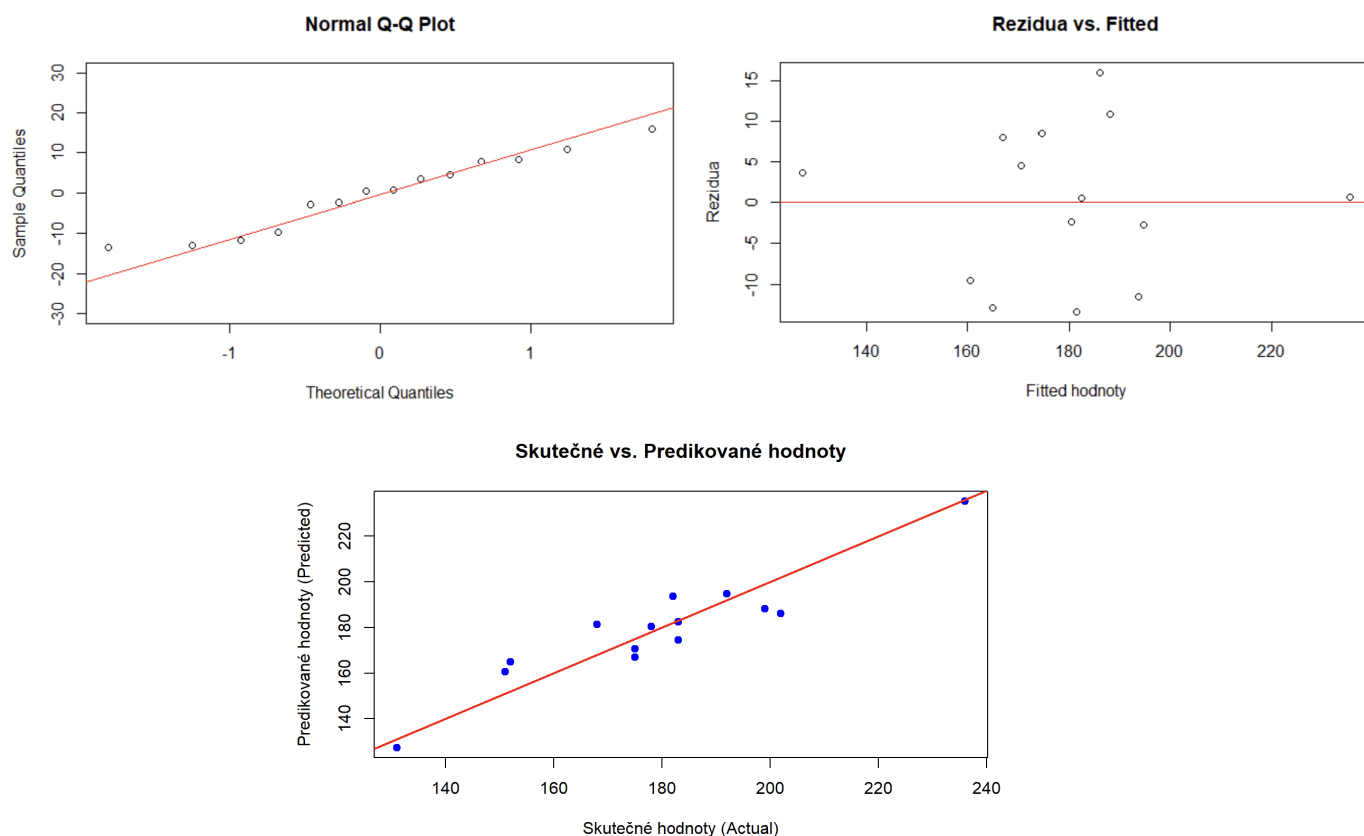
$$\text{Hospitalizace1000obyvatel} = 192.74017 + 3.09067 \cdot \text{Doktor1000obyvatel} + \\ + 3.90616 \cdot \text{Sestra1000obyvatel} - 0.29567 \cdot \text{PocetLidiNaLuzko} + \\ + 0.04383 \cdot \text{PracNeschopen1000obyvatel}$$

která říká, že čím více doktorů a sester je k dispozici, tím více lidí bude hospitalizováno, přičemž počet sester má větší vliv na počet hospitalizovaných, než počet doktorů. To opět naznačuje, že vyšší personální kapacita nemocnic umožňuje vyšší počet hospitalizovaných pacientů. Není překvapivé, že čím více lidí

bude na pracovní neschopence, tím více jich bude i hospitalizovaných (i když příspěvek není tak výrazný). Naopak, čím více lidí připadá na lůžko (tj. čím méně lůžek bude dostupných), tím méně lidí bude hospitalizovaných. Tento efekt je pravděpodobně způsoben neschopností nemocnic fyzicky pojmout větší počet pacientů. Může tedy dojít k tomu, že někteří nemocní, u kterých by bylo vhodné pozorování, se budou léčit doma a nebude možné efektivně reagovat na změnu jejich stavu.

7.2.1. Testování multivariantního modelu

Kvůli malému počtu vzorků, jsme pro testování modelu opět použili *Leave-One-Out Cross Validation*. Výsledky testování ukázaly: RMSE = 12.99, R-squared = 0.723, MAE = 11.38. Když si vytvoříme *predikce* podle našeho modelu (Obrázek 12), získáme RMSE = 8.98 a R-squared = 0.865. I když při křížové validaci došlo k lehkému zhoršení oproti prvotním výpočtům, lze celkově říci, že model je silný a poskytuje dobré predikce. Obrázek 12 ukazuje vizuální reprezentaci testování. V případě grafu pro *Normal Q-Q plot*, jsou body rozloženy podél přímky, ale v krajních hodnotách se odchylují. To naznačuje, že rozdělení reziduí nemusí být zcela normální (i když *Shapiro-Wilk* potvrdil normalitu). V grafu *Rezidua vs. Fitted* jsou body rozptýlené náhodně kolem nulové přímky bez zřetelných vzorců, což souhlasí s homoskedasticitou. Graf *Skutečné vs. Predikované* hodnoty ukazuje výkonnost modelu. Body jsou rozmístěny v blízkosti diagonální přímky. Zároveň nepozorujeme výrazné nadhodnocení nebo podhodnocení, což naznačuje nepřítomnost systematických chyb. Na základě těchto tří grafů můžeme konstatovat, že navržený lineární regresní multivariantní model je celkově vhodný pro naše data.



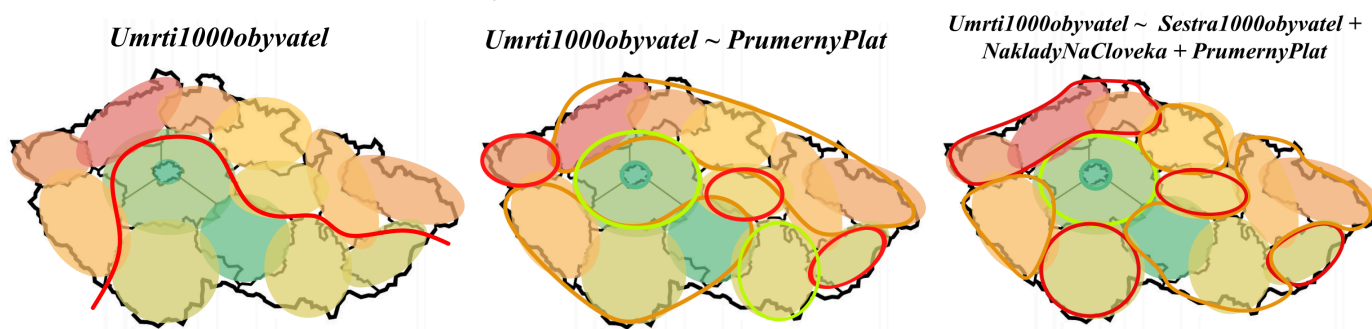
Obrázek 12: Q-Q plot a Rezidua vs. Fitted graf pro multivariantní model *Hospitalizace1000obyvatel* a graf *Skutečné vs. Predikované* hodnoty ukazující výkonnost

I tak stále existují mezery ve vysvětlení variability, což je očekávané vzhledem k povaze zdravotnických dat, která mohou být ovlivněna celou řadou faktorů. Tento výsledek naznačuje, že existují jiné, nelineární či skryté proměnné, které mohou dále ovlivnit výsledky hospitalizací.

8. Shlukování a analýza hlavních komponent

Naším dalším cílem je pokusit se identifikovat pomocí našich modelů vzory v regionech a porovnat je s výsledky podle závislých proměnných. V případě modelů o více prediktorech jsme použili *analýzu hlavních komponent* (PCA) pro snížení dimenze do dvoudimenzionálního prostoru. Jelikož máme proměnné s daty v různých jednotkách, použili jsme standardizovanou PCA metodu, která zajistí, že všechny proměnné budou mít stejnou váhu při výpočtu hlavních komponent. Dále jsme aplikovali metodu *k-means* pro identifikaci shluků mezi regiony, přičemž optimální počet shluků byl určen pomocí *Elbow Method*. Ve většině případů se ukázalo nejvýhodnější rozdělení do čtyř shluků.

8.1. Analýza hlavních komponent a shlukování krajů podle úmrťí v nemocnicích na 1000 obyvatel

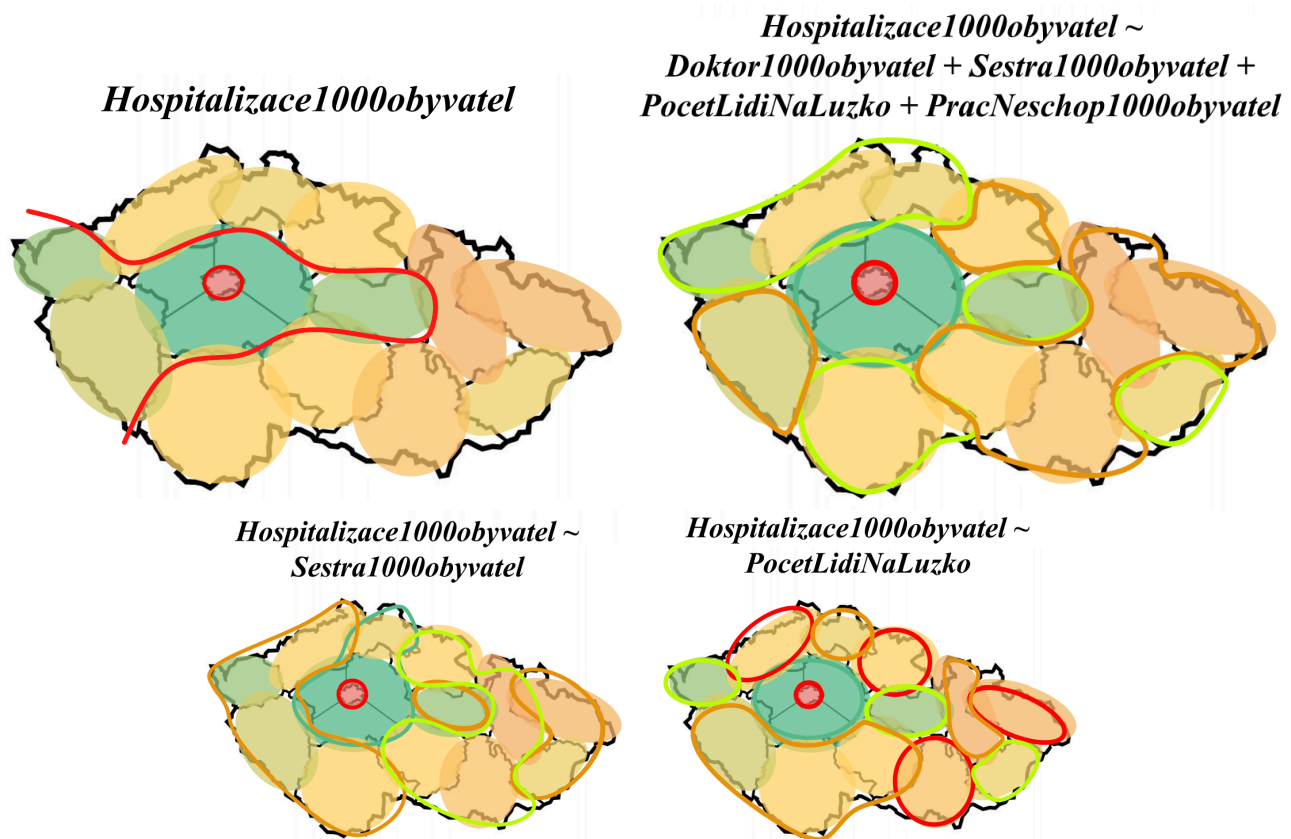


Obrázek 13: Výsledky shlukování pro *Umrti1000obyvatel*

Z hlediska míry úmrťí na 1000 obyvatel (*Umrti1000obyvatel*) lze kraje rozdělit do dvou regionů: severního a jižního. Jižní oblast republiky vykazuje nižší úmrtnost (viz Obrázek 13). Tuto skutečnost lze přisuzovat mnoha faktorům. Nejprve jsme se ji pokusili vysvětlit pomocí univariální analýzy, která jako nejvýznamnější proměnnou identifikovala *PrumernyPlat*. Následně jsme použili multivariální model, jehož dimenzi jsme pro lepší grafickou prezentaci snížili metodou *PCA*. Díky této redukci jsme zachovali 97.5 % variability původních dat. Po provedení *k-means* jsme rozdělili kraje do 4 skupin, které jsme porovnali s *Umrti1000obyvatel*. Vidíme, že v obou případech byly kraje PHA a STC zařazeny mezi oblasti s nižší úmrtností. V rámci univariální analýzy se k nim přidal také kraj JHM. Ostatní kraje byly zařazeny mezi oblasti s vyšší úmrtností. Ačkoli severní kraje byly vždy správně identifikovány jako regiony s vyšší úmrtností, tento výsledek byl spíše důsledkem toho, že většina krajů byla obecně označena jako "horší". Lze tedy říci, že pomocí univariálního ani multivariálního modelu nelze smysluplně rozdělit kraje do skupin pro hlubší interpretaci.

8.2. Analýza hlavních komponent a shlukování krajů podle počtu lidí hospitalizovaných v regionu na 1000 obyvatel

Z Obrázku 14 je patrné, že nejvíce hospitalizovaných lidí je zaznamenáno v Praze (červená barva). To může být způsobeno tím, že pacienti ze Středočeského kraje, ale i z dalších okolních regionů, jsou v případě závažnějších problémů převážně často do pražských nemocnic. Této myšlence napovídá i skutečnost, že právě Středočeský kraj má nejnižší míru hospitalizace (tmavě zelená barva), zatímco po Praze mají nejvyšší míru hospitalizace východní kraje. Na základě míry hospitalizace lze kraje rozdělit na dvě skupiny, které jsou naznačeny červenou rozděľující linií. U krajů KVK, PLK, STC a PAK (zkratky v Tabulce 1) je míra hospitalizace na 1000 obyvatel oproti zbytku republiky poměrně nízká.



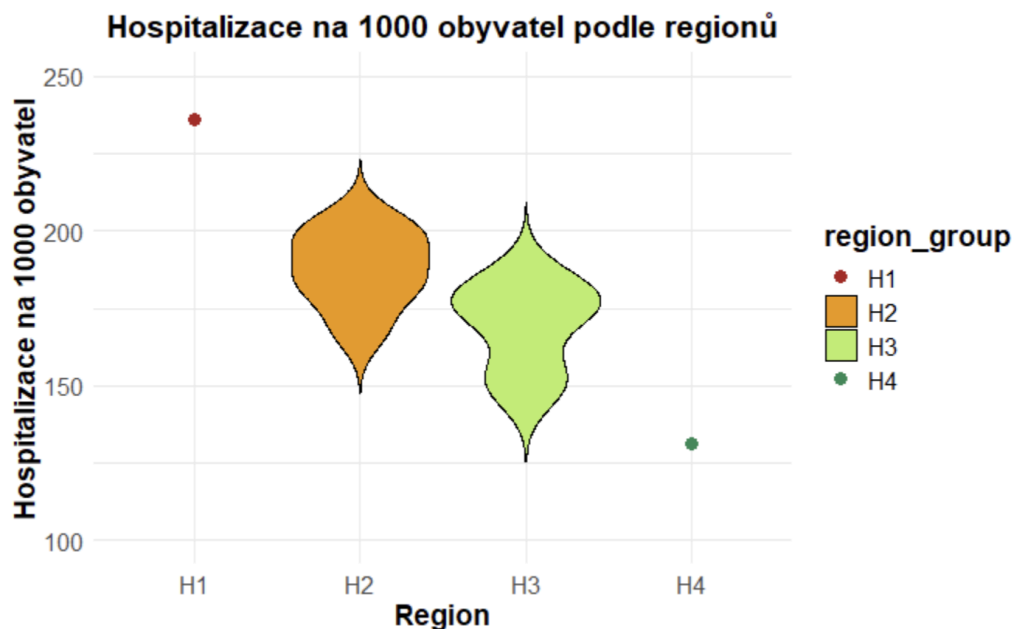
Obrázek 14: Výsledky shlukování pro *Hospitalizace1000obyvatel*

V pravé horní části Obrázku 14 lze vidět rozdělení krajů na základě multivariantního modelu pro *Hospitalizace1000obyvatel* z bodu 7.2., vytvořené pomocí *PCA* a *k-means*. Ve spodní části obrázku pak vidíme shluky, které vznikly z univariátních modelů (*Sestra1000obyvatel* a *PocetLidiNaLuzko*). Obecně lze usoudit, že ve všech modelech jsou jisté podobnosti.

V případě multivariantního modelu pro *Hospitalizace1000obyvatel*, se při redukci pomocí *PCA* do dvoudimenzionálního prostoru zachovalo 94.6 % vysvětlené variability dat. *K-means* následně rozdělilo kraje do čtyř skupin. PHA a STC byly správně zařazeny do samostatných “extrémních” skupin. Zbytek krajů byl rozdělen na dvě skupiny a při bližším pohledu vidíme, že poměrně přesně. Za výraznější chybu se dá považovat to, že *k-means* určilo PLK do horší skupiny než JHC, přestože PLK vykazuje menší míru hospitalizací.

Univariátní modely se zaměřují na vliv jednotlivých proměnných na počet hospitalizací izolovaně. Z Obrázku 14 je patrné, že ačkoliv došlo ke správnému určení některých krajů, celková úspěšnost nedosahuje úspěšnosti multivariantního modelu. To potvrzuje význam analýzy kombinující více faktorů pro dosažení přesnějšího rozdělení.

Zajímavostí je, že multivariantní model pro *Hospitalizace1000obyvatel* označil shodně regiony jako multivariantní model pro *Umrtil000obyvatel*. Výsledky *ANOVA* testu ukázaly, že existuje statisticky významný rozdíl mezi čtyřmi skupinami regionů (zobrazené pomocí *violin plot* v Obrázku 15) získaných z modelu *Hospitalizace1000obyvatel ~ Doktor1000obyvatel + Sestra1000obyvatel + PocetLidiNaLuzko + PracNeschop1000obyvatel* s *p-value* 0.001. Model ukazuje, že kombinace těchto proměnných dokáže efektivně rozlišit regiony podle míry hospitalizace.



Obrázek 15: Violin graf pro regiony vůči *Hospitalizace1000obyvatel*

9. Závěr

Cílem této práce bylo identifikovat klíčové faktory a na jejich základě vytvořit model pro mapování kvality regionálního zdravotnictví v České republice. Ačkoli je hodnocení kvality zdravotnictví komplexní problém, rozhodli jsme se hodnotit jej na základě míry úmrtnosti v nemocnicích a dodatečně podle míry hospitalizace. Pro modelování těchto proměnných jsme využili univariantsní a multivariantsní analýzu s prediktory z různých odvětví. Nejvíce nás zajímal vliv platového ohodnocení doktorů a zdravotních sester na míru úmrtí v nemocnicích (*Umrsti1000obyvatel*). Bohužel jsme nedokázali žádnou statisticky významnou souvislost se závislou proměnnou - tedy nezdá se, že by platové ohodnocení mělo vliv na míru úmrtí (ani na míru hospitalizace). Což naznačuje, že zdravotnický personál vykonává zodpovědnou práci i bez ohledu na míru finančního ohodnocení. Také se ukázalo, že nelze najít jednoduchý vztah mezi mírou úmrtí a hospitalizace s proměnnými jako je hodnocení kvality života (*KvalitaZivota*) a subjektivní hodnocení zdraví obyvatel (*ZdravotniStavObyvatel*). Celkově se ukázalo, že pro modelování těchto zdravotních ukazatelů je vhodnější multivariantsní analýza, která umožňuje zohlednit kombinovaný vliv faktorů a přispívá k přesnějším modelům. Ačkoli jsme měli k dispozici data splňující předpoklady pro lineární regresi, naše analýza byla zatížena malým počtem vzorků, což mohlo ovlivnit spolehlivost a statistickou významnost výsledků.

Na míru úmrtí měl jako samostatný prediktor největší vliv průměrný plat obyvatel v regionu, avšak nejednalo se o příliš silný model. Z vybraných prediktorů nešel sestavit statisticky významný multivariantsní model pro *Umrsti1000obyvatel* (nejlépe si však vedl model z prediktorů: *Sestra1000obyvatel*, *NakladyNaCloveka*, *PrumernyPlat*), což ukazuje na vliv kombinace socioekonomických faktorů, ale zároveň na potřebu identifikace komplexnějších faktorů, které jsme neměli k dispozici. Mnohem lépe byl sestaven model pro hospitalizaci, která měla mnoho vhodných prediktorů i v univariantsní analýze. Nejvíce variability v datech (86.5 %) *Hospitalizace1000obyvatel* vysvětlil model se statistickou významností (p-value 0.0006) s prediktory *Doktor1000obyvatel*, *Sestra1000obyvatel*, *PocetLidiNaLuzko* a *PracNeschopen1000obyvatel*. Počet zdravotnického personálu a počet dostupných lůžek zvyšuje míru hospitalizace, což ukazuje na roli dostupnosti zdravotní péče v počtu přijatých pacientů.

Na závěr jsme rozdělili kraje do skupin na základě našich modelů, což nám umožnilo provést hlubší interpretaci. Nebylo však překvapením, že modely pro *Umrți1000obyvatel* nedokázaly dobře identifikovat skupiny v krajích oproti závislé proměnné, která rozdělila zemi na severní část (s vyšší úmrtností) a jižní část (s nižší úmrtností). Oproti tomu si dobře vedl multivariantní model pro *Hospitalizace1000obyvatel*, který kraje rozdělil do skupin podobně jako závislá proměnná. Tyto skupiny měly mezi sebou statisticky významné rozdíly. Praha byla v samostatné extrémní skupině, jelikož míra hospitalizací je zde nejvyšší, kdežto u krajů v okolí Prahy je míra hospitalizace výrazně nižší. Naopak vyšší míra hospitalizace je pak na východě země. Toto rozdělení ukazuje na zvýšenou zátěž vůči Praze. Bylo by tedy vhodné posilnit nejen personální a materiální zdravotnické kapacity v hlavním městě, ale také pokusit se mu odlehčit podpořením zdravotnictví v okolních krajích.

Mezi kraji existují značné rozdíly a strategie rozvoje zdravotnictví by měla zahrnovat proporcionální rozdělení zdravotnických zdrojů mezi kraje a zaměřit se na zlepšení dostupnosti zdravotní péče (dostupnost lůžek) a zajištění odpovídajících kapacit ve zdravotnických zařízeních (počet zdravotních sester a doktorů).

10. Team Report

Každý člen týmu se aktivně podílel na vypracování tohoto reportu, sestavení datasetu, analýze a interpretaci dat a prezentaci práce.

Vendula Hrnčířová 30%

Milan Španko 25%

Matěj Dynda 25%

Stepan Gorichev 20%

11. Reference

- [1] [Fórum: Má Česko dostatek lékařů? | MT](#)
- [2] [Nízké finanční ohodnocení a nedostatek personálu trápí české zdravotníky nejvíce – mediJob](#)
- [3] [Novinka od 1. 1. 2024 – služby ve zdravotnictví až 24 hodin!](#)