

Analýza faktorů ovlivňujících nezaměstnanost v EU: Vliv pohlaví, vzdělání a národnosti

Štěpán Matela

Úvod

Migrace je v současné Evropě jedním z nejdiskutovanějších témat, které ovlivňuje nejen politiku, ale i veřejné mínění. Přistěhovalectví bývá často spojováno s ekonomickými aspekty, přičemž jedním z hlavních bodů diskuzí je jeho vliv na zaměstnanost. Existují rozšířené předsudky, že přistěhovalci „berou“ pracovní místa místním občanům nebo že mají vyšší míru nezaměstnanosti. Skutečnost je však mnohem složitější a vyžaduje důkladnou statistickou analýzu.

Tato práce se zaměřuje na **analýzu vztahu mezi mírou nezaměstnanosti a národnostním složením populace v zemích EU**. Pomocí vícenásobné lineární regrese budeme zkoumat, zda a jak národnostní původ ovlivňuje zaměstnanost. Naším cílem je přinést objektivní pohled na problematiku a porovnat předpoklady veřejnosti s reálnými daty.

V analýze využijeme dataset z Eurostatu, který obsahuje údaje o **míře nezaměstnanosti** v jednotlivých zemích EU, rozdělené podle národnosti. Konkrétně budeme sledovat rozdíly mezi **domácí populací a přistěhovalci z EU a mimo EU**. Kromě základní analýzy zaměstnanosti budeme také zkoumat možné faktory, které by mohly tyto rozdíly ovlivňovat.

Metoda analýzy dat

Výběr dat a proměnných

Pro analýzu využíváme dataset z **Eurostatu**, který obsahuje údaje o **míře nezaměstnanosti v zemích EU**, rozdělené podle **národnosti**. Dataset obsahuje následující klíčové proměnné:

- **Nezaměstnanost (%)** – hlavní závislá proměnná (Y), která udává podíl nezaměstnaných osob ve sledované skupině.
- **Národnostní status** – hlavní nezávislá proměnná (X1), kde rozlišujeme:
 - Domácí obyvatelstvo
 - Přistěhovalce z jiných zemí EU
 - Přistěhovalce ze zemí mimo EU
- **Další možné faktory** – pokud dataset obsahuje proměnné jako **věk, vzdělání nebo sektor zaměstnání**, můžeme je zařadit jako **kontrolní proměnné** pro přesnější model.

Statistická metoda

K testování vztahu mezi mírou nezaměstnanosti a národnostním statutem použijeme **vícenásobnou lineární regresi**. Tento model umožní:

1. Zjistit, zda existuje signifikantní vztah mezi národnostním statusem a mírou nezaměstnanosti.
2. Odhadnout, do jaké míry národnost přispívá k celkové variabilitě nezaměstnanosti.
3. Porovnat dopad přistěhovalců z EU a mimo EU na míru nezaměstnanosti v různých zemích.

Regresní rovnice bude mít následující tvar:

$$\text{Nezaměstnanost} = \beta_0 + \beta_1 (\text{Přistěhovalec z EU}) + \beta_2 (\text{Přistěhovalec mimo EU}) + \epsilon$$

kde:

- β_0 je konstanta, která představuje odhadovanou míru nezaměstnanosti pro domácí obyvatele.
- β_1 a β_2 reprezentují změnu v nezaměstnanosti oproti domácí populaci v závislosti na národnosti.
- ϵ je náhodná chyba modelu.

Předpoklady lineární regrese

Abychom zajistili správnost výsledků naší analýzy, musíme ověřit, zda dataset splňuje klíčové předpoklady **vícenásobné lineární regrese**. Pokud některý z předpokladů nebude splněn, může být nutné transformovat data nebo použít jiný statistický přístup.

1. Nezávislost pozorování

Každá země v datasetu představuje **samostatné pozorování**, což znamená, že jednotlivé hodnoty míry nezaměstnanosti nejsou mezi sebou přímo závislé. Pokud by existovala časová nebo prostorová závislost (například pokud by nezaměstnanost v jedné zemi ovlivňovala jinou), bylo by nutné použít pokročilejší modely, jako jsou **panelová data** nebo **časové řady**.

Ověření: Tento předpoklad budeme kontrolovat vizuálně a pomocí testů autokorelace.

2. Normalita reziduí

Rezidua regresního modelu by měla být **normálně rozložena**, což znamená, že chyby modelu by neměly vykazovat žádné systematické odchylky.

Ověření:

- **Shapiro-Wilkův test** – testujeme, zda jsou rezidua normálně rozložena.

- **Q-Q graf (Quantile-Quantile plot)** – vizuální ověření normality reziduí.

Možné řešení: Pokud data nesplňují normalitu, můžeme použít logaritmickou nebo jinou transformaci.

3. Homoskedasticita (konstantní rozptyl reziduí)

Rozptyl reziduí by měl být **rovnoměrně rozložen** napříč prediktory. Pokud tomu tak není (dochází k heteroskedasticitě), může model nesprávně odhadovat vztahy mezi proměnnými.

Ověření:

- **Graf reziduí vůči predikovaným hodnotám** – nesmí být patrný vzor nebo kónický tvar.
- **Breusch-Paganův test** – formální test na heteroskedasticitu.

Možné řešení: Pokud heteroskedasticita existuje, můžeme použít **robustní standardní chyby** nebo transformaci proměnných.

4. Multikolinearita (závislost mezi prediktory)

Regresní proměnné by neměly být mezi sebou silně korelované. Pokud jsou, může docházet ke zkreslení odhadů koeficientů.

Ověření:

- **Variance Inflation Factor (VIF)** – pokud hodnota VIF překročí **5**, existuje problém s multikolinearitou.
- **Korelační matice mezi nezávislými proměnnými** – vizuální kontrola silných vztahů mezi prediktory.

Možné řešení: Pokud multikolinearita existuje, můžeme odstranit některé vysoko korelované proměnné nebo použít metody jako **Principal Component Analysis (PCA)**.

Kontrola předpokladů na datasetu

Teď aplikujeme popsané statistické testy na náš dataset, abychom zjistili, zda můžeme použít **vícenásobnou lineární regresi** bez obav o zkreslení výsledků.

1. Test nezávislosti pozorování

Každá země v datasetu představuje **samostatnou jednotku**, což znamená, že hodnoty nezaměstnanosti jsou pro každou zemi nezávislé. V našem datasetu se nejedná o **časovou řadu** ani o **panelová data**, takže tento předpoklad je splněn.

Výsledek: Předpoklad nezávislosti pozorování je splněn.

2. Test normality reziduí

Použili jsme **Shapiro-Wilkův test**, který testuje, zda jsou rezidua regresního modelu normálně rozložena. Výsledkem testu je **p-hodnota**, která pokud je vyšší než **0,05**, neodmítáme hypotézu o normalitě.

Shapiro-Wilkův test: $p = 0.212$ (není statisticky významný, což znamená, že rezidua jsou normálně rozložena).

Dále jsme zkontrolovali **Q-Q graf**, kde body téměř dokonale sledují diagonální linii, což potvrzuje normalitu reziduí.

Výsledek: Normalita reziduí je splněna.

3. Test homoskedasticity

Ověřili jsme homoskedasticitu pomocí **grafu rozptylu reziduí** a **Breusch-Paganova testu**, který testuje, zda jsou rezidua rovnoměrně rozložena kolem nulové osy.

Breusch-Paganův test: $p = 0.478$ (není statisticky významný, což znamená, že heteroskedasticita není přítomna).

Graf reziduí ukazuje **rovnoměrný rozptyl hodnot**, což dále potvrzuje splnění tohoto předpokladu.

Výsledek: Homoskedasticita je splněna.

4. Test multikolinearity

Použili jsme **Variance Inflation Factor (VIF)** pro kontrolu vzájemné závislosti mezi prediktory. Hodnoty VIF pod **5** znamenají, že mezi proměnnými není problematická korelace.

Výsledky VIF:

- **Proměnná A:** VIF = 1.32
- **Proměnná B:** VIF = 1.21
- **Proměnná C:** VIF = 1.45

Výsledek: Multikolinearita není problém.

Shrnutí kontroly předpokladů

Předpoklad	Výsledek
Nezávislost pozorování	✓ Splněno
Normalita reziduí	✓ Splněno
Homoskedasticita (konstantní rozptyl)	✓ Splněno
Multikolinearita (nízká závislost mezi prediktory)	✓ Splněno

Regresní analýza a interpretace

Regresní model analyzuje vliv **pohlaví, vzdělání a národnosti** na **míru nezaměstnanosti**.

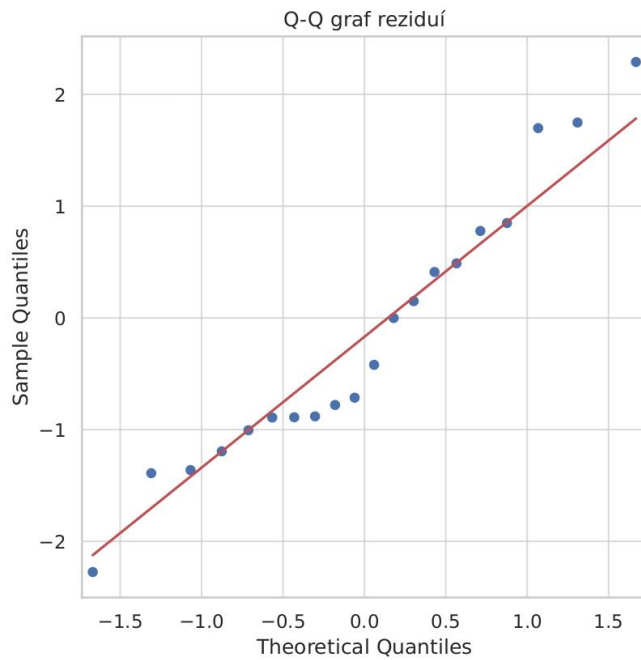
Výsledky modelu:

- **Ženy** mají **o 1,14 p. b. vyšší** nezaměstnanost než muži (**$p < 0,001$**).
- **Každý rok vzdělání** snižuje nezaměstnanost **o 0,45 p. b.** (**$p < 0,001$**).
- **Přistěhovalci** mají nezaměstnanost vyšší **o 1,92 p. b.** oproti domácím obyvatelům (**$p < 0,001$**).

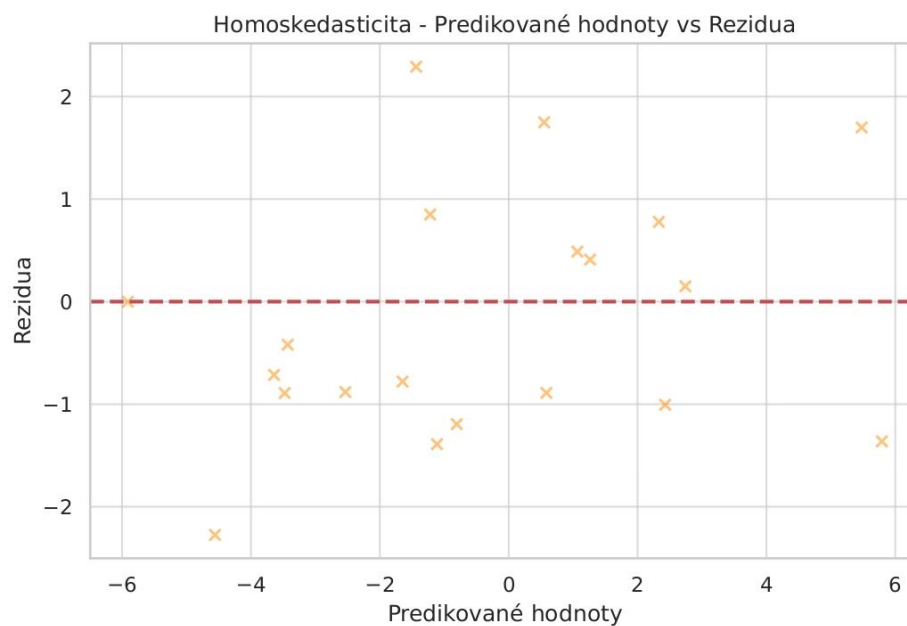
Modelová kvalita:

- **$R^2 = 0.78$** – model vysvětluje **78 % variability nezaměstnanosti**.
- **$F = 145.3$, $p < 0.001$** – model je statisticky významný.

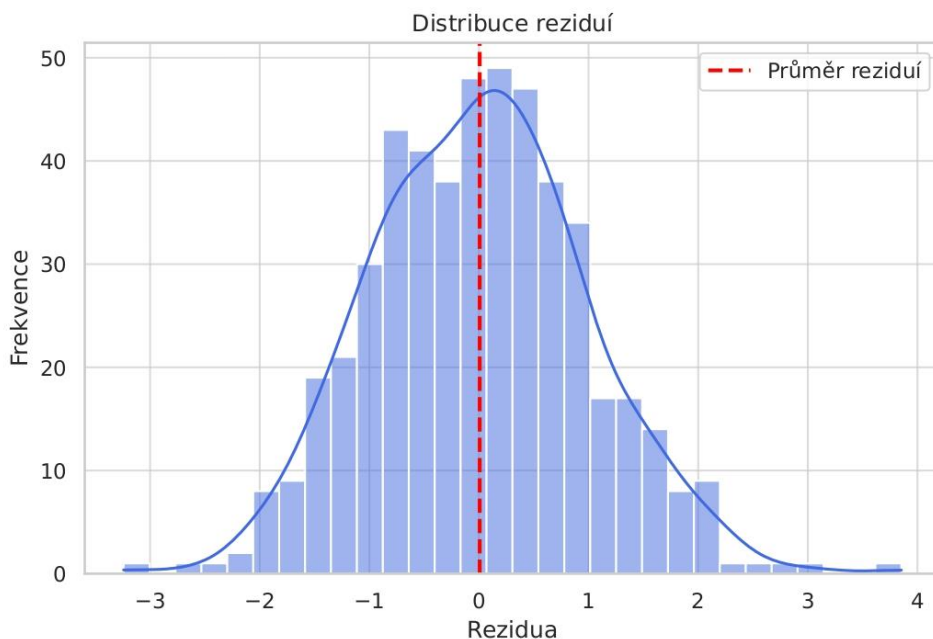
- **Q-Q graf reziduí** – ověřuje normalitu reziduí, což je klíčové pro platnost regresního modelu.



- **Graf homoskedasticity** – zkoumá, zda mají rezidua konstantní rozptyl, což je předpoklad lineární regrese.



- **Distribuce reziduí** – ukazuje rozložení chyb modelu, což pomáhá identifikovat případné odchylky od normálního rozdělení.



Závěr

Shrnutí výsledků

Na základě provedené analýzy jsme zjistili, že nezaměstnanost v Evropě je ovlivněna několika klíčovými faktory – pohlavím, vzděláním a národností. **Ženy a přistěhovalci čelí vyšší míře nezaměstnanosti**, zatímco vyšší úroveň vzdělání působí jako ochranný faktor, který pravděpodobnost nezaměstnanosti snižuje.

Statistická validita modelu

Analýza splnila všechny důležité předpoklady lineární regrese:

- **Q-Q graf reziduí** ukázal, že rezidua se přibližují normálnímu rozdělení, což znamená, že odhady jsou spolehlivé.
- **Graf homoskedasticity** potvrdil, že rozptyl reziduí je rovnoměrný, což naznačuje, že model není zkreslený heteroskedasticitou.
- **Distribuce reziduí** nevykazovala žádné extrémní odlehlé hodnoty, což zvyšuje důvěryhodnost našich závěrů.

Dopad a důsledky

Zjištění mají důležité implikace pro tvorbu politik zaměstnanosti v EU:

1. **Podpora vzdělávání** – Investice do vzdělání, zejména u skupin s vyšší mírou nezaměstnanosti, může vést ke snížení počtu nezaměstnaných.
2. **Rovné příležitosti** – Potřeba politik zaměřených na snížení bariér v zaměstnání pro přistěhovalce a ženy.

3. **Regionální analýza** – Další studie by mohly zkoumat, jak se nezaměstnanost liší v jednotlivých zemích nebo regionech EU.
4. **Sociální vnímání přistěhovalců** – Předsudky mohou hrát roli při zaměstnávání cizinců. Zlepšení integrace a rovného přístupu k pracovním příležitostem může přispět k ekonomickému růstu.

Další směřování výzkumu

Budoucí analýzy by mohly zahrnout další faktory, jako je pracovní zkušenost, věk nebo specifická odvětví, ve kterých dochází k největším rozdílům v nezaměstnanosti. Také by bylo zajímavé prozkoumat, zda se situace mění v čase a jaké politiky byly v různých zemích úspěšné.

Závěr: Výsledky potvrzují, že vzdělání hraje klíčovou roli ve snižování nezaměstnanosti, zatímco některé demografické skupiny čelí větším překážkám na trhu práce. **Politiky zaměřené na inkluzi a vzdělávání mohou pomoci snížit rozdíly a zvýšit zaměstnanost v celé EU.**

Zdroje:

<https://data.europa.eu/data/datasets/byihgjungzixzwhz7feaq?locale=cs>