VÍCEROZMĚRNÝ PRAVDĚPODOBNOSTNÍ MODEL ROZDĚLENÍ PŘÍJMŮ ČESKÝCH DOMÁCNOSTÍ*

Ivana Malá, Vysoká škola ekonomická v Praze

doi: 10.18267/j.polek.1040

Úvod

Sledování velikosti a vývoje příjmů a mezd je základní součástí úvah o ekonomické situaci států a jejím vývoji. Podrobné údaje o příjmech (mzdách) jsou východiskem pro analýzy a úvahy expertů v různých oblastech a správních orgánech, stejně tak jsou v centru zájmu veřejnosti. Příjmy domácností, analyzované v tomto textu, jsou odrazem sociálně ekonomického vývoje společnosti, ale také silně ovlivňují její stav, neboť ovlivňují chování domácností jako základních ekonomických a sociálních subjektů. Sledování vývoje příjmů a výdajů domácností umožňuje analyzovat situaci, se kterou se tyto společně hospodařící skupiny jednotlivců setkávají, a na jejímž základě plánují své krátkodobé i dlouhodobé chování. Znalost pravděpodobnostního rozdělení příjmů (nejen pouze jejich výše nebo proměnlivosti) umožňuje odhadovat další charakteristiky a hledat odpovědi na otázky, které jsou (nebo mohou být) kladeny. Na velikosti příjmů domácností závisí také různé charakteristiky nebo indexy, které se snaží kvantifikovat kvalitu života nebo charakterizují chudobu či aspoň její nebezpečí (ohrožení domácnosti pádem do chudoby). Proto je hledání vhodných modelů pro příjmy stále aktuálním problémem, ke kterému je v bohaté literatuře věnované tématu přistupováno nejrůznějšími způsoby. Neméně důležitým problémem, který je třeba řešit, je konstrukce spolehlivých předpovědí budoucího vývoje příjmů, které by bylo možné využít pro odhady spotřeby domácností nebo třeba daňového výnosu.

Údaje charakterizující velikost a vývoj příjmů českých domácností pro jednotlivé roky, stejně jako jejich analýzy z nejrůznějších pohledů, publikují pravidelně Český statistický úřad (ČSÚ, 2014) a ministerstvo práce a sociálních věcí (MPSV, 2014). V literatuře je možné najít velké množství prací, které se příjmy a jejich rozdělením zabývají, jejich přehled přesahuje možnosti tohoto textu. Existuje také množství statistických metod a postupů, které mohou pomoci při analýze dat týkajících se příjmů a poskytnout informace, které by podpořily rozhodování expertů různých ekonomických a společenských zaměření. Směs rozdělení je použita pro modelování příjmů (a jejich nerovnosti) ve Velké Británii v práci Flachaire, Nunez (2007). Kromě sledování příjmů v jednotlivých zemích lze nalézt také studie, zabývající se srovnáním příjmů v různých státech nebo regionech, například studie Pittau, Zelli (2006).

* Výzkum byl podpořen projektem IP 400040 v rámci institucionální podpory vědy Fakulty informatiky a statistiky Vysoké školy ekonomické v Praze.

Z prací týkajících statistických modelů rozdělení mezd nebo příjmů v České republice a používajících jednorozměrné logaritmicko-normální rozdělení nebo model konečných směsí rozdělení uveďme aspoň Bílková (2012), Bartošová (2006), Malá (2013) nebo Marek (2010). Jiný přístup (ekonometrický) je zvolen v práci Večerník (2013). Obdobný model vícerozměrného normálního rozdělení jako v předkládané studii byl (pro příjmy z let 2005–2008) publikován v článku Bartošová, Longford (2014). Článek obsahuje analýzu ekvivalizovaného příjmu podle modifikované metodiky Organizace pro hospodářskou spolupráci a rozvoj (dále OECD) a zabývá se konstrukcí směsi ve shodném modelu, stejně jako tento text.

Rozdělení příjmů (jakéhokoliv typu) v populaci je velmi nehomogenní. Model předkládaný v tomto textu vychází z předpokladu, že v silně nehomogenní populaci domácností je možné nalézt homogennější podmnožiny (dále komponenty), příjmy uvnitř těchto komponent zkoumat odděleně a nakonec rozdělení příjmů pro všechny domácnosti dohromady konstruovat pomocí váženého průměru komponentních s vhodnými vahami, které by odrážely zastoupení jednotlivých komponent v celé populaci. Takový model poskytuje informace o jednotlivých komponentách a jejich vývoji, o zastoupení jednotlivých komponent mezi všemi domácnostmi a také lepší model příjmů pro všechny domácnosti, než by poskytlo jedno jediné rozdělení. Jednou z možností je použít pro tvorbu jednotlivých složek známé (pozorované) údaje, jako vzdělání osoby v čele domácnosti (případě partnera nebo partnerky), bydliště, zaměstnání (osoby v čele nebo partnera či partnerky), například Malá, 2013. Pokud jsou takové vysvětlující proměnné vhodně zvoleny, může model směsi poskytnout vylepšení proti jednodušším modelům. V tomto textu ovšem budeme předpokládat, že příslušnost domácností do jednotlivých komponent nelze pozorovat (není známá). Cílem je posoudit, zda lze opravdu takové podmnožiny nalézt (při volbě přibližně normálního rozdělení pro logaritmy analyzovaných příjmů) a zda výsledný model je lepší než v případě jednoho rozdělení, a dále posoudit, zda je takový model schopen konkrétní domácnosti do jednotlivých komponent zařadit (klasifikovat). Poslední cíl je vlastně úlohou shlukové analýzy. Model je odhadnut pro ekvivalizované čisté příjmy domácností (podle metodiky OECD) a pozornost je věnována také jiným možnostem určení počtu ekvivalentních jednotek a závislosti postavení domácností ve výběru vzhledem k různým metodikám.

Předpokládejme, že analyzovaná data tvoří údaje o čtyřech následujících letech 2007–2010 (doba, po kterou domácnosti zůstávají v šetření Životní podmínky; ČSÚ, 2014) a že máme k dispozici domácnosti, které byly zařazeny do šetření ve všech sledovaných letech. Jedná se tedy o čtyři opakovaná pozorování příjmů a velkého počtu dalších charakteristik domácnosti. Jednou z možností, jak sledovat rozdělení příjmů v jednotlivých letech, je hledat model pro každý rok zvlášť, v takovém případě ovšem nevyužíváme znalosti individuálního vývoje příjmů jednotlivých domácností ve sledovaných letech. Proto je v tomto textu využito čtyřrozměrné normální rozdělení pro popis rozdělení logaritmů celkových čistých celkových ekvivalizovaných příjmů tak, aby bylo možné zohlednit sdílení příjmů a výdajů členy jedné domácnosti, ale také

skutečnost, že větší domácnost pro stejný životní standard potřebuje větší prostředky. Zkoumané příjmy je možné chápat jako příjem standardizované domácnosti, nebo také jako ekvivalizovaný příjem na jednoho člena. Vícerozměrný model samozřejmě poskytuje úplnou informaci o všech marginálních rozděleních, a tedy také o jednorozměrných rozděleních příjmů v jednotlivých letech.

Zmíněný ekvivalizovaný příjem není jedinou možností, jak lze zohlednit velikost a strukturu domácnosti. Další možností je uvažovat přímo počet členů nebo počet jednotek podle OECD, které také zohledňují počet členů a jejich věk, nicméně nedávají tak velkou váhu sdílení výdajů členy domácnosti jako modifikovaná metodika OECD. Vztahy mezi velikostí ekvivalizovaných příjmů určených podle těchto tří definic jsou zkoumány v první části textu. Neexistuje jedna všeobecně používaná metodika, která by zajišťovala stejný životní standard (EUROSTAT, 2014), metodika v tomto textu je shodná s metodikou v analýzách Eurostatu, dříve označovaná jako metodika podle Evropské Unie (dále EU).

Sledované čtyřleté období začíná po vstupu České republiky do EU v hospodářsky příznivých letech před ekonomickou krizí a pokračuje do roku 2010, kdy útlum způsobený krizí přecházel do recese. Proto ve sledovaných letech není možné předpokládat pozvolný plynulý vývoj (růst) očekávaný v ekonomicky příznivých letech (například Bartošová, Longford, 2014 pro roky 2005–2008). V letech 2006–2008 rostly mzdové příjmy (kolem 7 % ročně), situace na trhu práce byla velmi příznivá a panoval optimismus ekonomických subjektů i domácností vzhledem k budoucnosti. Ekonomická konjunktura vedla k velkému zvýšení prosperity domácností v České republice. Toto období (začínající již v roce 2004) bylo vystřídáno ekonomickou krizí a následujícím obdobím útlumu. V roce 2008 byla zavedena superhrubá mzda a rovná 15% daň z příjmu fyzických osob, byl upraven systém slev a odpočtů od základu daně a změněny podmínky pro pobírání dávek státní sociální podpory i jejich vyplácené částky. Rok 2009, krizový pro ekonomiku jako celek, neznamenal pro české domácnosti (na rozdíl od firemní sféry) významnou negativní změnu. Na české domácnosti měl takový vliv spíše další vývoj, který byl spojen s restriktivní hospodářskou politikou (SOCR, 2014).

1. Metodologie

Uvažujme tedy čtyřrozměrné náhodné vektory $\mathbf{X}$ (ekvivalizované celkové čisté roční příjmy českých domácností v letech 2007–2010 v Kč) a $\mathbf{Y} = \ln \mathbf{X}$ (vektor logaritmů příjmů v Kč). Budeme předpokládat, že náhodný vektor logaritmů ekvivalizovaných příjmů $\mathbf{Y}$ má (čtyřrozměrné) pravděpodobnostní rozdělení popsané směsí K (v tomto textu volíme $K = 2-4$) čtyřrozměrných normálních rozdělení, která budeme značit $N_4(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ kde $\boldsymbol{\mu}_j$ jsou vektory středních hodnot, $\boldsymbol{\Sigma}_j$ jsou kovarianční matice komponent, $j = 1, ..., K$. Hustoty komponentních rozdělení označíme $f_j(\mathbf{y}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = f_j(\mathbf{y}; \boldsymbol{\psi})$ kde vektor $\boldsymbol{\psi}$ obsahuje všechny neznámé parametry modelu. Hustota f modelu směsi (hustota rozdělení vektoru $\mathbf{Y}$) má v takovém případě tvar váženého aritmetického průměru hustot jednotlivých komponent s vahami π_j

$$f(\mathbf{y};\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j \frac{1}{\sqrt{(2\pi)^4 \left|\boldsymbol{\Sigma}_j\right|}} \exp\left\{-\left(\mathbf{y}-\boldsymbol{\mu}_j\right)' \boldsymbol{\Sigma}_j^{-1} \left(\mathbf{y}-\boldsymbol{\mu}_j\right)\right\}, \ \mathbf{y} \in R^4 . \quad (1)$$

Vektor $\boldsymbol{\psi}$ obsahuje $K-1$ vah π_j jednotlivých složek (K-tá váha je rovna doplňku součtu předcházejících vah do jedničky), K vektorů neznámých parametrů komponentních rozdělení obsahujících složky vektorů středních hodnot $\boldsymbol{\mu}_j$ (celkem $4K$ parametrů) a K kovariančních matic $\boldsymbol{\Sigma}_j$ (celkem $10K$ parametrů vzhledem k symetrii kovarianční matice). Pokud neklademe další omezení na tvar kovariančních matic, jedná se celkem o $15K-1$ parametrů.

Předpoklad o stejných kovariančních maticích, složek nebo dokonce diagonálních kovariančních maticích (to znamená těžko obhajitelný předpoklad nezávislosti logaritmů příjmů v jednotlivých letech) snižuje počet parametrů a výrazně zjednodušuje numerické hledání odhadů parametrů (podle McLachlan; Peel, 2000). V případě prezentovaného modelu s pouze malým počtem složek se tyto modely ukázaly být nevhodné, neboť nevystihovaly dobře analyzovaná data (v případě volby kovariančních matic nezávislých na komponentách i diagonálních kovariančních matic) nebo neposkytovaly informaci o předpokládané závislosti příjmů v jednotlivých letech (v případě volby diagonálních kovariančních matic). Dosažení dobré aproximace empirického rozdělení zmíněnými modely je možné, neboť teoreticky lze jakékoliv rozdělení aproximovat libovolně přesně i zmíněnými modely s omezením na kovarianční strukturu (McLachlan; Peel, 2000), bylo by ale nutné použít výrazně většího počtu komponent. V takovém případě by nebylo možné nalézt interpretaci pro komponenty modelu ani dostatečně přesně odhadnout parametry. Z vícerozměrného výrazu (1) lze nalézt jednorozměrná rozdělení logaritmů ročních příjmů jako směs normálních rozdělení, a tedy pro příjmy v jednotlivých komponentách dostáváme logaritmicko-normální rozdělení. Parametry těchto rozdělení lze vyčíst z (1) a jejich odhady (složky vektorů $\boldsymbol{\mu}_j$ a diagonální prvky kovariančních matic) jsou obsaženy v odhadu vektoru $\boldsymbol{\psi}$.

Pro $j = 1, \ldots, K$ označíme $\mathbf{Y}_j$ náhodné vektory s rozdělením $N_4(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Potom můžeme pro střední hodnotu a kovarianční matici logaritmu příjmů $\mathbf{Y}$ psát

$$E(\mathbf{Y}) = \sum_{j=1}^{K} \pi_j E\left(\mathbf{Y}_j\right) = \sum_{j=1}^{K} \pi_j \boldsymbol{\mu}_j \ \text{ a } \ \boldsymbol{\Sigma}_{\mathbf{Y}} = \sum_{j=1}^{K} \pi_j \boldsymbol{\Sigma}_j + \sum_{j=1}^{K} \pi_j \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{\mathbf{Y}}\right)\left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{\mathbf{Y}}\right)' . \quad (2)$$

Vektor středních hodnot vektoru $\mathbf{Y}$ je tedy opět váženým aritmetickým průměrem vektorů středních hodnot složek. Druhá část vzorce (2) je zobecněním známého vztahu pro rozklad rozptylu, který počítá celkový rozptyl náhodné veličiny z rozptylů vnitroskupinových a meziskupinových. Záleží tedy nejen na kovariančních maticích (první část vzorce pro $\boldsymbol{\Sigma}_{\mathbf{Y}}$ v (2)), ale také na odchylkách komponentních středních hodnot od celkové střední hodnoty (druhá část vzorce výrazu pro $\boldsymbol{\Sigma}_{\mathbf{Y}}$).

Pro nalezení maximálně věrohodného odhadu neznámých parametrů se používá numerický EM algoritmus (McLachlan; Peel, 2000, Boldea; Magnus, 2009). Přesnost odhadů je možné najít pomocí bootstrapu (McLachlan; Peel, 2000; Benagli *et al*., 2009). Všechny numerické postupy spojené s odhady pro data v řádu několika

tisíc pozorování umožňují použít asymptotické vlastnosti odhadů, jsou však časově i výpočetně velmi náročné a je třeba řešit numerické problémy s odhadem spojené. Pro porovnání jednotlivých modelů (pro různý počet komponent) je možné použít informační kritéria, která jsou schopna zohlednit kvalitu proložení dat a počet parametrů modelu. V tomto textu byly modely porovnány pomocí Akaikeova informačního kritéria, které je založeno na hodnotě logaritmické věrohodnostní funkce v nalezeném minimu a počtu parametrů.

Použitý model nepracuje s žádnými vysvětlujícími proměnnými (jako by mohlo být například vzdělání jednotlivých členů domácnosti, počet ekonomicky aktivních členů, počet dětí nebo velikost a poloha bydliště domácnosti), a není tedy známa (pozorována) příslušnost domácností ke složkám a dokonce ani počet těchto složek. Na základě odhadnutého modelu (odhadu $\hat{\boldsymbol{\psi}}$ parametru $\boldsymbol{\psi}$ je ovšem možné odhadnout pro každou hodnotu příjmu $\mathbf{y}$ posteriorní pravděpodobnosti $P\left(j\middle|\mathbf{y}\right),\ j=1,\ldots,K$ příslušnosti domácnosti do jednotlivých složek podle vzorce

$$P\left(j\middle|\mathbf{y}\right)=\frac{\pi_j f_j(\mathbf{y};\boldsymbol{\psi})}{f(\mathbf{y};\boldsymbol{\psi})}. \tag{3}$$

Tento postup spojuje předkládanou metodu se shlukovou analýzou, nahrazujeme jen slovo komponenta (složka) slovem shluk, též Bartošová, Longford (2014). Při shlukování byla domácnost zařazena do shluku s největší posteriorní pravděpodobností.

Všechny potřebné výpočty byly provedeny v programu *R* (RCORE, 2014), pro odhad neznámého vektoru parametrů vícerozměrných modelů byl použit EM algoritmus implementovaný v balíčku *MIXTOOLS* (MIXTOOLS, 2014; Benaglia *et al.*, 2009).

2. Data a výsledky

V následující analýze jsou použita data z výběrového šetření Životní podmínky (národní modul evropského šetření European Union Statistics on Income and Living Conditions; EU-SILC) prováděného Českým statistickým úřadem v letech 2008–2011. Tato šetření obsahují údaje o příjmech domácností z let 2007–2010, byly použity pouze údaje domácností, které byly zahrnuty do všech těchto čtyř šetření (data neobsahují chybějící pozorování). Cílem šetření Životní podmínky je dlouhodobě a pravidelně získávat srovnatelná data o sociální situaci domácností, která jsou díky jednotné metodice porovnatelná i s dalšími zeměmi EU (ČSÚ, 2014). V analýze bylo použito 4 873 domácností s údaji o příjmech a charakteristikách počtu osob a jejich věkové struktuře, obsažené v údajích o počtu členů a přepočtených ekvivalentních jednotkách podle modifikované metodiky OECD (dále označíme počet jednotek *msj*). Podle této metodiky má první dospělá osoba má váhu 1, další osoby starší 13 let 0,5 a děti do 13 let 0,3. V metodice podle OECD (počet jednotek označíme dále *sj*) má první dospělá osoba váhu 1, další osoby starší 13 let 0,7 a děti do 13 let 0,5). Ve zkoumaných letech 2007–2010 došlo v České republice k malému poklesu počtu ekvivalentních jednotek (pro počet členů domácnosti z 2,30 na 2,27, pro počet ekvivalentních modifikovaných

jednotek OECD z 1,58 na 1,56) podobně jako ve zkoumaných domácnostech (ČSÚ, 2014)).

Čistý ekvivalizovaný příjem domácností určený podle modifikované metodiky OECD, jehož rozdělení budeme modelovat, označíme *CPMSJ*, ekvivalizovaný příjem určený podle metodiky OECD označíme *CPSJ* a příjem na jednoho člena domácnosti (též per capita) *CPPC*. Všechny tyto různé typy ekvivalizovaných příjmů lze zapsat ve tvaru

čistý roční příjem domácnosti (*v Kč*)*/počet ekvivalentních jednotek*. (4)

Uvážíme-li definici ekvivalentních jednotek, platí $msj \leq sj \leq pc$ kde rovnosti je dosaženo pro jednočlenné domácnosti. Vzhledem k (4) pak platí

$$CPMSJ \geq CPSJ \geq CPPC.$$

Lze předpokládat, že kromě toho, že jsou ekvivalizované příjmy takto seřazeny podle velikosti, jsou všechny sledované ekvivalizované příjmy (podle různých definic) také vázány (vždy pro každý rok zvlášť i v jednotlivých letech). Závislost v čase pro jednotlivé komponenty je popsána pomocí modelu, závislost v jednotlivých letech lze popsat pomocí korelačních koeficientů. Pokud předpokládáme přibližně normální rozdělení logaritmů příjmů (logaritmicko-normální rozdělení je považováno za přijatelný model pro rozdělení příjmů (Bílková, 2012; Bartošová, 2006)), vhodným popisem pro lineární závislost jsou korelační koeficienty uvedeny v tabulce 1.

Tabulka 1
Korelační koeficienty mezi logaritmy ekvivalizovaných příjmů *CPMSJ, CPPC* a *CPSJ* pro jednotlivé roky

proměnné	2007	2008	2009	2010
ln*CPPC*, ln*CPMSJ*	0,865	0,854	0,865	0,863
ln*CPPC*, ln*CPSJ*	0,952	0,948	0,953	0,953
ln*CPMSJ*, ln*CPSJ*	0,976	0,974	0,975	0,975

Zdroj: vlastní výpočty

Z tabulky 1 je vidět, že všechny hodnoty koeficientů jsou velmi vysoké a pohybují se v intervalu 0,85–0,98, pro všechny roky je nejvyšší hodnota zaznamenána mezi příjmy *CPMSJ* a *CPSJ* (poslední řádek tabulky 1). Je třeba také vzít v úvahu, že logaritmus zkoumané veličiny (4) je roven rozdílu logaritmů obou proměnných. Logaritmus celkového příjmu je shodný pro všechny tři příjmy a hodnotami v intervalu 8,5–14 převyšuje logaritmy počtu ekvivalentních jednotek, které nabývají hodnot od 0 (pro jednočlenné domácnosti) do 2,4 (pro největší domácnosti s 11 členy).

Jiný pohled na závislost sledovaných ekvivalizovaných příjmů poskytuje postavení jednotlivých domácností v uspořádaném výběru jednotlivých ekvivalizovaných příjmů. Předpokládejme, že byly pro všechny tři sledované ekvivalizované příjmy (a všechny čtyři roky) určeny kvartily v jednotlivých letech a u každé domácnosti byla

(pro všechny typy příjmů a všechny roky) nalezena část výběru, ve které se domácnost nalézá (čtvrtina nejmenších příjmů (I, od minima do dolního kvartilu), čtvrtina nižších středních (II, od dolního kvartilu do mediánu), čtvrtina vyšších středních (III, od mediánu do horního kvartilu), čtvrtina vysokých příjmů (IV, od horního kvartilu do maximální hodnoty). V tabulce 2 jsou uvedena procenta domácností, které jsou ve stejných čtvrtinách podle všech příjmů, a dále procenta těch, které jsou ve skupině nejmenších (skupina I), ve skupině největších (skupina IV) a ve středu (skupiny II a III). Uvedená procenta jsou stabilní v jednotlivých letech, necelých 60 % domácností je ve stejné čtvrtině příjmů pro všechny tři definice, podobné procento platí pro čtvrtinu nejmenších příjmů. Více domácností (necelých 70 %) je v prostřední polovině a nejvyšší shoda je pro vysoké příjmy ze čtvrtiny IV (téměř 80 %).

Tabulka 2
Procenta domácností ve shodných čtvrtinách výběru v jednotlivých letech (I–IV) a dále v dělení I, II + III, IV

rok	stejná čtvrtina všechny příjmy	čtvrtina I všechny příjmy	čtvrtina IV všechny příjmy	čtvrtiny II a III všechny příjmy
2007	57,1 %	57,5 %	81,3 %	66,7 %
2008	55,2 %	54,7 %	77,6 %	66,0 %
2009	56,4 %	56,3 %	78,4 %	67,0 %
2010	56,8 %	57,0 %	79,4 %	68,0 %

Zdroj: vlastní výpočty

Pokud jde o závislost v čase, ve stejné skupině (definované kvartily) jako v roce 2007 do roku 2010 zůstalo 60 % domácností (a vždy ve dvou následujících letech přes 88 %) pro všechny příjmy.

V tabulce 3 jsou uvedeny popisné charakteristiky sledovaného ekvivalizovaného příjmu. Za základ bereme výběrové hodnotu pro ekvivalizovaný příjem *CPMSJ*, pro který jsou uvedeny hodnoty výběrových charakteristik polohy (aritmetický průměr a medián) a variability (směrodatná odchylka *S*, kvartilová odchylka *Q* (polovina kvartilového rozpětí) a variační koeficient *V*). Pro další dva příjmy jsou v tabulce 3 uvedena procenta hodnot pro příjmy *CPSJ* a *CPPC* (s výjimkou posledního řádku tabulky, kde je uvedena procentní změna hodnoty ve sloupci od roku 2007 do 2010), kde základem je odpovídající hodnota pro *CPMSJ* (levá část tabulky). Příjmy jsou uvažovány bez zohlednění inflace, ve sledovaných letech byla celková inflace rovna 7,8 % (ČSÚ, 2014). Ve sledovaných letech rostly v domácnostech ve výběru charakteristiky polohy i variability příjmů, celkový nárůst byl větší než inflace. Je vidět, že medián příjmu *CPSJ* je kolem 86 % mediánu ekvivalizovaného příjmu podle modifikované metodiky OECD ve všech letech a medián příjmu na jednu osobu je přibližně 75 % mediánu tohoto příjmu.

Tabulka 3
Popisné charakteristiky příjmu *CPMSJ* (Kč), variační koeficient *V* (levá část), procenta mediánu a kvartilové odchylky pro *CPSJ a CPPC* (100% je odpovídající hodnota pro *CPMSJ*).

rok	*CPMSJ* (Kč)					*CPSJ*		*CPPC*	
	průměr	medián	*S*	*Q*	*V*	medián	*Q*	medián	*Q*
2007	175 315	156 414	81 149	40 020	46,3 %	86,5 %	76,7 %	74,2 %	60,5 %
2008	189 403	168 760	93 699	42 009	49,5 %	86,9 %	78,7 %	74,4 %	62,6 %
2009	195 418	173 408	102 597	41 908	52,5 %	87,5 %	79,1 %	75,4 %	67,1 %
2010	197 113	176 839	92 724	43 316	47,0 %	87,8 %	79,6 %	75,1 %	67,8 %
Změna	+12 %	+13 %	+14 %	+8 %	-	+15 %	+12 %	+14 %	+21 %

Zdroj: vlastní výpočty

Z tabulky 3 je také patrné, že rozdělení příjmu *CPMSJ* je kladně zešikmené (průměrné hodnoty jsou větší než mediány, koeficienty šikmosti (neuvedeno v tabulce) jsou kladné a nabývají hodnot od 2,9 do 6,5 pro všechny roky a všechny typy ekvivalizovaných příjmů). V případě logaritmování hodnot (tak jako v tabulce 1) se koeficienty šikmosti zmenší na hodnoty od −0,1 do 0,4, stále ještě ale není vhodné předpokládat, že logaritmy pocházejí z normálního rozdělení (to by odpovídalo jednosložkovému modelu směsi, volba

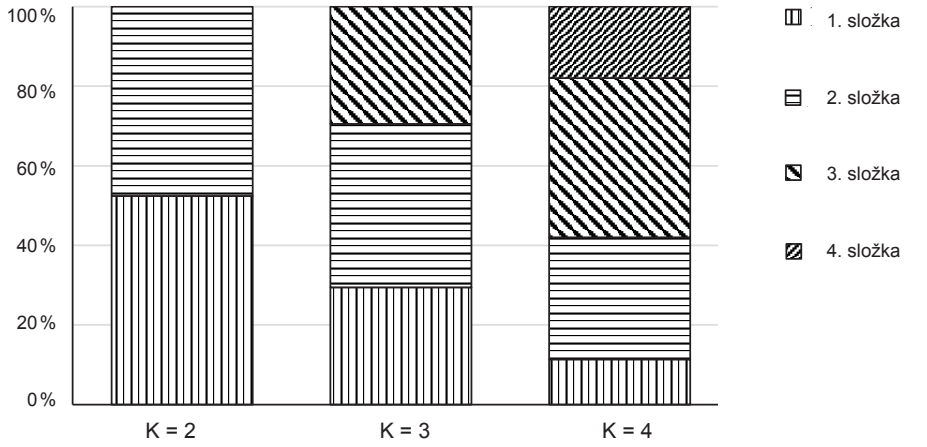
Pro aplikaci modelu směsi s neznámou příslušností ke komponentám je třeba zvolit (nebo odhadnout) vhodný počet komponent K. K tomuto účelu je možné použít grafické metody, bootstrapový test nebo, tak jako v předkládané analýze, informační kritéria. Analyzovanými daty byly proloženy modely směsi (1) pro $K = 1$ (pouze pro srovnání) až pro $K = 9$ a byl sledován průběh hodnot Akaikeho informačního kritéria (minimální hodnoty bylo dosaženo pro $K = 7$) a Bayesovského informačního kritéria (minimum pro $K = 8$). V obou kritériích docházelo k rychlému poklesu hodnoty do čtyř komponent, pak se pokles zpomalil. Předkládaný model je možné považovat pouze za užitečný popis rozdělení (též Bartošová; Longford, 2014), proto byly jako kompromis mezi kvalitou modelu a numerickými možnostmi zvoleny jako optimální modely se třemi a čtyřmi komponentami (odpovídá doporučením z Bartošová; Longford, 2014). Tyto modely umožňovaly nalézt dostatečně přesné odhady parametrů (44 pro $K = 3$ a 59 pro $K = 4$ a přehledně interpretovat výsledky. Hodnoty Akaikeho kritéria jsou rovny 7 221 pro $K = 2$, 5 942 pro $K = 3$ a 4 315 pro $K = 4$. Nalezené složky jsou umělé shluky popsané pravděpodobnostním rozdělením a není zřejmá jejich interpretace. Pro tento text bylo zvoleno posouzení odhadnutých středních hodnot komponent, velikost variability jednotlivých složek a dále intenzita závislosti mezi sledovanými roky posuzovaná pomocí korelační matice.

Na obrázcích 1–3 jsou nalezené složky seřazeny podle odhadnutého středního příjmu *CPMSJ* v roce 2007 od nejmenší do největší hodnoty (střední hodnoty závisejí na parametrech $\boldsymbol{\mu}_j$ i na diagonálních prvcích kovariančních matic $\boldsymbol{\Sigma}_j$ viz Bílková, 2012; Malá, 2013). Rozdělení pravděpodobností mezi jednotlivé složky je znázorněno na obrázku 1, odhadnuté střední hodnoty složek a jejich vývoj v čase obsahuje

obrázek 2. Hodnoty odhadů hodnot vektorů **μ** (střední hodnoty složek logaritmů celkových příjmů) je ukázán na obrázku 3.

Odhadnuté váhy směsí byly modely odhadnuty (až do tří komponent) jako přibližně stejně velké, až při čtyřech složkách byly skupiny středně příjmové více zastoupeny než skupiny vysoko a nízkopříjmové (obrázek 1).

Obrázek 1
Rozložení odhadů pravděpodobností složek

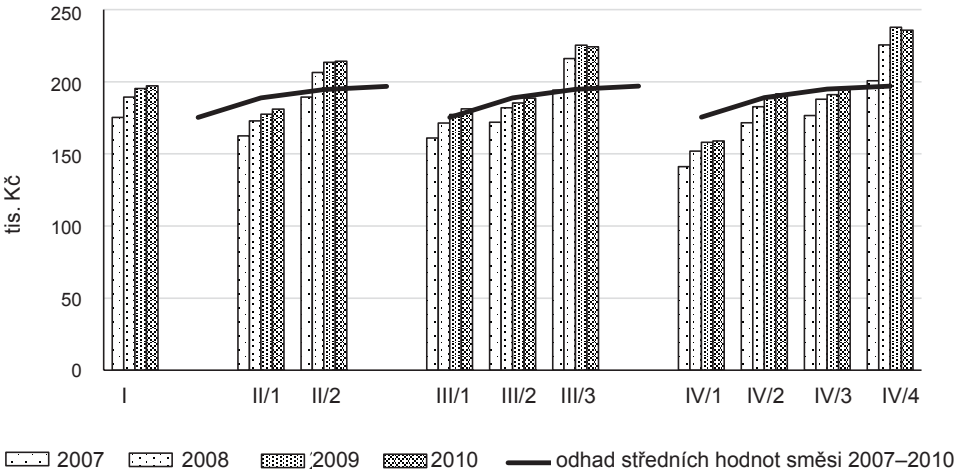


Zdroj: vlastní výpočty

Odhadnuté střední hodnoty komponent ve směsi jsou pro počet komponent 2–4 znázorněny na obrázku 2. Sloupcový graf obsahuje odhady středních hodnot složek a jejich vývoj v čase. Z grafu je patrné, že model nacházel pouze jednu nadprůměrnou komponentu, ostatní byly podprůměrné. Je to proto, že na rozdíl od předpokladu, model nevyděloval jednu malou skupinu s vysokými příjmy, ale poměrně rozsáhlou skupinu domácností s vyššími příjmy (jak také ukazuje obrázek 1). V případě zvýšení počtu komponent model dělil domácnosti středně příjmové, případně byla posunuta skupina nízkopříjmová dolů (viz také obrázek 3).

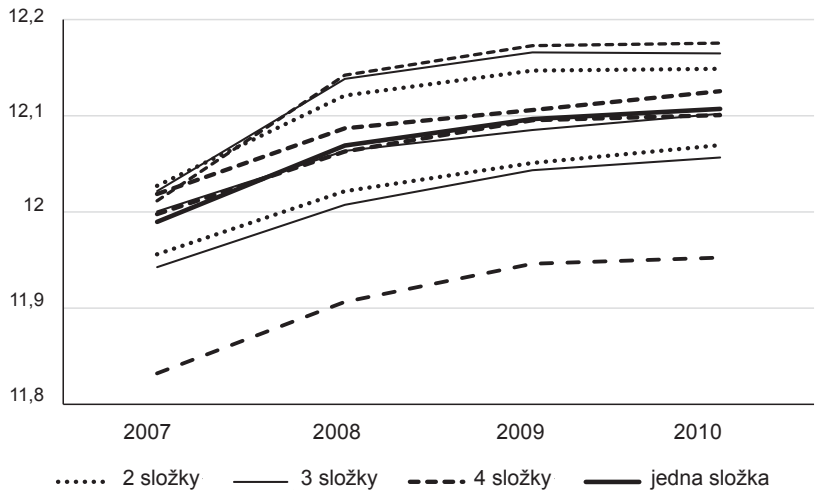
Pro dvě složky byly odhady parametrů rozptylu logaritmů v první složce menší (kolem 0,08), zatímco pro druhou složku byly kolem 0,24. Pro tři složky měly menší dvě složky přibližně stejnou hodnotu odhadů rozptylů kolem 0,08–0,1, zatímco pro třetí složku jsou odhadnuté rozptyly větší než 0,3. Pro čtyři složky měla nejmenší odhady (ve všech letech) parametru rozptylu logaritmů (kolem 0,06), další dvě složky kolem 0,1 a největší rozptyl má složka vysokých příjmů s rozptylem logaritmů větším než 0,3. V první nejmenší složce jsou také největší korelace příjmů v jednotlivých letech. V korelačních maticích (není uvedeno v tomto textu) je dobře vidět síla závislosti mezi příjmem domácností v jednotlivých letech. V první a třetí komponentě jsou korelace mezi jednotlivými roky vyšší, dochází k pomalému vývoji (podle dalšího růstu) příjmů. V dalších složkách jsou větší odhadnuté rozptyly příjmů a také závislost mezi roky je slabší.

Obrázek 2
Odhady středních hodnot složek rozdělení směsi, *K* = 2−4



Zdroj: vlastní výpočty

Obrázek 3
Odhady středních hodnot složek logaritmů



Zdroj: vlastní výpočty

Ukažme nyní na obrázku 3 průběh odhadů čtyřrozměrných středních hodnot logaritmů složek $\boldsymbol{\mu}_j$, j = 1, ..., K, K = 2 – 4. Pro srovnání je v obrázcích uvedena také střední hodnota maximálně věrohodný odhad vektoru střední hodnoty logaritmů příjmů, jestliže bychom předpokládali, že tvoří výběr ze čtyřrozměrného normálního rozdělení. Hodnoty středních hodnot pro malý počet komponent (do 4) stále rostou, je vidět složka s největšími hodnotami téměř společná pro všechny modely. Naopak čím větší počet složek, tím nižší jsou hodnoty vektoru pro domácnosti s nejnižšími příjmy. Průměr logaritmů příjmů (plná čára na obrázku 2 znázorňující průměrný vývoj) zhruba vystihuje prostřední složka pro model se třemi složkami a druhá složka v prezentovaném modelu (tabulka 3) se čtyřmi složkami.

V tabulce 4 jsou uvedeny odhadnuté korelace mezi logaritmy příjmů *CPEJ* v jednotlivých letech určené z odhadů korelačních matic pro model se čtyřmi komponentami. Korelace v první skupině jsou vyšší než v dalších složkách a znamenají vyšší stabilitu v příjmech v této nízkopříjmové složce.

Tabulka 4
Kovariance mezi jednotlivými roky ve složkách,

	2007	**2008**	**2009**	**2007**	**2008**	**2009**	**2007**	**2008**	**2009**
2008	0,963	1		0,651	1		0,567	1	
2009	0,950	0,973	1	0,385	0,632	1	0,285	0,501	1
2010	0,409	0,407	0,802	0,379	0,626	0,961	0,131	0,295	0,583

Model směsi s nepozorovanou příslušností k jednotlivým umělým komponentám je vlastně úlohou shlukové analýzy, kdy předpokládáme zvolený počet shluků K (používáme shluk místo složky) a snažíme se je v datech identifikovat. Příslušnost pozorování k jednotlivým shlukům (složkám) lze na základě modelu posoudit pomocí odhadu posteriorních pravděpodobností (3), že dané pozorování přísluší k danému shluku. Výsledky zařazení do komponent pomocí posteriorních pravděpodobností jsou znázorněny v tabulce 5. Pro každou domácnost ve výběru byla nalezena komponenta, která měla největší odhadnutou posteriorní pravděpodobnost. Hodnotu této pravděpodobnosti vezmeme jako charakteristiku toho, s jakou jistotou model pozorování zařazuje. V prvním sloupci tabulky 5 jsou uvedena procenta pro případ, ve kterém jsou všechny komponenty stejně pravděpodobné a pozorování jsou do nich umisťována náhodně se stejnými pravděpodobnostmi. Dále tabulka obsahuje procenta domácností ve výběru s maximálními posteriorními pravděpodobnostmi většími než 0,5, 0,6, 0,7, 0,8 a 0,9. Jak lze očekávat, s rostoucím počtem složek (ve sloupcích) procento domácností klesá, nicméně procento domácností, pro které existuje dominantní komponenta (komponenta s odhadnutou posteriorní pravděpodobností větší než 0,5) nekleslo v žádném modelu pod 95 %. Největší pravděpodobnost zařazení má nad 0,90 pro všechny modely aspoň 40 %, pro model se dvěma komponentami je to až 7 z 10 domácností.

Tabulka 5
Hodnoty odhadnutých posteriorních pravděpodobností pro zařazení domácností do komponent

	100/*K*	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
***K* = 2**	50,0 %	100,0 %	94,7 %	88,9 %	81,2 %	68,2 %
***K* = 3**	33,3 %	98,5 %	87,7 %	74,3 %	60,0 %	44,2 %
***K* = 4**	25,0 %	95,4 %	83,6 %	72,1 %	59,2 %	41,8 %

Zdroj: vlastní výpočty

Všechny modely tedy dovolují poměrně s jistotou klasifikovat pozorování do nalezených umělých složek. Rozdíl mezi modelem se třemi a čtyřmi komponentami není velký (srovnání předposledního a posledního řádku v tabulce 5).

Závěr

České domácnosti tvoří vzhledem k příjmům velmi nehomogenní množinu, předkládaný model se snaží ve výběru domácností nalézt umělé (nepozorovatelné) podmnožiny, které jsou homogennější, a pro tyto podmnožiny nalézt vhodný popis rozdělení příjmů. Součástí modelu je také odhad velikosti těchto podmnožin, respektive jejich zastoupení v množině všech domácností.

Předkládaný vícerozměrný pravděpodobnostní model odhaduje sdružené (čtyřrozměrné) rozdělení ekvivalizovaných příjmů domácností ve čtyřech následujících letech, a umožňuje tak úsudky založené na jakýchkoliv marginálních rozděleních, speciálně na jednorozměrných příjmech v jednotlivých letech. Model také umožňuje odhadnout závislosti mezi příjmy ve sledovaných letech a také na základě posteriorních pravděpodobností složek jednotlivá pozorování zařadit do složek.

Byl odhadnut model s dvěma až osmi složkami, jako kompromis mezi kvalitou modelu (posuzovanou pomocí informačních kritérií) a možností odhadnout velké množství parametrů byly vybrány pro popis rozdělení logaritmů příjmů modely se třemi nebo čtyřmi komponentami s vícerozměrným normálním rozdělením (v souladu s Bartošová; Longford, 2014).

Pro odhad ve vícerozměrném modelu je potřeba analyzovat opakovaná individuální pozorování, což snižuje možnost aplikace modelu. Dalším problémem (statistickým) je velké množství parametrů, které je třeba odhadnout, a to, že v řešeném problému nelze rozumně využít v literatuře navrhované postupy, jak snížit počet parametrů, případně zlepšit identifikaci modelu. V práci McLachlan; Peel, 2000 je navrhováno použít stejné kovarianční matice pro všechny komponenty, nebo použít matice diagonální. V případě malého počtu komponent jsou modely tohoto typu výrazně horší než model předkládaný (posuzováno například Akaikeovým informačním kritériem) a odhadnuté kovarianční matice jsou různé (rozptyly i kovariance jsou různé v různých komponentách) a korelační koeficienty nejsou malé (existuje závislost) mezi jednotlivými roky. Hledání odhadů bylo časově velmi náročné a bylo spojeno s poměrně velkými numerickými

problémy. Největším problémem bylo nalézt vhodné počáteční aproximace pro odhadované parametry, zvláště kovarianční matice a pravděpodobnosti jednotlivých komponent.

Model konečných směsí má přímou spojitost se shlukovou analýzou, nahradíme-li pojem komponenta slovem shluk. Proto je zajímavé, jak dobře je model směsi schopen zařadit domácnosti do jednotlivých komponent (tabulka 5). Podle tabulky modely se třemi i čtyřmi složkami umožňují poměrně s jistotou zařazovat jednotlivé domácnosti do složek.

Předkládané modely je možné považovat za užitečný popis rozdělení příjmů obecně (v této práci v letech 2007–2010), i když testy dobré shody zamítají hypotézu o vhodnosti rozdělení. Na základě předkládaného modelu lze také, s jistou opatrností, sestrojit krátkodobé předpovědi na následující období.

Literatura

BARTOŠOVÁ, J. 2006. Logarithmic-normal model of income distribution in the Czech Republic. *Austrian Journal of Statistics*. 2006, Vol. 35, No. 2&3, pp. 215–222.

BARTOŠOVÁ, J.; LONGFORD, N. T. 2014. A study of income stability in the Czech Republic by Finite Mixtures. *Prague Economic Papers.* 2014, Vol. 23, No. 3, pp. 330–248.

BENAGLIA, T.; CHAUVEAU, D.; HUNTER, D. R.; YOUNG, D. 2009. An R package for analyzing finite mixture models. *Journal of Statistical Software*. 2009, Vol. 32, No. 6, pp. 1–29. Dostupné na: http://www.jstatsoft.org/v32/i06/.

BÍLKOVÁ, D. 2012. Recent development of the wage and income distribution in the Czech republic. *Prague Economic Papers*. 2012, Vol. 21, No. 2, pp. 233–250. doi: 10.18267/j.pep.421.

BOLDEA, O.; MAGNUS, J. R. 2009. Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association. 2009,* Vol. 104, No. 488, pp. 1539–1549. doi: 10.1198/jasa.2009.tm08273.

ČSÚ. 2014. *Životní podmínky 2008, 2009, 2010, 2011.* Český statistický úřad. http://www.czso.cz. [červen 2014].

FLACHAIRE, E.; NUNEZ, O. 2007. Estimation of the income distribution and detection of subpopulations: an explanatory model. *Computational Statistics & Data Analysis*. 2007, Vol. 51, No. 7, pp. 3368–3380.

MALÁ, I. 2013. Použití konečných směsí logaritmicko-normálních rozdělení pro modelování příjmů českých domácností. *Politická ekonomie*. 2013, Vol. 61, No. 3, pp. 356–372.

MAREK, L. 2010. Analýza vývoje mezd v ČR v letech 1995–2008. *Politická ekonomie*. 2010, Vol. 58, No. 2, pp. 186–206.

McLACHLAN, G. J.; PEEL, D. 2000. *Finite mixture models*. New York: Wiley series in Probability and Mathematical Statistics, 2000. ISBN 978-0-471-00626-8. doi: 10.1002/0471721182.

MIXTOOLS. 2014. *Package 'mixtools'.* http://cran.r-project.org/web/packages/mixtools/mixtools.pdf. [červen 2014].

MPSV. 2014. *Analýza vývoje příjmů a výdajů domácností ČR*. Ministerstvo práce a sociálních věcí. ČR. http://www.mpsv.cz/cs/. [září 2014].

PITTAU, M. G.; ZELLI, R. 2006. Empirical evidence of income dynamics across EU regions. *Journal of Applied Econometrics*. 2006, Vol. 21, No. 5, pp. 605–628. doi: 10.1002/jae.855.

R Core Team. 2012. *A language and environment for statistical computing*. Vienna, Austria : the *R* Foundation for Statistical Computing. http://www.R-project.org/. [červen 2014].

SOCR. 2014. *Domácnosti v ČR: příjmy, spotřeba, úspory a dluhy v letech 1993 až 2012*. Svaz obchodu a cestovního ruchu ČR. http://www.socr.cz/file/2115/analyza-csu-domacnosti-v-cr---prijmy--spotreba--uspory-a-dluhy-1993-2012.pdf [červen 2014].

VEČERNÍK, J. 2013. The changing role of education in the distribution of earnings and household income. *Economics of Transition.* 2013, Vol. 21, No. 1, pp. 111–133.

MULTIVARIATE PROBABILITY MODEL FOR INCOMES OF THE CZECH HOUSEHOLDS

Ivana Malá, University of Economics, Prague, W. Churchill Sq. 4, CZ – 130 67 Prague 3
(malai@vse.cz)

Abstract
The equivalised total net annual incomes of the Czech households (in CZK) in 2007–2010 are analysed in the text. The set of all households is very nonhomogeneous (with respect to incomes) and the aim of the analysis is to determine more homogeneous subsets (components) and to describe the distribution of incomes in these components. The components are supposed to be artificial, the membership of households in components is not known (or observable). A multivariate mixture of normal distributions (four dimensional component distributions) is used to describe a vector of logarithms of incomes, models with 2 to 9 components are fitted. Maximum likelihood estimates of unknown parameters were found with the use of EM algorithm. Akaike information criterion was used (accompanied by bootstraped test) and models with 3 or 4 components were selected to be acceptable for the description of distribution of incomes. Cluster analysis was performed in order to classify households into components and good performance of the model was found.

Keywords
distribution of incomes, multivariate normal distribution, finite mixture of distributions, EM algorithm, cluster analysis, maximum likelihood estimate

JEL Classification
C46, D31, C38