REGRESNÍ MODEL PRO PREDIKCI BUDOUCÍCH POČTŮ POJIŠTĚNCŮ VZP ČR

Jaromír Běláček

Anotace

V příspěvku[1] [1] byl představen tzv. mechanizmový model pro predikci budoucích počtů pojištěnců VZP ČR, založený na aditivních dopočtech extrapolovaných kvartálních přeregistračních sald (mezi zdravotními pojišťovnami) a multiplikativních korekturách vzhledem k aktualizovaným variantám demografické projekce ČSÚ (z r. 2013).

V tomto příspěvku bude popsán regresní model, který odhaduje regresní parametry metodou nejmenších čtverců v zobecněném lineárním modelu, kde váhy retrospektivních pozorování jsou přímo úměrné exponenciálně klesajícím diskontním faktorům (koeficientům α) jako u exponenciálního vyhlazování (běžné pro jednodušší polynomiální modely).

Nově představený model rozšiřuje původní mechanizmový o čtyři optimalizované nezávisle proměnné (přeregistrační a demo proměnnou, a dále o formální absolutní a lineární člen, poslední dva kumulativně překvalifikované na lineární a kvadratický). Kalibrace modelu (volba diskontu α) byla provedena zvlášť pro každou z 18ti pětiletých věkových skupin na základě maximalizace nelineárního průběhu procenta rozptylu vysvětlovaného modelem.

Vizuální porovnávání výsledků predikce pro tři varianty možného vývoje přeregistrací pojištěnců je operacionalizováno v rámci uživatelské aplikace v prostředí MS Excel. Diskontovaný regresní model umožňuje jednodušší rozpočet extrapolovaných variant predikce pro systém všech 14ti krajů ČR do r. 2032. Vzhledem k rychle se měnící dynamice počtu pojištěnců ve stávajícím období je ale zapotřebí nastavení parametrů regresního modelu rok od roku operativně přepočítávat.

Klíčová slova

odvozené demografické projekce, zobecněné lineární modely, exponenciální vyrovnávání, koeficient determinace, odhad metodou nejmenších čtverců

Vývoj řešené problematiky

Úlohu predikce budoucích počtů pojištěnců můžeme považovat za klíčovou pro kalkulaci budoucí profitability pojišťovny. Ve všeobecné zdravotní pojišťovně se tato úloha tradičně řeší za účelem tvorby ročních *zdravotně pojistných plánů* metodikou expertního stanovení budoucích meziročních změnových indexů ve srovnání s minulým obdobím. Tuto heuristickou metodu lze úspěšně

1 Běláček J.: Predikce budoucích počtů pojištěnců VZP ČR – data, metodika a výsledky. In.: Sborník příspěvků MEDSOFT 2018. Hotel Academic, Roztoky u Prahy, 20.–21.3.2018, vyd.: Creative Connections ve spolupráci s Zeithamlová Milena – Agentura Action M, Praha, ISSN 1803-8115, 2018, str.7–19

použít i pro predikci počtů pojištěnců v pětiletých věkových skupinách, nikoli však pro delší časové období. Důvodem jsou zářezy ve věkové struktuře obyvatelstva ČR, které se vizualizují i v ročních časových řadách počtů pojištěnců VZP ČR, když jejich průběhy zobrazíme v rámci jednotlivých věkových skupin. Pokud máme navíc k dispozici retrospektivní údaje o vývoji průměrných příjmů a průměrných výdajů v příslušných věkových skupinách, může být kalkulace celkových příjmů a výdajů zdravotní pojišťovny schůdnou záležitostí i ve střednědobém výhledu (řekněme pro období budoucích 3–5 let).

Predikce počtů pojištěnců pro střednědobý horizont musíme ale založit na sofistikovanějších algoritmech než na indexování. Jeden z nich vycházející z metodiky tzv. *odvozených demografických projekcí* byl použit např. pro odhady budoucích demografických struktur pacientů ZZ AGEL v kraji Olomouckém, Moravskoslezském a v Praze (zde ale pro málo reprezentativní vzorek pacientů). Výsledky byly poprvé prezentovány na IX. Sympoziu AGEL v Olomouci v říjnu 2015 (viz [1]) a publikačně shrnuty v následujících dvou letech (viz [2–3]). Pro demograficky reprezentativnější počty budoucích pojištěnců VZP ČR byl tento model aplikován a posléze rozšířen ještě o časové řady počtů pojištěnců získaných z databáze přeregistrací (tedy pojištěnců migrujících mezi zdravotními pojišťovnami). Pro verzi ročních časových řad byl představen v příspěvku [4] jako tzv. *mechanizmový model* predikce počtu pojištěnců VZP ČR (pro kvartální verzi je tento model představen ve stati 2.1).

Paralelně s mechanizmovým modelem byly na stejné datové bázi prováděny experimenty se *zobecněnými lineárními modely* (obecně viz [6], str. 233), když váhy retrospektivních pozorování byly voleny přímo úměrné exponenciálně klesajícím diskontním faktorům α (0<α≤1) – vzorce ve stati 2.2. Pro formálně jednodušší polynomiální regresní modely tento přístup koresponduje s metodikou tzv. *exponenciálního vyhlazování* (viz [6], str. 57–72), kde pro kalibraci α na bázi tzv. *in-sample přepočtů* lze využít explicitně odvozených sofistikovaných vzorců. (Zcela obecně bylo in-sample ověřování – tzv. fit – regresních modelů pro obdobnou problematiku použito pro modelování budoucích příjmů ze zdravotního pojištění, nezávisle na věku pojištěnců – viz [5], str. 2). V rámci uživatelské aplikace vytvořené ve VZP ČR pod MS Excel však nebylo jednoduché toto provést při kalibraci α právě v případě zobecněných lineárních modelů; namísto toho byla tedy α ve všech 18ti věkových skupinách kalibrována na základě maximalizace nelineárního průběhu procenta rozptylu vysvětleného regresním modelem (podrobněji ve stati 3. a 4.).

2 Metodika

2.1 Mechanizmový model

Mechanizmový model pro predikci počtu pojištěnců VZP ČR (na anuální bázi pro každé čtvrtletí) byl představen v příspěvku [4]. S ohledem na pozdější transkripci dat do unifikované kvartální verze můžeme budoucí počty pojištěnců

P_x^t pro každou věkovou skupinu "x"(= '0-4', … ,'85+') generovat na základě předpisu

$$\hat{P}_x^t = (\hat{P}_x^{t-1} + S_x^t) \cdot (D_x^t/D_x^{t-1}),\ t = 1, 2, \ldots. \qquad (1)$$

kde $\hat{P}_x^0 \equiv P_x^0$ značí známé vstupní počty pojištěnců ve věku „*x*" v prahovém čase t=0 (ke konci vstupního čtvrtletí), D_x^t je střední stav žijících osob (nebo součet zdravotních pojištěnců v ČR ve věkové skupině „*x*") přepočtený k témuž kvartálu jako $\hat{P}_x^t$ a S_x^t značí saldo přeregistrací předpokládané (nebo extrapolované) pro období mezi roky *t-1* a *t* .

Při speciální volbě $S_x^t = 0$ pro všechny budoucí časy *t* se model (1) redukuje na běžný *model odvozené demografické projekce*. (Při stávající dynamice pojistného kmene VZP ČR dokonce tento předpoklad může být pro nejbližší období i jedním z reálných scénářů pro nejbližší budoucí vývoj.) Ale můžeme uvažovat i alternativu pouze *čisté projekce kumulativních přeregistrací*, kterou získáme formální volbou $D_x^t/D_x^{t-1} = 1$ pro všechny budoucí časy t. Přestože tento druhý scénář vlastně pro žádnou věkovou skupinu „*x*" není reálný, může být dobře využit při účely posuzování možného vzájemného i synergického vlivu obou „vysvětlujících komponent" tj. *demografické a přeregistrační* na budoucí vývoj počtů pojištěnců.

Z hlediska formálního vyjádření (1) je zahrnutí obou vysvětlujících složek mechanizmového modelu zjevně nesymetrické. Multiplikativní demografický index (D_x^t/D_x^{t-1}) v rámci tohoto výpočetního předpisu umožňuje operativní přenos dynamiky demografie ČR homogenně na celý pojistný kmen VZP ČR (ve věkových skupinách); ten se ale bez zahrnutí přeregistrační složky již dříve ukázal jako ne zcela postačující. Aditivní přeregistrační složka S_x^t se v rámci vzorce (1) adaptuje na složku demografickou odpovídajícím způsobem – ale přeregistrovaní pojištěnci by se samozřejmě nemuseli řídit stejnými homogenizačními demografickými předpoklady jako celý pojistný kmen. (Na tomto místě je třeba si uvědomit, že zatímco demografická složka reflektuje celorepublikové procesy úmrtnosti, migrace obyvatelstva a v nejnižší věkové skupině také porodnosti; přeregistrační složka se týká pouze meziročního přerozdělování pojištěnců mezi zdravotními pojišťovnami. Formálně jde tedy o nezávislé komponenty.) Symetrie obou vysvětlujících proměnných v modelu bychom dosáhli například formálním přepisem (1) do tvaru

$$\hat{P}_x^t = (\hat{P}_x^{t-1} + S_x^t) + (\hat{P}_x^{t-1} + S_x^t) \cdot (D_x^t/D_x^{t-1} - 1),\ t = 1, 2, \ldots. \qquad (1a)$$

kde celý druhý sčítanec můžeme považovat za kvartální *saldový* přírůstek počtu pojištěnců, který spadá na konto celorepublikového *demografického vývoje*. Multiplikativní člen ve vzorci (1a) má interpretaci *relativního přírůstku* spadajícího na konto demografie ČR mezi sousedními kvartály.

Takto koncipovaný model má přínos pro porozumění smysluplnosti zobecněného regresního modelu popsaného ve stati 2.2.

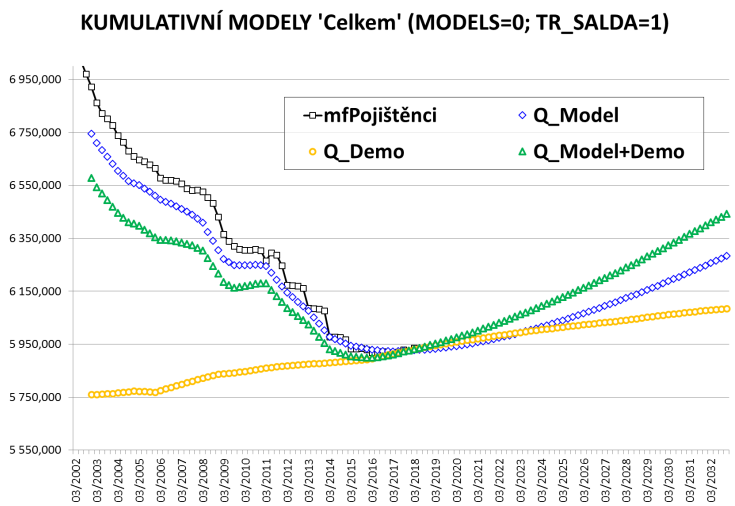
Máme-li pro věkové skupiny k dispozici i retrospektivní kvartální údaje o počtech pojištěnců, tj. P_x^t pro *t = -1, … , -T,* a rovněž údaje o demografii a o minulých přeregistracích, lze výše uvedený rekurentní vzorec formalizovat rovněž

retrospektivně. Vyjdeme-li z téhož prahového času *t=0*, můžeme generovat analogickou posloupnost modelových hodnot směrem do minulosti předpisem

$$\hat{P}_x^t = (\hat{P}_x^{t+1} + S_x^t) \,.\, (D_x^t/D_x^{t-1}) \,,\, t = -1, -2, \ldots, -T \,. \quad (2)$$

Ukázka finální aplikace modelů (1) a (2) vyscítaných z věkových skupin pro úroveň celkových počtů pojištěnců VZP ČR je pro střední variantu extrapolace hodnot S_x^t do budoucnosti prezentována na Grafu č. 1.

Graf 1 – Vývoj kvartálních počtů pojištěnců VZP ČR za období 03/2002 – 12/2017 a časové řady jejich teoretických počtů generovaných rekurentními mechanizmovými modely (pro střední variantu extrapolace přeregistračního salda do roku 2032)



Graf 1 – Skutečné kvartální počty pojištěnců VZP ČR (v tisících) jsou zobrazeny černými čtverci jako časová řada 'mfPojištěnci', hodnoty generované kvartálním mechanizmovým modelem (1) a (2) jsou vyobrazeny zelenými trojúhelníky (řada 'Q_Model+Demo'). Modelovou řadu lze dále rozložit na součet čistě demografické složky (řada 'Q_Demo' – žlutá kolečka) vytvořené jednoduchým indexováním (metodika odvozené demografické projekce) a složky čistě přeregistrační (řada 'Q_Model' – modré kosočtverečky) při hypotetickém nulovém kvartálním saldu demografie. Kvartální demografické komponenty modelu byly získány interpolací ze středních stavů obyvatelstva ČR v pětiletých věkových skupinách (viz portál ČSÚ) a ze střední varianty demografické prognózy ČSÚ z r. 2013. Historická kvartální salda přeregistrací pro pojištěnce VZP ČR pro období 03/2002 – 12/2017 byla převzata z interních databází VZP a z registru pojištěnců ČR; pro období 03/2018 – 12/2032 jsou na grafu budoucí přeregistrační salda reprezentována střední variantou extrapolace z ročních časových řad 2012–17. (Tato varianta předpokládá trvale mírný logaritmický nárůst salda přeregistrací pro celou dobu predikovaného období – podrobněji v [4] v odst. 5.3). Na grafu 1 zobrazené modelové řady vznikly jako součet metodicky stejně odvozených křivek pro 18 pětiletých věkových skupin dohromady.

2.2 Zobecněný lineární model

Při značení jako v části 2.1 lze základní zobecněný lineární regresní model vyjádřit ve formě

$$\Delta(P_x^t;\theta) = b_x + c_x \,.\, t + d_x \,.\, S_x^t + e_x \,.\, \Delta(D_x^t) + \varepsilon(0, wt \,.\, \sigma_x^2), t = 1, \ldots., t_0, \ldots, T \quad (3)$$

kde $\Delta(P_x^t;\theta) \equiv (P_x^t - P_x^{t-1})$ značí retrospektivní diference počtu pojištěnců mezi konci sousedících kvartálů, $\theta \equiv (b_x, c_x, d_x, e_x)$ vektor lineárních regresních parametrů, t_0 práh projekce a $\varepsilon(0, w_t . \sigma_x^2)$ označuje chybu měření (s nulovou střední hodnotou a rozptylem ve tvaru součinu $w_t \,.\, \sigma_x^2$ pro vahový faktor w^t funkcionálně nezávislý na systémovém parametru σ_x^2). Výraz $\Delta(D_x^t)$ reprezentuje kvartální diference absolutního přírůstku pojištěnců příslušných demografické složce. Ve vzorci (3) jsme použili prospektivní (pro $t > t_0$) i retrospektivní (pro $t < t_0$) aproximace typu $\Delta(D_x^t) \approx \hat{P}_x^{t-1} \,.\, (D_x^t/D_x^{t-1}-1)$ tzn. jako v mechanizmovém modelu, jen oproti druhému sčítanci v (1a) zanedbáváme rozpočet demografického faktoru na přeregistrační člen S_x^t ; tím se do jisté míry zbavujeme nadbytečné kolinearity mezi S_x^t a $\Delta(D_x^t)$. Za důležitou považujeme vlastnost, že všechny členy na pravé straně v (3) jsou nyní součástí standardizovaného „aditivního" regresního modelu, který zahrnuje dřívějších mechanizmové modely jako svůj speciální případ (při fixní volbě $b_x = c_x = 0$ a $d_x = e_x = 1$).

Odhady $\hat{\theta}$ vektoru parametrů θ v modelu (3) získáváme minimalizací váženého součtu čtverců

$$\Sigma_t = 0, \ldots, -T \, w_t \,.\, [\Delta(Px^t;\theta) - (b_x + c_x \,.t + d_x \,.\, S_x^t + ex \,.\, \Delta(D_x^t))]^2 \quad (4)$$

jako tzv. *Aitkenův odhad* (viz v [6], str. 233). Váhy ve vzorci (4) jsou v daném případě definovány jako $w_t \equiv \alpha_x^{-t}$ pro diskontní faktory α_x ($0 < \alpha_x \leq 1$) – to znamená, že s postupným vzdalováním od prahové hodnoty (v čase t=0) dostávají historická data exponenciálně se snižující vliv. (Pouze při $\alpha_x = 1$ odhady $\hat{\theta}$ vektoru parametrů θ korespondují s klasickou *neváženou* lineární regresí, kdy všechny minulé hodnoty časové řady mají stejnou váhu.) Dosazením optimalizovaných odhadů do vzorce (4) získáme odhady pro rozptylový parametr σ_x^2, jehož prostřednictvím kvalifikujeme statistickou významnost parametrů v zobecněných lineárních modelech a odhadnout pásy spolehlivosti kolem predikovaných regresních přímek pro každou věkovou skupinu *"x"* (= '0-4',…,'85+') zvlášť.

Pro účely finální predikce počtů pojištěnců nakonec použijeme kumulativní přepočet

$$\begin{aligned} \hat{P}_x^t &= \hat{P}_x^{t-l} + \Delta(\hat{P}_x^t;\hat{\theta}) = P_x^{t0} + \Sigma_{l=t0,\ldots,t} \, \Delta(\hat{P}_x^1;\hat{\theta}) = \\ &= P_x^0 + \Sigma_l \, (\hat{b}_x + \hat{c}_x \,.\, l + \hat{d}_x \,.\, S_x^l + \hat{e}_x \,.\, \Delta(D_x^l)) = \\ &= P_x^{t0} + \hat{b}_x \,.t + \hat{c}_x \,.\, (t+1)t/2 + \hat{d}_x \,.\Sigma_t \, S_x^t + \hat{e}_x \,.\Sigma_t \, \Delta(D_x^t) \,, \\ &\quad t = t_0+1 \,, \ldots, T \end{aligned} \quad (5)$$

kde $\Delta(\hat{P}_x^t;\hat{\theta})$ jsou hodnoty střední hodnoty regresní funkce z (3) – tedy výrazu na pravé straně rovnice (3) bez posledního členu $\varepsilon(0, w_t \,.\, \sigma_x^2)$ – v bodě $\hat{\theta} \equiv (\hat{b}_x, \hat{c}_x, \hat{d}_x, \hat{e}_x)$ v budoucích časech *t*. Hodnota P_x^{t0} (vstupní počet pojištěnců v prahovém čase projekce) ve výpočetním předpisu (5) figuruje jako absolutní

člen (a_x) pětiparametrické lineární kumulativní regresní funkce, který v rámci minimalizace součtu čtverců ad (4) není zapotřebí odhadovat. Tímto postupem se dosáhne spojitého napojení regresní funkce na počet pojištěnců v prahovém čase t_0, což má význam zejména pro krátkodobé prognózování.

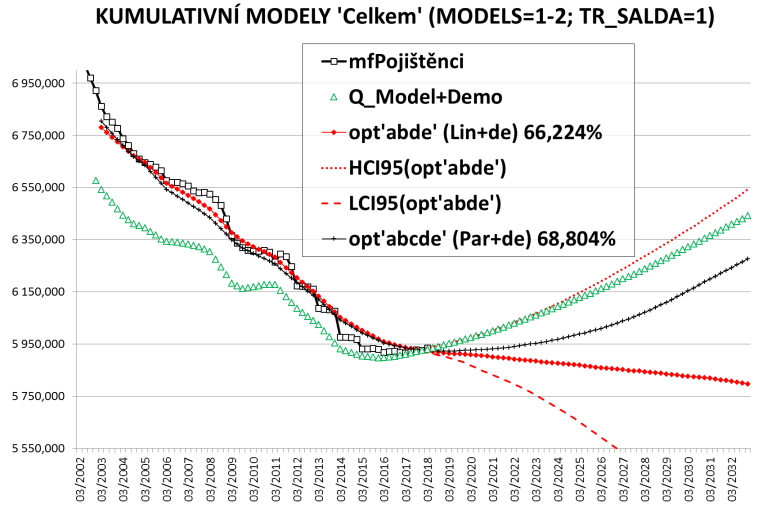
Rovnice (3) resp. (5) reprezentuje ovšem v každé věkové skupině hned několik lineárních regresních modelů, když některé z parametrů b_x, c_x, d_x nebo e_x uvažujeme jako rovné nule. Nejefektivnější (nepřeparametrizované) regresní modely bychom získáme hierarchickým testováním hypotéz o jejich nulových hodnotách prostřednictvím normálních statistických testů (zde máme na mysli testy založené na předpokladu normálního t.j. Gaussovského rozdělení chybových členů $\varepsilon(0, w_t \cdot \sigma_x^2)$ v rovnici (3)). Jelikož ale o významu demografické a přeregistrační proměnné v podstatě nemáme (po experimentech s mechanizmovými modely) pochybnosti, zajímá nás ponejvíce testování hypotézy 'H_0: c_x=0' ověřující legitimitu kvadratického členu u časové proměnné *t*. Diskusi k této problematice, kde se ukázala jako nejpodstatnější volba diskontního parametru α_x, provádíme v závěrečné stati 3.

3 Výsledky

Více než technologie odhadu parametrů v regresních modelech výše je důležitá věrohodnost predikčních křivek ve vztahu k jejich skutečnému budoucímu vývoji. Spolehlivost predikčních odhadů můžeme měřit např. porovnáním modelových hodnot s retrospektivním vývojem v minulosti. Pokud máme k dispozici dostatečně flexibilní třídu predikčních odhadů, můžeme se pokusit predikovanou časovou řadu v dané třídě parametrických odhadů optimalizovat. Takovou možnost nemáme v případě mechanizmových modelů představených v rámci stati 2.1; naopak u zobecněných regresních modelů umožňuje volba váhových-diskontních faktorů α_x až nečekaně variabilní a senzitivní scénáře budoucího vývoje pojištěnců VZP ČR. Pouze částečně indikují možné problémy údaje v Tabulce 1, kde jsou uvedeny důležité parametry pro dva nejobecnější predikční modely – 'abcde' (plně odhadnutý čtyřparametrický regresní model (5) s kvadratickým členem u časové proměnné) a 'abde' (tříparametrický model s kvadratickým členem fixně definovaným jako nula).

V Tabulce 1 jsou pro oba modely uvedena procenta vysvětleného rozptylu (R^2), které pro normální regresní modely již samy osobě představují klíčovou charakteristiku pro posouzení vhodnosti výběru regresních proměnných. V případě zobecněných lineárních modelů se ale koeficienty determinace (R^2) ukázaly jako silně závislé na volbě diskontního parametru α_x ($0<\alpha_x\leq 1$). Vzhledem k spojitě-nelineárním průběhům procenta vysvětleného rozptylu v jednotlivých věkových skupinách byly tedy za optimální považovány obvykle ty volby α_x, kde R^2 dosáhl svého maxima. Jak vyplývá z údajů v Tabulce 1, v řadě případů odpovídala maxima hodnotám α_x=1 (tedy modelům nevážené lineární regrese), ale u několika věkových skupin byla v rozpětí prakticky aplikovatelných hodnot α_x ($0{,}65<\alpha_x$) shledána i hodnotově srovnatelná lokální maxima dvě – takže finální výběr α_x byl učiněn až po pečlivějších expertizních (a vizualizačních) analýzách.

Graf 2 – Vývoj kvartálních počtů pojištěnců VZP ČR za období 03/2002–12/2017 a časové řady jejich teoretických počtů generovaných dvěma alternativami regresních modely (pro střední variantu extrapolace přeregistračního salda do roku 2032) ve srovnání s mechanizmovým modelem a pásy spolehlivosti



Graf 2 – Historické kvartální počty pojištěnců VZP ČR (v tisících) jsou zobrazeny černými čtverci (časová řada 'mfPojištěnci'). Pro plně parametrizovaný regresní model z (3) – označme jej formálně 'abcde' – optimalizovaný v rámci kumulativních vzorců (5) a vyšcítaný z úrovně všech 18 věkových skupin – je na Grafu 2 znázorněný černými křížky. Jednodušší alternativní model 'abde' (odhadnutý nezávisle při volbě c_x=0) je zobrazen plnými ležatými (červenými) kosočtverci. Na obrázku jsou vyznačeny rovněž 95%ní pásy spolehlivosti (v oblasti vymezené křivkami 'HCI95(opt'abde')' a 'LCI95(opt'abde')', což jsou horní a dolní meze pro model 'abde' odhadnutý účelově pro jedinou součtovou skupinu '0+'). Termínem „optimalizované" míníme modely, které zaznamenaly co možná nejvyšší % vysvětleného rozptylu (R^2) při systematických změnách diskontního parametru ax ($0<\alpha_x\leq 1$), zvlášť v rámci každé pětileté věkové skupiny – podrobněji viz Tabulka 1. Procenta (68,804% resp. 66,224%) uvedená pro oba alternativní modely v Legendě na obrázku byla vyčíslena jako prostý průměr v rámci všech 18ti věkových skupin („x"='0-4', … ,'85+'). Pro vizuální porovnání výsledků je na Grafu 2 zobrazena i časová řada hodnot predikovaných mechanizmovým modelem (1) a (2) z Grafu 1 (řada 'Q_Model+Demo' – zelené trojúhelníky).

Již z údajů v Tabulce 1 jasně vyplývá, že na většině řádků se výsledky pro oba srovnávané modely příliš neliší. To znamená, že ani přírůstek R^2 evidovaný ve prospěch čtyřparametrického modelu 'abcde' není statisticky významný. Tento fakt bychom potvrdili i na úrovni indikací statistické významnosti individuálních parametrů tohoto modelu, kterých by v celkovém souhrnu ani nebylo více než pro redukovaný model 'abde'. (Statisticky významné koeficienty c_x u kvadratických členů modelu 'abcde' jsme shledali pouze ve skupinách '5–9'

a '35–39' – proto tyto údaje z úsporných důvodů ani detailně neuvádíme). Naopak statistická významnost formálních parametrů v redukovaném modelu 'abde' (viz indikace hvězdičkami v prostředních sloupcích Tabulky 1) jasně vypovídá o významu lineárního členu (zejména ve všech vyšších věkových skupinách, počínaje '45+', ale také u '0–4' a '10–19'). Významnost přeregistračního členu je nejsilnější ve skupinách '5–9' (p=0,002), '10–14' (p=0,019) a v několika dalších. Ve všech pětiletých věkových skupinách ale dominuje svojí statistickou významností aditivní člen demografický (p<0,001).

Je důležité si všimnout, že na úrovni modelů odhadovaných marginálně pro celý pojistný kmen (uvažovaný jako agregátní skupina '0+') vychází demografický člen d_x jako nesignifikantní (na rozdíl od lineárního a přeregistračního). Tato skutečnost je způsobena nejen formálně nízkými procenty vysvětleného rozptylu na úrovni celkových počtů, ale jednoduše i tím, že dřívější vysoké meziroční úbytky počtů pojištěnců VZP fakticky nelze vysvětlovat dlouhodobě stabilizovaným vývojem celkových počtů obyvatelstva ČR. To bylo základním „kamenem úrazu" již při hledání metodické koncepce vhodné třídy predikčních regresních, neboť demografická složka se výrazně uplatňuje až s přechodem k jednotlivým věkovým skupinám.

Ve třech věkových skupinách ('15–19', '65–69', '80–84') byly za optimální diskontní faktory α_x pro redukovaný model ('abde') použity hodnoty znatelně nižší než u úplného modelu. U skupiny '15–19' to vede dokonce k výrazně vyšším hodnotám R^2 ve prospěch redukovaného modelu (v situaci téhož α_x by ale musela platit opačná nerovnost). Ve skupině '5–9' byla hodnota α_x=0,69 (pro oba modely) vybrána až jako "2. lokální maximum shledané v rámci analýzy spojitého průběhu R^2", a to z důvodu jasně věrohodnějšího průběhu predikované regresní křivky. V obdobném smyslu bylo v pořadí až 2. lokální maximum u R^2 pro α_x vybráno u skupin '5–9' a '80–84', protože v rámci regionálních rozpočtů z ČR na 14 krajů ČR by hodnoty extrapolované pro 1. lokální maximum v individuálních regionech překračovaly rámec věrohodných budoucích krajských podílů VZP na trhu zdravotního pojištění.

Tabulka 1 – Základní parametry dvou nejvýznamnějších zobecněných regresních modelů ad (3) optimalizovaných za účelem predikce počtu pojištěnců VZP ČR 'Celkem" a v rámci pětiletých věkových skupin (optimální diskontní faktory α_x, % vysvětleného rozptylu a vyznačení statistické významnosti parametrů modelu 'abde', tj. modelu odhadnutého za podmínky c_x=0)

"x"	α('abde')	R^2(‚abde')	$b_x\|c_x$=0	$d_x\|c_x$=0	$e_x\|c_x$=0	α('abcde')	R^2(‚abcde')
'0+'	0,85	16,59%	-13,16*	1,661	0,619*	1,00	13,28%
'0–4'	1,00	27,44%	0,212	0,861***	0,477	1,00	32,14%
'5–9'	0,69	52,68%	-0,674	1,335***	0,895**	0,69	52,68%
'10–14'	0,99	82,47%	-1,03***	1,217***	0,623*	0,99	82,75%
'15–19'	0,86	76,50%	-0,961**	1,161***	0,476*	1,00	64,91%
'20–24'	1,00	39,75%	-0,241	1,208***	0,691	1,00	41,69%
'25–29'	1,00	47,21%	-0,288	1,04***	0,231	1,00	49,12%
'30–34'	1,00	66,04%	0,081	0,789***	0,528	1,00	66,18%
'35–39'	0,93	72,07%	-0,619	0,719***	0,159	0,94	73,45%
'40–44'	0,99	49,01%	-0,464	0,877***	0,826	1,00	49,47%
'45–49'	1,00	83,03%	-1,079***	1,109***	0,683*	1,00	83,23%
'50–54'	0,96	71,66%	-1,006**	0,994***	0,69	0,96	71,77%
'55–59'	1,00	79,01%	-0,77**	1,027***	0,68*	1,00	79,09%
'60–64'	1,00	79,93%	-0,895**	1,039***	0,408	1,00	79,98%
'65–69'	0,75	85,02%	-1,24***	1,145***	0,968*	1,00	83,65%
'70–74'	1,00	93,05%	-1,274***	1,034***	-0,025	1,00	93,15%
'75–79'	0,94	85,73%	-0,654***	0,898***	1,613*	0,94	85,82%
'80–84'	0,86	36,66%	-0,568***	1,235***	-0,42	1,00	82,29%
'85+'	1,00	64,78%	0,397*	0,751***	2,348	1,00	67,10%
mean	0,943	66,225%				0,973	68,804%

Legenda: V jednotlivých sloupcích tabulky jsou uspořádány: optimální diskontní faktory ax a % vysvětleného rozptylu pro dva nejdůležitější zobecněné lineární modely ('abde' a 'abcde') a odhadnuté formální parametry modelu 'abde' s vyznačením jejich statistické významnosti vůči teoretickým nulovým hodnotám (resp. ** nebo *** značí statistickou významnost na hladině významnosti 0,001 resp. 0,01 nebo 0,05); na základě těchto parametrizací byly vytvořeny predikční křivky regresních modelů zobrazených na Grafu 1.*

4 Závěry

V rámci excelovské aplikace vyvinuté na půdě VZP v letech 2017–18 se automaticky generují všechny časové řady predikované na základě mechanizmových i zobecněných regresních modelů, a to dokonce pro všechny hierarchické submodely, které lze odvodit z modelu 'abcde' pro libovolně zvolená α_x ($0<\alpha_x\leq 1$) v 18ti pětiletých věkových skupinách. Vizualizované průběhy predikčních křivek pro specifické regresní submodely umožňují srovnání s korespondujícími mechanizmovými predikčními modely (za korespondující lze považovat nikoli pouze 'abde' resp. 'abcde' vs. 'Q_Model+Demo' z Grafu 1, ale také např. 'abd' resp. 'abcd' vs. 'Q_Demo' z Grafu 1 zahrnující pouze demografickou složku

anebo analogicky regresní modely 'abe' resp. 'abce' vs. 'Q_Model' akceptující pouze složku přeregistrační). Pro optimalizované parametry dvou nejvýznačnějších regresních modelů ('abde' resp. 'abcde') se agregátní výsledky pro ČR rozpočítávají na úroveň 14ti krajů ČR, kde má VZP již historicky vzešlé velmi rozdílné podíly na trhu zdravotního pojištění. Z těchto úhlů pohledu tedy predikce založené na zobecněných regresních modelech jasně překonávají dřívější modely mechanizmové, které jsou zde nyní zahrnuty jako speciální případy.

Složitost systematického metodického zpracování této predikční úlohy spočívá v tom, že v jednotlivých pětiletých věkových skupinách se aktuální dynamika vývoje počtu pojištěnců nachází vždy v odlišné fázi růstu nebo poklesu kmene pojištěnců, která je podmíněna aktuálními zářezy ve věkové struktuře obyvatelstva ČR. Navíc se v jednotlivých věkových skupinách uplatňují lokální, ale spojité přeregistrační trendy odvíjené od specifické situace na trhu zdravotního pojištění. Zobecněné regresní modely použité výše mají schopnost velmi senzitivně formalizovat demografickou i přeregistrační složku, u deseti pětiletých věkových skupin – povětšinou těch nejstarších – je formálně nezbytné zahrnout do modelu i formální (sestupný) lineární trend, ve skupině '85+' se formální lineární trend indikuje statisticky významně jako vzestupný.

Největší problémy s použitím zobecněných lineárních modelů byly spojeny s kalibrací diskontních parametrů α_x, které byly v rámci třídy modelů 'abde' a 'abcde' optimalizovány v každé věkové skupině expertním výběrem z nejvýše dvou hodnot, kde procenta vysvětleného rozptylu (R^2) nabývala svého lokálního maxima. Tento postup maximalizuje informaci obsaženou v rámci demografické a přeregistrační nezávisle proměnné. Další rozvoj modelu je možný korekturami výše stanovených kalibračních parametrů αx na základě znalosti budoucího kvartálního vývoje počtů pojištěnců VZP ve skupinách.

Literatura

[1.] Běláček J, Fiala T, Parma M, Foks R: Civilizační nemoci, věkové stárnutí obyvatelstva a data o pacientech ZZ AGEL. IX. Symposium AGEL, Olomouc 1.– 2.10. 2015, poster

[2.] Běláček J, Fiala T, Parma M, Foks R, Murtingerová K: Projekce nemocnosti v kontextu stárnutí obyvatelstva a poskytovaných zdravotních služeb v ČR 2012–14. Forum Statisticum Slovacum 4/2015, 120–128, Slovenská Štatistická a Demografická Společnost, Eds.: Chajdiak J, Luha J, Madarász Š, Rev.: Chajdiak J, Luha J, Korónyi S

[3.] Běláček J, Fiala T, Parma M, Michna P, Lukeš K, Murtingerová K: Projekce budoucí potřeby a spotřeby zdravotní péče z perspektivy stárnutí ambulantních pacientů ZZ AGEL 2012-14. In.: Sborník příspěvků MEDSOFT 2017, 4–18. Hotel Academic, Roztoky u Prahy, 21.–22.3.2017

[4.] Běláček J.: Predikce budoucích počtů pojištěnců VZP ČR – data, metodika a výsledky. In.: Sborník příspěvků MEDSOFT 2018, 7–19. Hotel Academic, Roztoky u Prahy, 20.–21.3.2018

[5.] Bělohradský A-Šolc Z: Predikce příjmů veřejného zdravotního pojištění. Metodické compendium. Ministerstvo financí České republiky, 2018, http://www.mfcr.cz/studie

[6.] Cipra T.: Analýza časových řad s aplikacemi v ekonomii. Praha, SNTL/Alfa,1986

[7.] Projekce obyvatelstva ČR do r. 2100, ČSÚ, 2013; https://www.czso.cz/csu/czso/projekce-obyvatelstva-ceske-republiky-do-roku-2100-n-fu4s64b8h4

[8.] Věkové složení obyvatelstva ČR, ČSÚ, 2012, …, 2016; https://www.czso.cz/csu/czso/vekove-slozeni-obyvatelstva-2016

[9.] Projekce obyvatelstva v krajích ČR – do r. 2050, https://www.czso.cz/csu/czso/projekce-obyvatelstva-v-krajich-cr-do-roku-2050-ua08v25hx9

Kontakt

Jaromír Běláček, RNDr., CSc.
VZP ČR
Orlická 4
130 00 Praha 3
e-mail: *jaromir.belacek@vzp.cz*