ZÁKLADNÍ POJMY A PRINCIPY KONSTRUKCE MODELŮ TYPU VĚK-OBDOBÍ-KOHORTA

Jindra Reissigová[1] – Jitka Rychtaříková[2]

THE BASIC CONCEPTS AND PRINCIPLES OF THE CONSTRUCTION OF AGE-PERIOD-COHORT MODELS

Abstract
The aim of the article is to examine the age-period-cohort models that are used to evaluate the trends of various population indicators (e.g. mortality, fertility). This approach is mainly used when we have no available data on the potential risk or protective factors (e.g. lifestyle) affecting population indicators. The advantages and disadvantages of age-period-cohort modelling are described and their use is illustrated on the basis of examples of Czech male mortality.

Keywords: Lexis diagram, age-period-cohort, identification problem, generalised linear model, prediction, male mortality, Czech Republic

Demografie, 2015, **57: 21–39**

1. ÚVOD

Již od 70. let minulého století se modelují trendy nejrůznějších demografických, sociologických a epidemiologických ukazatelů v závislosti na věku, období a kohortě. V tomto smyslu hovoříme o modelech typu věk-období-kohorta (Age-Period-Cohort models, APC). Od té doby do současnosti se tento přístup analýzy dat neustále vyvíjí a zdokonaluje. Metody modelování APC jsou publikovány v řadě příspěvků a shrnuty v monografiích (*Hobcraft et al.*, 1982; *Caselli – Capocaccia*, 1989; *Wilmoth*, 2006; *Yang – Land*, 2013; *O'Brien*, 2014).

Modely APC se aplikují především v *deskriptivních studiích*, jež si kladou za cíl popsat rozložení onemocnění, úmrtí či jiných událostí v populaci a rozlišit vlivy věku, období a kohorty na hodnoty zkoumaných ukazatelů. Výpočty se obvykle opírají o data pocházející z národních registrů, běžné evidence a dalších zdrojů. Jako příklad jejich použití můžeme uvést studie, které se zabývaly analýzou trendů nemocnosti, úmrtnosti, anebo plodnosti v České republice (*Gelnarová et al.*, 2007; *Reissigová – Tomečková*, 2008; *Katrňák*, 2009). Výsledky modelů APC také slouží jako podklady pro tvorbu hypotéz, které například mohou ukazovat na možný příčinný vztah mezi určitými faktory a rozvojem nemoci. Platnost těchto hypotéz se posléze může ověřit v *analytických* nebo *intervenčních studiích*, které jsou designovány tak, aby potencionální příčinný vztah dokázaly vyhodnotit (*Bencko et al.*, 2003). Kromě toho se modely APC také využívají k predikci budoucích

1) Ústav informatiky AV ČR, v.v.i. v Praze, Oddělení medicínské informatiky a biostatistiky. Pod Vodárenskou věží 271/2, 180 07 Praha 8, reissigova@cs.cas.cz.

2) Přírodovědecká fakulta UK v Praze, Katedra demografie a geodemografie.

trendů ukazatelů. Například finský onkologický registr (Finnish Cancer Registry) použil modelování APC pro odhad vývoje onkologických onemocnění až do roku 2020 (*Finnish Cancer Registry*, 2009).

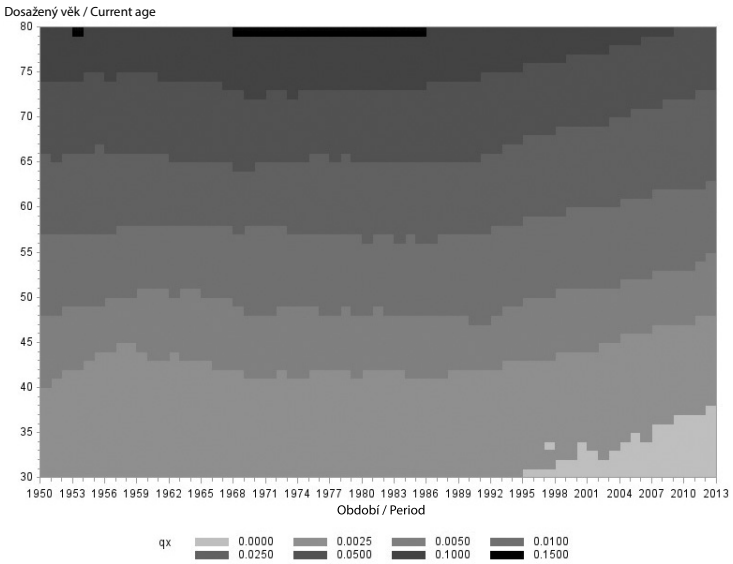
Tento článek je určen všem, kteří se zabývají analýzou trendů a mají alespoň základní znalosti statistických metod. Jeho cílem je seznámit čtenáře s podstatou modelování trendů metodou APC (část 2–4), která je jednou z používaných metod statistického vyhodnocování trendů populačních ukazatelů. Přehled základních používaných přístupů tohoto modelování je prezentován spolu s analýzou úmrtnosti mužů v České republice (část 5). Jsou také popsány doporučované postupy pro výběr správného modelu (část 6). V závěru (část 7) se shrnují možnosti využití modelování APC.

2. MOTIVAČNÍ PŘÍKLADY: UKAZATELE ÚMRTNOSTI MUŽŮ V ČR

V České republice docházelo po druhé světové válce k významným změnám v úmrtnostních trendech (*Rychtaříková*, 2005). Tyto změny se lišily podle věku i v čase a je otázkou do jaké míry se na nich podílel faktor kohorty, tj. příslušnosti k určité generaci definované rokem narození. V demografické literatuře byl zejména popsán vývoj české úmrtnosti v čase a podle věku (*Rychtaříková*, 2004; *Burcin – Kučera*, 2008). V námi studovaném období, po roce 1950, lze rozlišit v České republice tři rozdílné etapy. Do počátku šedesátých let střední délka života při narození mužů narůstala zejména v souvislosti s poklesem míry kojenecké úmrtnosti a také snižováním úrovně úmrtnosti mladších věkových skupin, což dokumentuje graf 1 (metoda konstrukce ukazatele q_x a popis dat viz příloha). Od poloviny šedesátých let do konce osmdesátých let 20. století lze pozorovat stagnaci, respektive rozšiřování vyšších úmrtnostních hladin do mladšího věku (graf 1). Koncem osmdesátých let začala zřetelněji narůstat naděje dožití a tento příznivý trend trvá dodnes. Na nedávném příznivém obratu se podílejí zejména střední a starší věkové skupiny (graf 1). Trendy úmrtnosti z generačního pohledu jsou méně známé (graf 2). Lze vystopovat dva odlišné vzorce od pravidelného trendu. První se týká generací mužů narozených během první světové války. Tito muži měli ve středním věku nižší úroveň úmrtnosti v porovnání se staršími, ale i mladšími generacemi. Druhá odchylka souvisí s nedávným zlepšováním úmrtnostních poměrů, které se postupně promítalo do snižování hodnot pravděpodobností úmrtí jednotlivých generací. Na tomto novém obratu se podíleli všichni muži od 30 do 80 let. Grafy 1 a 2 (mapy intenzit úmrtnosti na základě izočar neboli vrstevnic) znázorňují vždy kombinaci dvou proměnných: věku a období, respektive věku a generace, přičemž třetí dimenze, v prvním případě generace, ve druhém případě období je v pozadí, protože vypočítané pravděpodobnosti vyjadřují vždy kombinaci obou vlivů tj. jak období, tak kohorty (generace). Proto je důležité pomocí pokročilejších analytických metod dezintegrovat všechny tři efekty současně, což právě řeší modely APC.

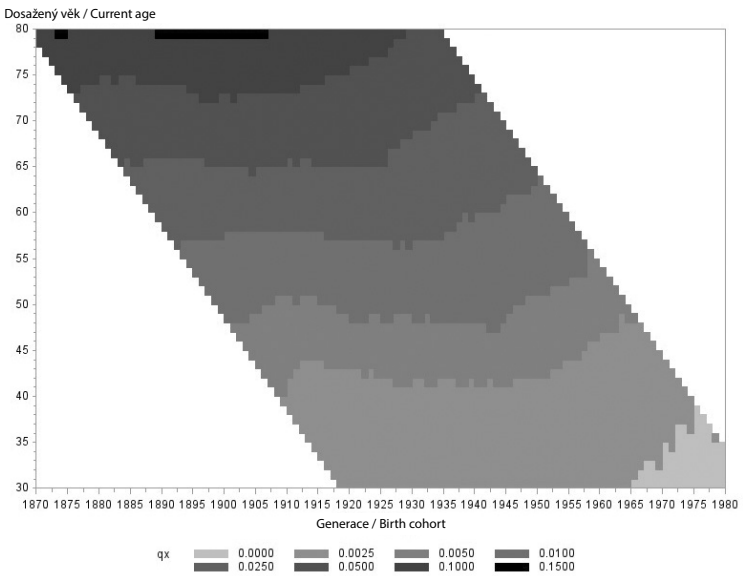
Jiným příkladem dat, k jejichž analýze se používají modely APC, jsou agregovaná data v tabulce 1. Na takto uspořádaná data se aplikují modely v případě, kdy počty událostí (např. úmrtí) pro jednotlivé roky věku a období jsou malé a jejich analýzou bychom dostali nestabilní odhady, anebo pokud roční data nejsou vůbec dostupná. V tabulce 1 jsou míry úmrtnosti mužů na ischemickou chorobu srdeční (ICHS) v České republice agregované do pětiletých věkových skupin a období. Například míra úmrtnosti v prvním sloupci a ve třetím řádku v tabulce 1 je 81,1. To znamená, že zemřelo 81,1 mužů z 100 000 mužů, kterým bylo 40–44 let ($a = 3$) v letech 1980–1984 ($p = 1$). Z toho vyplývá, že tito muži se museli narodit mezi 1.1.1935 a 31.12.1944. Jinými slovy řečeno, muž ze třetí věkové skupiny ($a = 3$) a prvního období ($p = 1$) náleží do sedmé kohorty narozených ($c = 7$), neboť platí $c = A - a + p = 9 - 3 + 1 = 7$, kde $A = 9$ (celkový počet věkových skupin). Všimněme si však, že roky narození sousedních kohort z tabulky 1 se vzájemně překrývají, a proto jsou kohorty (generace) definovány v tomto případě jen aproximativně na rozdíl od kohort v minulém příkladu. Jaký je vliv věku, období a kohorty na úmrtnost se v tomto případě vyjadřuje odpovídajícími kategoriemi (pětiletými pro věk a období, desetiletými pro kohorty), a proto se např. předpokládá, že vliv věků 30, 31, 32, 33 a 34 na míru úmrtnosti je stejný, což vyjadřuje věková kategorie 30–34.

Graf 1: Pravděpodobnosti úmrtí (q_x), muži, ČR, období 1950–2013
Probability of death (q_x), men, Czech Republic, period 1950–2013



Zdroj: ČSÚ.
Source: *Czech Statistical Office (CZSO).*

Graf 2: Pravděpodobnosti úmrtí (q_x), muži, ČR, generace 1870–1980
Probability of death (q_x), men, Czech Republic, 1950–2013 birth cohorts



Zdroj: ČSÚ.
Source: *Czech Statistical Office (CZSO).*

Tab. 1: Míry úmrtnosti na ICHS (na 100 tis.), muži, ČR, 1980–2004*)
Mortality rate from ischaemic heart disease (per 100 thous.), men, Czech Republic, 1980–2004*)

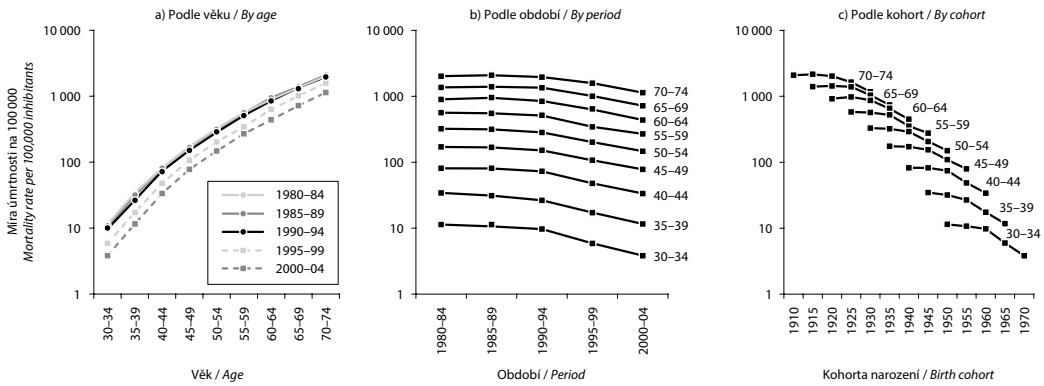
Dokončený věk *Completed age (a)*	Období / *Period (p)*					Generace *Birth cohort (c)*
	1980–1984 (1)	1985–1989 (2)	1990–1994 (3)	1995–1999 (4)	2000–2004 (5)	
30–34 (1)	11,3	10,6	9,7	5,9	3,8	
35–39 (2)	34,3	31,3	26,5	17,3	11,6	1965–1974 (13)
40–44 (3)	81,1	80,6	73,1	47,9	33,6	1960–1969 (12)
45–49 (4)	171,4	168,9	151,6	107,5	77,9	1955–1964 (11)
50–54 (5)	322,4	314,8	283,8	202,3	147,3	1950–1959 (10)
55–59 (6)	565,3	553,5	510,9	345,7	270,1	1945–1954 (9)
60–64 (7)	899,9	950,1	851,6	631,9	438,4	1940–1949 (8)
65–69 (8)	1 363,9	1 396,0	1 355,0	1 011,4	720,9	1935–1944 (7)
70–74 (9)	2 022,4	2 088,6	1 960,2	1 587,9	1 137,1	1930–1939 (6)
Generace / *Birth cohort (c)*		1905–1914 (1)	1910–1919 (2)	1915–1924 (3)	1920–1929 (4)	1925–1934 (5)

Zdroj: ÚZIS, ČSÚ.
Source: *Institute of Health Information and Statistics of the Czech Republic (IHIS), CZSO.*
Pozn.: *) Čísla v závorkách označují pořadí kategorie.
Note: **) Numbers in parentheses indicate the order of the category.*

Data z tabulky 1 jsou znázorněna v grafu 3. Je vidět, že každá ze tří složek (věk, období, kohorta narození, tj. generace) má na vývoj měr úmrtnosti na ICHS svůj specifický vliv (*Carstensen*, 2007). Míry úmrtnosti rostou s věkem (graf 3a), snižují se v čase (s obdobím pozorování, graf 3b) a s rokem narození (graf 3c).

Na ukazatele úmrtnosti (grafy 1, 2 a 3) může mít vliv celá řada rizikových a protektivních faktorů životního stylu, životního prostředí, životní úrovně anebo genové dědičnosti. Pokud nemáme k dispozici spolehlivá data o potencionálních faktorech (např. o kouření, tělesné aktivitě, stresu, znečištění ovzduší emisemi výfukových plynů) anebo tyto faktory není

Graf 3: Míry úmrtnosti mužů na ICHS, muži, ČR, 1980–2004*)
Mortality rate from ischaemic heart disease, men, Czech Republic, 1980–2004*)



Zdroj: Reissigová – Tomečková, 2008.
Source: *Reissigová – Tomečková, 2008.*
Pozn.: *) Kohorty narození jsou označeny středem 10letého intervalu, např. kohorta narození 1940 reprezentuje kohortu mužů narozených od 1. 1. 1935 do 31. 12. 1944.
Note: **) Birth cohorts are identified by the middle interval, e.g. the 1940 birth cohort represents the cohort of men born between 1 January 1935 and 31 December 1944.*

ani možné spolehlivě kvantifikovat za delší časové období, můžeme ukazatele úmrtnosti analyzovat právě na základě uvedených tří časových veličin: věku, období a kohorty.

A co vlastně vyjadřují jednotlivé časové veličiny? Vliv věku odráží biologický proces stárnutí člověka a vyjadřuje životní etapu, v níž se právě člověk nachází. Například starší osoby jsou více ohroženy úmrtím či kardiovaskulárními onemocněními. Obdobím se míní časový úsek, ve kterém osoby žijí. Jeho vliv se projevuje politickými, ekonomickými, technickými, sociálními a jinými změnami, které probíhají v daném období (např. změna společenského režimu, nové diagnostické metody či nová klasifikace onemocnění). Tyto změny jsou celospolečenského charakteru. Kohortou se zde rozumí generace jedinců narozených ve stejném kalendářním roce anebo jiném časovém úseku. Vliv kohorty je spojen s životním stylem, a tedy dlouhotrvajícími zvyky a návyky typickými pro určité generace (např. kouření, počet dětí anebo věk při porodu). Kohorta prochází různými změnami v čase jakoby společně. Jestliže nějaká změna ovlivní jednu kohortu, nemusí ovlivnit jinou, neboť změnu mohla zažít v jiném roce svého života (vliv věku) a za jiných podmínek (vliv období). Podrobněji jsou složky věku, období a kohorty například popsány (v českém jazyce) v článku (*Katrňák*, 2009), v němž se rozebírají i z pohledu panelového výzkumu (typ longitudinální studie, v níž se té samé skupině jedinců kladou opakovaně v určitých intervalech stejné otázky, např. volební preference).

Závěrem této části shrňme, že grafické znázornění (grafy 1, 2 a 3) je důležitým pomocníkem k pochopení dat, nedává však odpověď na otázku, jak dalece jsou ukazatele úmrtnosti ovlivněny věkem, jak dalece obdobím a jak dalece kohortou narození, neboť tyto tři veličiny působí na míry úmrtnosti simultánně. A právě odpověď na uvedenou otázku nám pomáhají dát modely APC, jež se snaží vliv věku, kohorty a období separovat a kvantifikovat. Modely APC jsou speciálním typem zobecněných lineárních modelů (generalised linear model), jak uvidíme v následující části 3 (*Pekár – Brabec*, 2009).

3. MATEMATICKÉ VYJÁDŘENÍ MODELŮ APC

Počet úmrtí d_{ap} v a-té věkové skupině a p-tém období ($a = 1,..,A$, $p = 1,..,P$) je možno považovat za veličinu s Poissonovým rozdělením. Poissonovým rozdělením se řídí náhodné veličiny, které vyjadřují, kolikrát nastane nějaká málo pravděpodobná událost v populaci většího rozsahu v určitém období, prostoru či jinak definovaném úseku. Přičemž pravděpodobnost výskytu jedné události je úměrná délce úseku a události se vyskytují nezávisle na sobě. Cílem modelování APC je popsat závislost počtu úmrtí d_{ap} (nebo počtu výskytů jiných událostí, které se řídí Poissonovským rozdělením) na veličinách věku (α_a, $a = 1,..,A$), období (β_p, $p = 1,..,P$) a kohortě (γ_c, $c = 1,..,C$) Poissonovým regresním modelem jako (*Clayton – Schifflers*, 1987)

$$\ln(d_{ap}) = \mu + \alpha_a + \beta_p + \gamma_c + \ln(n_{ap}) + \varepsilon_{ap},$$

kde μ je absolutní člen (intercept) čili průměrná zlogaritmovaná hodnota ukazatele bez ohledu na věk, období a kohortu, n_{ap} je velikost populace a ε_{ap} je náhodná chyba (odhadujeme ji jako klasické reziduum, jak si ukážeme později). V grafech 1 a 2 je velikost populace n_{ap} reprezentována počátečním stavem mužů (tj. počtem mužů ke dni 1.1. daného kalendářního roku), $A = 51$ (počet věkových tříd: 30–80), $P = 64$ (počet období: 1950–2013) a $C = 114$ (počet generací, tj. kohort: 1870–1983). V tabulce 1 je n_{ap} rovno střednímu stavu mužů (tj. počtu mužů ke dni 1. 7. daného kalendářního roku), $A = 9$, $P = 5$ a $C = 13$.

Poissonova regrese je speciálním typem zobecněných lineárních modelů. Algebraickými úpravami můžeme Poissonoův regresní model vyjádřit jako

$$\ln(r_{ap}) = \mu + \alpha_a + \beta_p + \gamma_c + \varepsilon_{ap},$$

$$r_{ap} = \exp(\mu + \alpha_a + \beta_p + \gamma_c + \varepsilon_{ap}),$$

kde na levé straně rovnice je místo absolutního počtu událostí d_{ap} (např. úmrtí) ukazatel $r_{ap} = d_{ap} / n_{ap}$ (např. pravděpodobnost úmrtí, míra úmrtnosti, ap). Zde je na místě upozornit na to, že krajní kohorty jsou založeny na menším počtu pozorování než ostatní kohorty. Například v tabulce 1 máme u kohort narozených 1905–1914 (γ_1) a 1965–1974 (γ_{13}) jen jeden údaj o úmrtnosti na rozdíl od kohorty 1935–1944 (γ_7), u níž máme pět údajů o úmrtnosti. Následkem toho intervalové odhady odpovídajících parametrů založených na menším počtu pozorování bývají méně spolehlivé (jejich

intervaly spolehlivosti jsou širší), jak uvidíme později v grafu 4c.

Jestliže počet událostí nelze považovat za veličinu s Poissonovým rozdělením, nýbrž za veličinu s binomickým rozdělením, vyjadřuje se závislost logistickou regresí, která také patří mezi takzvané zobecněné lineární modely, jako

$$\ln(\frac{r_{ap}}{1 - r_{ap}}) = \mu + \alpha_a + \beta_p + \gamma_c + \varepsilon_{ap}.$$

Poissonovo rozdělení je limitním případem binomického rozdělení. V praxi se binomické rozdělení zpravidla aproximuje Poissonovým rozdělením, pokud je pravděpodobnost výskytu události malá (nižší než 10 %) a sledovaný počet osob je vyšší než 30.

Parametry zobecněných lineárních modelů se zpravidla odhadují metodou maximální věrohodnosti. Jednoduše řečeno touto metodou se získají odhady parametrů, které jsou pro pozorovaná data (v našem případě míry nebo pravděpodobnosti, tj. kvocienty úmrtnosti) nejpravděpodobnější.

4. PROBLÉM IDENTIFIKACE

Jak již víme z Úvodu, každá z analyzovaných proměnných (věk, období, kohorta) má své teoretické opodstatnění. Navzdory tomuto faktu je však matematicky obtížné vliv jednotlivých složek změřit, a to z důvodu deterministické provázanosti složek. Když známe hodnoty dvou složek, dokážeme určit třetí. Například když víme, kolik je člověku let (*věk*) v určitém čase (*období*), dokážeme říci, kdy se narodil (*kohorta = období – věk*), a tedy k jaké kohortě narozených patří.

V důsledku uvedené kolinearity (každá složka je lineární funkcí zbylých dvou), je problematické odlišit vliv jednotlivých složek na analyzovanou událost. Model zahrnujícím současně věk, období a kohortu má více parametrů než může být z dat odhadnuto. Tím pádem neexistuje jednoznačné řešení odhadů parametrů a hovoříme o problému identifikace či identifikačním problému (identification problem). Z tohoto důvodu se někdy prezentuje jen podrobná grafická analýza dat a doporučuje se, aby se problém identifikace neřešil prostřednictvím statistických modelů, ale méně formalizovanými postupy jakými je například kontextuální analýza (*Glenn*, 2003).

5. ŘEŠENÍ PROBLÉMU IDENTIFIKACE

Problému identifikace bychom se mohli vyhnout, kdybychom se místo zobecněného lineárního modelu se třemi proměnnými omezili na model pouze s dvěma proměnnými (two-factor generalised linear model), např. na model s věkem a kohortou, $\ln(r_{ap}) = \mu + \alpha_a + \gamma_c + \varepsilon_{ap}$. V tomto případě totiž nejsou složky modelu deterministicky provázané (nejsou kolineární) a jejich odhady získáme klasickými statistickými metodami (*Pekár – Brabec*, 2009). Model s dvěma proměnnými se obvykle používá ve standardní demografické analýze nebo byl například použit při analýze sociální fluidity (přechod z jedné sociální vrstvy do druhé s ohledem na třídní původ) ve Švédsku (*Breen – Jonsson*, 2007). Na tomto místě, je však nutné zdůraznit, že k vyloučení proměnné z modelu musíme být věcné důvody podpořené statistickou analýzou dat (*Mason et al.*, 1973; *Kupper et al.*, 1985; *Fienberg – Mason*, 1985; *Holford*, 1991).

Jiné přístupy obcházejí problém identifikace tím, že se zaměří jen na analýzu vlivu kohorty, který vyjadřují jako parciální interakci vlivu věku a období, např. metoda mediánového vyhlazování (median polish method) (*Keyes – Li*, 2010). Navzdory složitosti problému bylo však také publikováno mnoho návrhů, jak se s problémem identifikace vypořádat v případě, že do hry vstupují všechny tři proměnné (věk, období, kohorta). V následujících částech některé tyto metody představíme.

5.1 ZOBECNĚNÉ LINEÁRNÍ MODELY S OMEZENÍMI

Když jsou proměnné (věk, období, kohorta) kategorizované veličiny, existuje řada možností, jak můžeme vyjádřit jejich vliv (*Řeháková*, 2008). V kontextu modelování APC se obyčejně jeden parametr věku, období a kohorty zvolí jako nulový, např. $\alpha_5 = 0$, $\beta_2 = 0$, a $\gamma_7 = 0$. Nulové parametry slouží jako referenční kategorie, vzhledem k nimž se porovnávají odhadované parametry ostatních kategorií. Tento typ vyjádření hodnot parametrů se odborně nazývá umělé kódování (dummy coding) kontrastů. Naznačeným způsobem bychom odhadovali a vyjadřovali parametry modelu, kdyby proměnné nebyly kolineární. Při existenci problému identifikace se však musí klást na hodnoty parametrů další požadavky, aby šlo hodnoty parametrů metodou maximální věrohodnosti odhadnout.

Již v 70. letech minulého století se doporučovalo zvolit dva parametry věku, období anebo kohorty jako shodné (*Mason et al.*, 1973). U jedné složky (např. u období β_p) se tak namísto jednoho omezení (referenční kategorie, např. druhé období, $\beta_2 = 0$) zvolí ještě rovnost hodnot parametrů dvou kategorií (např. druhého a třetího období, $\beta_2 = \beta_3$). Tím se zruší kolinearita mezi proměnnými a parametry modelu jsou odhadnutelné. Potom o modelech hovoříme jako o zobecněných lineárních modelech s omezeními (constrained generalised linear models, CGLM). Je však důležité zdůraznit, že různá omezení (tj. jaké parametry zvolíme sobě rovné) vedou k různým hodnotám odhadovaných parametrů. Z toho vyplývá, že kladené požadavky musí mít svá věcná opodstatnění, to však nebývá vždy nasnadě.

Při statistické analýze měr úmrtnosti na ICHS z tabulky 1 byl odhadnut model ve tvaru $r_{ap} = \exp(\mu + \alpha_a + \beta_p + \gamma_c + \varepsilon_{ap})$. Odhad μ je –5,89. Odhady parametrů α_a, β_p a γ_c jsou vyjádřeny ve formě relativních rizik v grafu 4. Referenční kategorií pro hodnocení vlivu věku na míru úmrtnosti byla zvolena pátá věková kategorie, tj. $\alpha_5 = 0$, $\exp(\alpha_5) = 1$. Jak je z grafu 4a vidět, se zvyšujícím věkem se zvyšovala míra úmrtnosti na ICHS.

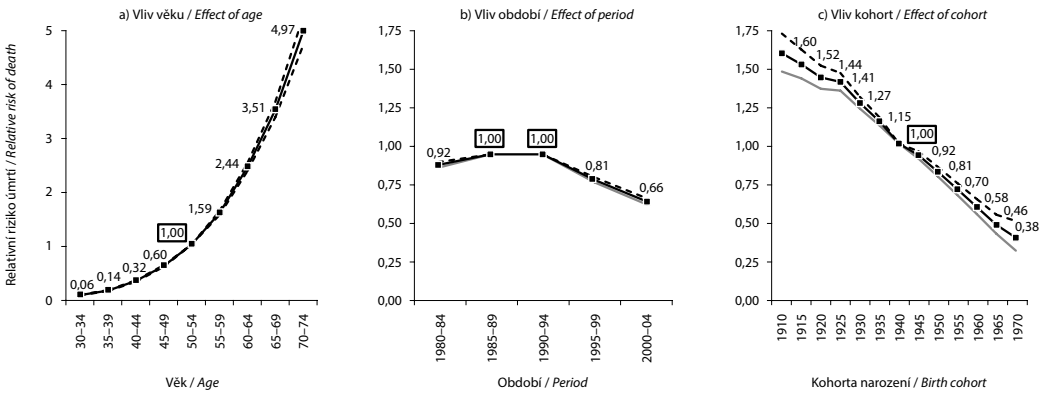
Referenční kategorií pro hodnocení vlivu období byla zvolena druhá kategorie ($\beta_2 = 0$), přičemž se předpokládala rovnost vlivu druhého a třetího období ($\beta_2 = \beta_3$), a proto $\exp(\beta_2) = \exp(\beta_3) = 1$. Jinými slovy, zvolily se jako shodné parametry období 1985–1989 a 1990–1994, referenční období bylo tedy desetileté období (1985–1994), a to z důvodu, že riziko úmrtí na ICHS bylo v těchto letech kolem pádu komunismu v roce 1989 přibližně stejné (*Reissigová – Tomečková*, 2008). Jak je vidět z grafu 4b, za tohoto předpokladu se riziko úmrtí v posledních letech snížilo. Za referenční kohortu v grafu 4c byla zvolena kohorta 1940. To je sedmá kohorta, jež vycházela z dostatečného počtu pozorování ($\gamma_7 = 0$, $\exp(\gamma_7) = 1$). Riziko úmrtí se snižovalo s kohortou narození (porovnáváno s referenční kohortou 1940).

Na základě modelu, jehož odhady parametrů jsou v grafu 4, můžeme odhadovat míry úmrtnosti v tabulce 1. Například pro věkovou skupinu 45–49 ($a = 4$) v letech 1980–84 ($p = 1$), to je kohortu 1930–39 ($c = A - a + p = 9 - 4 + 1 = 6$), platí

$$r_{41} = \exp(\mu + \alpha_4 + \beta_1 + \gamma_6) = \exp(\mu) \cdot \exp(\alpha_4) \cdot \exp(\beta_1) \cdot \exp(\gamma_6) = 0{,}0028 \cdot 0{,}60 \cdot 0{,}92 \cdot 1{,}15 = 0{,}001749,$$

Graf 4: Zobecněný lineární model s omezeními*): vlivy věku, období a kohort (95% interval spolehlivosti) na úmrtnost na ICHS, muži, ČR, 1980–2004

Generalised linear model with restrictions*): effects of age, period and cohorts (95% confidence interval) on ischaemic heart disease mortality, men, Czech Republic, 1980–2004



Zdroj: Reissigová – Tomečková, 2008.
Source: *Reissigová – Tomečková, 2008.*
Pozn.: *) Období 1985–1989 a 1990–1994 a kohorta narození 1940 byly vybrány jako referenční kategorie. Kohorty narození jsou označeny středem intervalu, např. kohorta narození 1940 reprezentuje kohortu mužů narozených 1. 1. 1935 do 31. 12. 1944.
Note: **) The 1985–1989 and 1990–1994 periods and the 1940 birth cohort were selected as the refence categories. Birth cohorts are identified by the middle interval, e.g. the 1940 birth cohort represents the cohort of men born between 1 January 1935 and 31 December 1944.*

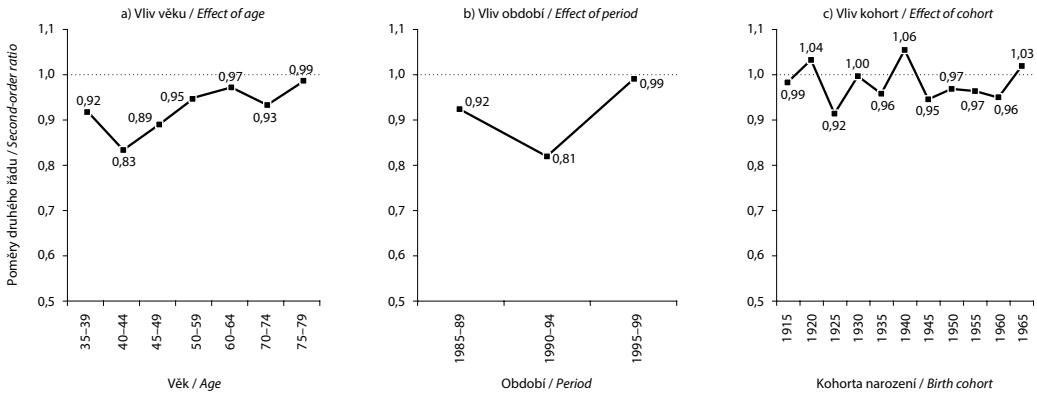
neboť exp(μ) = exp(–5,89) = 0,0028 a zbývající tři hodnoty exponenciální funkce jsou vyčísleny v grafu 4. To znamená, že modelem odhadovaná míra úmrtnosti se rovná 174,9 na 100 tis. mužů, přičemž skutečně pozorovaná je 171,4 (viz tab. 1), a tedy reziduum e_{41} = 171,4 – 174,9 = –3,5. Poznamenejme, že hodnota 0,001749 je vypočtena z nezaokrouhlených odhadů parametrů.

Aby řešení nebylo závislé na libovolně zvolených a často neobhajitelných omezeních hodnot, byla od 80. let minulého století navržena celá řada postupů, jak vyjádřit vlivy věku, období a kohorty. Jednou z možností je prezentovat místo odhadovaných parametrů (α_1, α_2, α_3, α_4 atd.) takzvané *rozdíly druhého řádu*, které jsou definovány jako ($\alpha_3 - \alpha_2$) – ($\alpha_2 - \alpha_1$), ($\alpha_4 - \alpha_3$) – ($\alpha_3 - \alpha_2$) atd. Pro různé modely (tj. modely s různými omezeními hodnot parametrů) jsou tyto rozdíly stejné (*Clayton – Shifflers*, 1987). To má na jedné straně význam, ale na druhé straně je nutné si uvědomit, že rozdíly druhého řádu popisují jen změny ve vývoji vlivu věku, období a kohorty a nepopisují vývoj jejich trendů. Jinak řečeno, hodnoty rozdílů druhého řádu, které se podstatně liší od nuly, upozorňují na to, kdy došlo k nějakému většímu zvratu (pozitivnímu nebo negativnímu) ve vlivu věku, období či kohorty na studovaný ukazatel (např. míru úmrtnosti). Uveďme příklad.

Touto metodou jsme vyhodnocovali míru úmrtnosti na ICHS z tabulky 1 (*Reissigová – Tomečková*, 2008). Výsledky jsou prezentovány grafu 5. Rozdíly druhého řádu jsme vyjádřili v ekvivalentní formě jako poměry, např. (exp(α_3)/exp(α_2))/(exp(α_2)/exp(α_1)), a potom jsme sledovali, které hodnoty se výrazně liší od hodnoty jedna. Například z grafu 4a vyplývá, že (87,7/38,6)/(38,6/15,6) = 0,92, jak je vidět v grafu 5a. Přelomovým věkem pro muže byl věk kolem 40/50 let (období andropauzy), kdy došlo k největšímu zpomalení nárůstu úmrtnosti (poměry druhého řádu v grafu 5a jsou 0,83 a 0,89). Úmrtnost se výrazně snížila po roce 1990 (poměr druhého řádu v grafu 5b je 0,81). To se vysvětluje pádem komunismu, jenž přinesl radikální změnu životního stylu a modernější léčbu. Na druhé straně nebyly pozorovány žádné větší výkyvy úmrtnosti mezi sousedícími kohortami, jelikož poměry druhého řádu kolísaly kolem hodnoty jedna, graf 5c. Tyto výsledky jsou víceméně v souladu se závěry, které byly učiněny na základě zobecněného lineárního modelu s omezeními, viz graf 4.

Obdobné výsledky bychom dostali, kdybychom parametry věku, období a kohorty odhadnuté při jakémkoli omezení jejich hodnot aproximovali

Graf 5: Poměry druhého řádu: vlivy věku, období a kohort*) na úmrtnost na ICHS, muži, ČR, 1980–2004
Second-order differences: effects by age, period and cohort*) on ischaemic heart disease mortality, men, Czech Republic, 1980–2004



Pozn.: *) Kohorty narození jsou označeny středem intervalu, např. kohorta narození 1940 reprezentuje kohortu mužů narozených 1. 1. 1935 do 31. 12. 1944.
Note: **) Birth cohorts are identified by the middle interval, e.g. the 1940 birth cohort represents the cohort of men born between 1 January 1935 and 31 December 1944.*

metodou lineární regrese (*Holford*, 1983). Jinými slovy, kdybychom lineární regresí popsali závislost odhadnutých parametrů věku na věku a analogicky totéž provedli pro období a kohorty (*Holford*, 1983). Odpovídající *rezidua* (rozdíly mezi odhadnutými hodnotami parametrů v modelu APC a jejich hodnotami aproximovanými lineární regresí) jsou totiž stejná při jakémkoli omezení jejich hodnot. Odlehlé hodnoty reziduí však opět pouze poukazují na to, kdy došlo ke změně trendu vlivu věku, období a kohorty, jako jsme ukázali v předešlé metodě. Touto metodou se analyzovala například úmrtnost na rakovinu prostaty v USA v letech 1935–1969 (*Holford*, 1983).

Jiná takzvaná sekvenční metoda vychází z předpokladu, že nejdůležitějším faktorem je věk, druhým kohorta a nejméně důležitým období (pozn. důležitost kohorty a období lze prohodit), (*Carstensen*, 2007). Tato metoda spočívá v tom, že se v prvém kroku do modelu vloží pouze parametry věku a kohorty, a potom se vypočítají rezidua, tj. rozdíly mezi pozorovanou hodnotou a hodnotou odhadnutou tímto modelem. V druhém kroku se do modelu zahrnou jen parametry období a jejich hodnoty se odhadnou podmíněně na hodnotách zmíněných reziduí.

Aby řešení problému identifikace nebylo závislé na dodatečných omezeních hodnot parametrů, hledaly se i jiné metody odhadu. Ty jsou však již výpočetně složitější a pouze se o nich stručně zmíníme.

5.2 ZOBECNĚNÉ ADITIVNÍ MODELY

V zobecněných aditivních modelech (generalised additive model, GAM) se vliv proměnných vyjadřuje skrze vyhlazovací funkce (smooth functions), takže problém identifikace odpadá. V těchto modelech se s věkem, obdobím a kohortou pracuje jako se spojitými veličinami. Zobecněné aditivní modely jsou rozšířením zobecněných lineárních modelů. Věk, období a kohorta jsou vyhlazovány například *polynomickými funkcemi* (*Verdecchia – De Angelis – Capocaccia*, 2002),

$$\ln(r_{ap}) = \mu + \sum_{i=1}^{I} \alpha_i a^i + \sum_{j=2}^{J} \beta_j p^j + \sum_{k=1}^{K} \gamma_k (p-a)^k + \varepsilon_{ap},$$

kde a označuje věk a p období (ostatní parametry definovány v části 3). Výraz $\beta_1 p$ nefiguruje v modelu, aby se zamezilo problému identifikace ($c = p - a$). Vyhlazování polynomickými funkcemi se například použilo při analýze pocitu štěstí lidí v USA, kdy se vliv věku na pocit štěstí vyjádřil kvadratickou funkcí (*Yang et al.*, 2008). K vyhlazování se vedle polynomických funkcí také používají spline funkce (*Heuer*, 1997; *Carstensen*, 2007; *Jiang – Carriere*, 2013),

$$\ln(r_{ap}) = \mu + f(a) + g(p) + h(c) + \varepsilon_{ap},$$

kde $f(a)$, $g(p)$ a $h(c)$ jsou spline funkce věku, období a kohorty. Jednoduše řečeno, spline funkce jsou po částech (na každé předdefinované části daného intervalu) polynomické funkce, které v krajních bodech (uzlech) na sebe navazují (piecewise polynomial function). Vyhlazování prostřednictvím spline funkcí se využilo například k analýze incidence zlomenin kyčle ve vztahu k politickým a ekonomickým událostem v Portugalsku v letech 2000–2008 (*Alves et al.*, 2013).

Zobecněné aditivní modely jsou užitečné především k analýze dat tabulovaných pro jednotlivé roky věku v ročních obdobích. Mohou se však aplikovat i na data agregovaná, např. do pětiletých intervalů jako v tabulce 1. Zobecněné aditivní modely bývají kritizovány za to že, volba vyhlazovací funkce nemusí být jednoduchá. Kromě toho je nutné zdůraznit, že zobecněné aditivní modely řeší problém identifikace přechodem od linearity k nelinearitě. Tím se modelování stává složitějším a někdy je to na úkor dat, neboť nebýt problému identifikace, zákonitosti v datech by šlo možná popsat lineárně.

5.3 MODELY CHARAKTERISTIK VĚKU, OBDOBÍ A KOHORTY

Modely charakteristik věku, období a kohorty (Age-Period-Cohort Characteristic Models, APCC) spočívají v tom, že se do zobecněného lineárního modelu APC místo věku, období nebo kohorty vkládají takzvané *zástupné proměnné* (proxy variables), které na rozdíl od těch původních nevykazují kolinearitu. Odůvodňuje se to tím, že věk, období a kohorta pouze zastupují neměřitelné primární příčiny, a ty mohou být popsány i jinými zástupnými proměnnými než jsou věk, období a kohorta. Například místo kohorty se do modelu vkládá proměnná vyjadřující relativní velikost kohorty (*Kahn – Mason*, 1987; *O'Brien*, 2000), nebo místo období se uvažuje míra nezaměstnanosti (*Pavalko et al.*, 2007). Winship a Harding navrhli

strategii, jak prostřednictvím zástupných proměnných specifikovat mechanismus působení věku, období a kohorty na sledovanou událost (*Winship – Harding*, 2008). Autoři k tomu využívají modelovaní pomocí strukturálních rovnic (structural equation model). Na jedné straně zástupné proměnné řeší problém identifikace, na druhé straně však vyvstává problém jiný. Zástupné proměnné (pokud je vůbec máme k dispozici) nemusí být dostatečně reprezentativní, aby postihly vliv věku, období nebo kohorty v celé jejich šíři, jak se o tom píše i v úvodu článku.

5.4 METODA INTRINSICKÉHO ODHADU

Metoda intrinsického (vnitřního) odhadu (intrinsic estimator, IE) se snaží odhadnout parametry modelu bez zástupných veličin s minimem doplňujících matematických předpokladů (*Fu*, 2000; *Yang et al.*, 2008). Zjistilo se, že každý odhad $\hat{b}$ vektoru parametrů $b = (\mu, \alpha_1,.., \alpha_{a-1}, \beta_1,.., \beta_{p-1},.., \gamma_1,.., \gamma_{a+p-2})^T$, může být rozložen do dvou částí jako $\hat{b} = B + t \cdot B_0$, kde B je takzvaný intrinsický odhad vektoru parametrů b, t je realné číslo specifické pro daný odhad $\hat{b}$ a B_0 je vlastní vektor designové matice X. Každou kategorizovanou proměnnou s I kategoriemi (v našem případě věk, období, kohorta) je možné popsat $I - 1$ takzvanými designovými proměnnými, jejichž hodnoty závisí na typu zvolených kontrastů a tvoří designovou matici. Protože vlastní vektor B_0 není závislý na libovůli řešitele ani na hodnotách závislé proměnné, ale jen a pouze na matici X, odhadované hodnoty parametrů nezávisí na žádných libovolně stanovených omezeních.

Touto metodou byla hodnocena například data týkající se (živě) narozených dětí v České republice od konce druhé světové války do roku 2007 (*Katrňák*, 2009). Metoda intrinsického odhadu se také využila v amerických studiích náboženské aktivity a víry (*Schwadel*, 2011), těžkého epizodického pití (*Keyes – Miech*, 2013), psychické úzkosti (*Keyes et al.*, 2014) a interpersonální důvěry (*Clark – Eisenstein*, 2013).

Přestože se do této metody vkládají velké naděje, i ona má své kritiky. Těm se omezení intrinsického modelu nezdají nezanedbatelná, zůstávají podle nich abstraktní a jsou těžko srozumitelná (*O'Brien*, 2011; *Luo*, 2013). Tomu zastánci metody oponují tím, že modely s intrinsickým odhadem vykazují lepší výsledky než zobecněné lineární modely s omezeními a zobecněné aditivní modely (*Yang et al.*, 2004; *Yang et al.*, 2008; *Fu et al.*, 2011; *Yang – Land*, 2013).

5.5 HIERARCHICKÝ ZOBECNĚNÝ LINEÁRNÍ MODEL

Pokud máme k dispozici data například z opakovaných průřezových studií (repeated cross-sectional studies) můžeme zkusit aplikovat hierarchický (víceúrovňový) zobecněný lineární model (hierarchical (multilevel) generalised linear model, HGLM) (*Yang – Land*, 2006; *Yang et al.*, 2006). V něm se věk a případně další použité veličiny zjištěné ve studii (např. vzdělání, příjem) považují za fixní (fixed effects) a období a kohorta za veličiny s náhodnými efekty (random effects). Pojem fixní vyjadřuje, že se s veličinami pracuje na individuální úrovni. Jejich vliv se považuje za stejný napříč obdobími a kohortami. Náhodné efekty vyjadřují skupinovou podstatu veličiny, a to že hodnoty závislé proměnné jsou v rámci jednotlivých období a kohort korelovány.

Autoři této metody prokládají (aproximují) vliv věku kvadratickou regresí, aby se odstranila lineární závislost mezi věkem, obdobím a kohortou (*Yang – Land*, 2006). Hierarchický model APC specifikují jako model náhodných křížených efektů (Cross-Classified Random Effects Model, CCREM). To znamená, že vliv každého období je odvozen jako průměrný napříč všemi kohortami a naopak vliv každé kohorty jako průměrný napříč všemi obdobími. Hierarchický zobecněný lineární model se použil například k rozboru postoje vůči předmanželskému pohlavnímu styku ve Spojených státech od roku 1975 do roku 2008 (*Elias et al.*, 2013). Někteří kritici této metody nedoporučují aplikovat modely HGLM, neboť z jejich pohledu kvadratické proložení věku a náhodné efekty problém identifikace neřeší (*Bell – Jones*, 2014).

Přestože se modely HGLM používají především pro data z opakovaných průřezových studií, pro názornost vyjádříme hierarchický zobecněný lineární model pro data prezentovaná v grafech 1 a 2. Parametry modelu jsou odhadnuty v grafu 6 výše popsanou metodou, která je také dostupná z internetu (http://yangclaireyang.web.unc.edu/age-period-cohort-analysis-new-models-methods-and-empirical-applications/chapter-7/) (*Yang – Land*, 2006). Věk figuroval v modelu jako fixní veličina, období a kohorta jako

veličiny s náhodnými efekty. Zobrazené výsledky znázorňují známou skutečnost lineárního narůstání (logaritmické měřítko) intenzity úmrtnosti s věkem, zde mezi 30 a 80 roky, tento vliv je nejsilnější. Vliv období potvrzuje již dříve popsané trendy, a to snižování intenzity úmrtnosti do počátku šedesátých let 20. století, které je zde vzhledem k uvažované věkové skupině 30–80 let méně výrazné, protože prodlužování naděje dožití při narození souviselo v tomto období především se snižováním míry kojenecké úmrtnosti. Období od počátku šedesátých let do konce osmdesátých let 20. století je známé zhoršováním úmrtnostních poměrů středního a vyššího věku, což průběh odhadnutého parametru potvrzuje. Poslední časová fáze je ve znamení velmi výrazného poklesu. Toto je rovněž v souladu s rychlejším snižováním mužské úmrtnosti pozorované tentokrát ve starším a středním věku. Průběh parametru měřícího vliv kohorty je nové zjištění, které není v rozporu s dřívějšími výsledky získanými jinou metodou (*Rychtaříková et al.*, 1994). Kohortní vliv je, pokud jde o úmrtnost, nejslabší. Na území českých zemí, muži narození v letech 1910–1920 měli relativně nižší úroveň úmrtnosti než muži narození dříve nebo později. Naopak ti, co se narodili v období druhé světové války a těsně po ní, budou pravděpodobně v dalším přežívání méně favorizovaní. Počáteční a koncové kohorty nejsou hodnoceny, protože hodnoty parametru vycházejí z malého počtu pozorování.

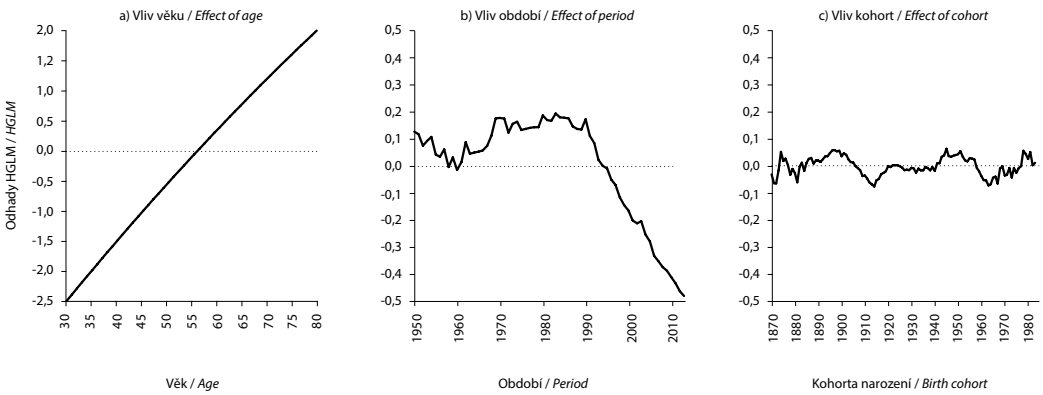
5.6. BAYESOVSKÉ ODHADY

V literatuře jsou popsány nejen maximálně věrohodné odhady vlivu věku, období a kohorty, ale i jejich bayesovské odhady. Bayesovské metody předpokládají, že odhadované parametry (α_a, β_p, γ_c) jsou náhodné veličiny s nějakým rozdělením (není to tedy pevná časově neměnná konstanta jako u maximálně věrohodných odhadů). Apriorní představa o tomto rozdělení (apriorní rozdělení) parametrů se kombinuje s rozdělením pozorovaných dat. O výsledném rozdělení mluvíme jako o aposteriorním rozdělení. To je určeno ze vztahu, který využívá Bayesovu větu (Thomas Bayes – anglický matematik 18. století), a proto hovoříme v tomto smyslu o bayesovských zobecněných lineárních modelech (Bayesian generalised linear model). Někdy se vychází z předpokladu, že sousední parametry věku, období a kohorty se mění postupně. Jinými slovy se předpokládá, že rozdíly $\alpha_a - \alpha_{a+1}$, $\beta_p - \beta_{p+1}$, $\gamma_c - \gamma_{c+1}$ ($a = 1,.., A-1$, $p = 1,.., P-1$, $c = 1,.., C-1$) jsou blízké nule. Za tímto účelem se definuje, že vektor parametrů má apriorně mnohorozměrné normální rozdělení. Tento postup se použil při analýze trendu počtu vražd v USA (*Nakamura*, 1986). Bayesovské odhady se použily i v řadě jiných studií (*Berzuini et al.*, 1993; *Berzuini – Clayton*, 1994; *Bray*, 2002; *Bashir – Estève*, 2001).

6. DOPORUČOVANÉ POSTUPY MODELOVÁNÍ APC

Základním předpokladem každé správné statistické analýzy jsou spolehlivá data a jejich dostatečné množství.

Graf 6: Hierarchický zobecněný lineární model: vlivy věku, období a kohort na pravděpodobnost úmrtí, muži, ČR, 1950–2013
Hierarchical generalised linear model: effects of age, period and cohort on probability of death, men, Czech Republic, 1950–2013



V našem případě mohou být data ovlivněna administrativními změnami (např. změna klasifikace onemocnění) anebo je může znehodnotit nespolehlivá registrace onemocnění (např. pohlavní choroby se často zatajují, neboť nemocní bývají spojování s dehonestující pověstí). Abychom mohli vyhodnotit vliv kohorty na sledovanou událost, doporučuje se mít k dispozici data alespoň za posledních 20 let. Než začneme s odvozováním modelu, nejprve bychom si měli data graficky zobrazit, abychom provedli jejich kontrolu a udělali si základní představu o jejich časových trendech (grafy 1, 2 a 3).

6.1 ODVOZOVÁNÍ MODELU

Obecně se doporučuje testovat, které z proměnných (α_a, β_p, γ_c) budou v modelu zastoupeny, hierarchicky (*Clayton – Schifflers*, 1987). Tento postup shrnuje tabulka 2 (*Arbyn et al.*, 2002; *Carstensen*, 2007). Protože nejdůležitější proměnnou je zpravidla věk, začíná se s Modely 1 až 3.1 (Modely 2.1 a 2.2 jsou ekvivalentní). V dalším kroku se do modelu přidá období (Model 3.2), resp. kohorta (Model 3.3). Na závěr se zkoumá simultánní vliv věku, období a kohorty (Model 4). Jestliže se k analýze použijí například zobecněné aditivní modely (spline funkce), doporučuje se postupovat analogicky podle tabulky 2 (parametry α_a, β_p, γ_c v tabulce 2 se nahradí spline funkcemi věku $f(a)$, období $g(p)$ a kohorty $h(c)$). Jaký model je nejvhodnější pro daná data, se rozhoduje na základě statistických metod, které jsou shrnuty v části 6.2.

Modely 1–4 jsou ilustrovány na mírách úmrtnosti na ICHS (tab. 1) v grafu 7. Modely se vlastně odlišují, jak popisují na logaritmické stupnici trendy měr ve věkových skupinách: Model 1 – konstantní trendy, Model 2.1 (Model 2.2) – stejná směrnice lineárních trendů (přímky rovnoběžné), Model 3.1 – různé směrnice lineárních trendů (přímky různoběžné), Model 3.2 (vliv období) – stejné nelineární trendy (křivky rovnoběžné), Model 3.3 (vliv kohort) – stejné nelineární trendy (křivky rovnoběžné) a Model 4 (vliv období a kohort) – různé nelineární trendy (křivky různoběžné). Pro uvedené míry úmrtnosti je nejvhodnější Model 4 (*Reissigová – Tomečková*, 2008). Odhadované míry úmrtnosti na základě tohoto Modelu 4 jsou v grafu 7f; odpovídající pozorované míry úmrtnosti jsou v grafu 3c.

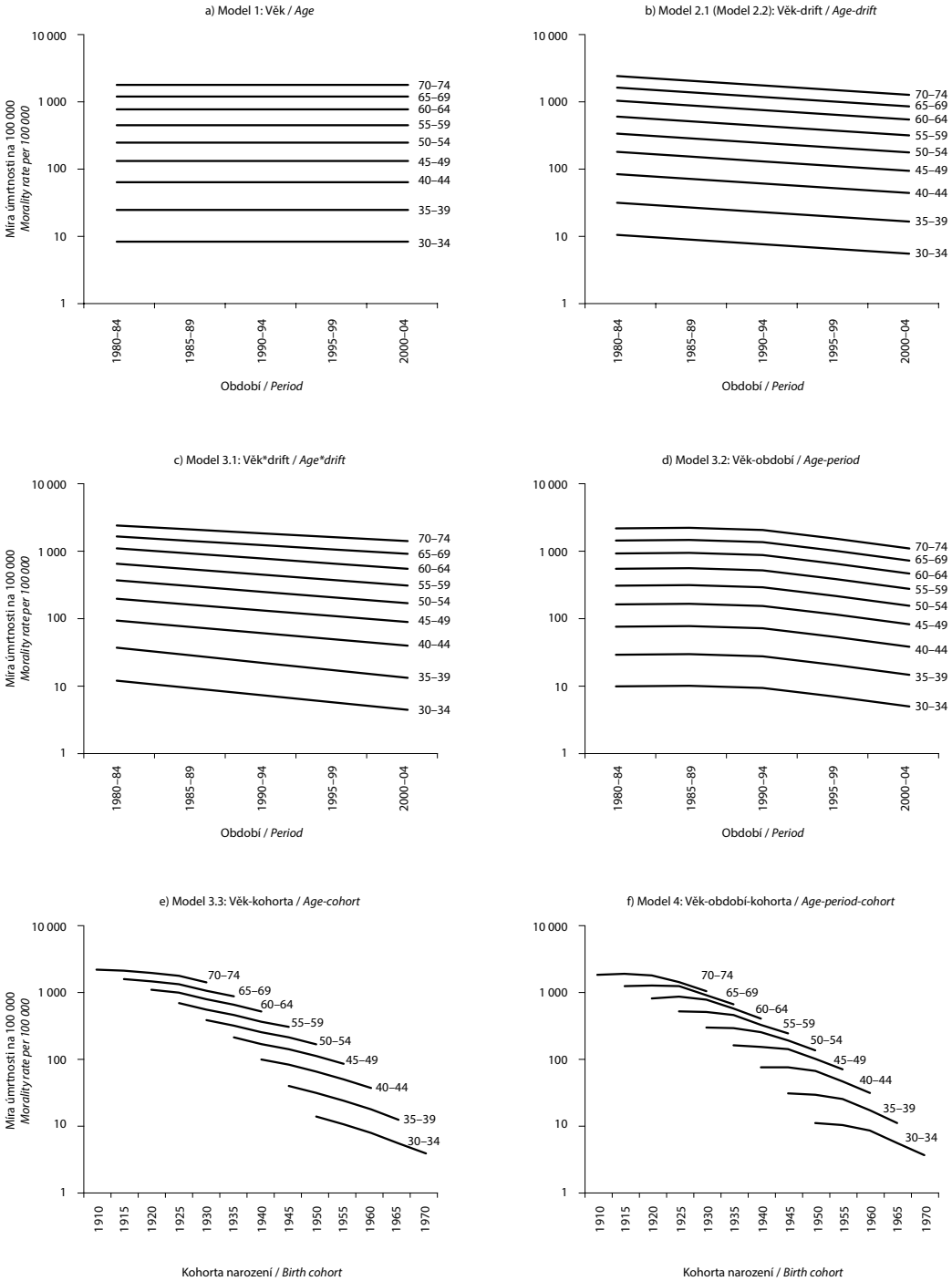
Pokud data nejlépe popisuje právě Model 4, nastává problém identifikace parametrů (tj. jakým způsobem vyjádřit vliv věku, období a kohort). V poslední době se doporučuje parametry modelu vyjadřovat metodou intrinsického odhadu (*Fu*, 2000; *Yang et al.*, 2008). Nicméně je ku prospěchu věci odhadnout parametry modelu více metodami a výsledky mezi sebou porovnat. Byla publikována celá řada studií, které porovnávají různé přístupy modelování dat (*Yang et al.*, 2004; *Smith*, 2008).

Tab. 2: Hierarchický postup testování modelů
Hierarchical procedure for testing models

Číslo *Number*	Model / *Model*	Matematické vyjádření za platnosti Poissonova rozdělení *) *Mathematical expression of the Poisson distribution *)*
1	Věk / *Age*	$\ln(r_{ap}) = \mu + \alpha_a + \varepsilon_{ap}$
2.1	Věk-drift / *Age-drift*	$\ln(r_{ap}) = \mu + \alpha_a + \delta_{period}(p - p_0) + \varepsilon_{ap,}$ kde δ_{period} je lineární trend (drift, slope) a p_0 je referenční kategorie *where δ_{period} is the linear trend (drift, slope) and p_0 is the reference category*
2.2	*Věk-drift* / Age-drift	$\ln(r_{ap}) = \mu + \alpha_a + \delta_{cohort}(c - c_0) + \varepsilon_{ap,}$ kde δ_{cohort} je lineární trend (drift, slope) a p_0 je referenční kategorie *where δ_{cohort} is the linear trend (drift, slope) and p_0 is the reference category*
3.1	Věk*drift / *Age*drift*	$\ln(r_{ap}) = \mu + \alpha_a \cdot \delta_{age.period}(p - p_0) + \varepsilon_{ap,}$ kde $\delta_{age.period}$ je lineární trend (drift, slope) a p_0 je referenční kategorie *where $\delta_{age.period}$ is the linear trend (drift, slope) and p_0 is the reference category*
3.2	Věk-období / *Age-period*	$\ln(r_{ap}) = \mu + \alpha_a + \beta_p + \varepsilon_{ap}$
3.3	Věk-kohorta / *Age-cohort*	$\ln(r_{ap}) = \mu + \alpha_a + \gamma_c + \varepsilon_{ap}$
4	Věk-období-kohorta *Age-period-cohort*	$\ln(r_{ap}) = \mu + \alpha_a + \beta_p + \gamma_c + \varepsilon_{ap}$

Pozn.: *) a je věková skupina, $a = 1,.., A$ (A počet věkových skupin); p je období, $p = 1,.., P$(P počet období). Za platnosti binomického rozdělení by místo $\ln(r_{ap})$ byl $\ln(r_{ap}/(1 - r_{ap}))$.
Note: **) a is the age group, $a = 1,.., A$ (A is the number of age groups); p is the period, $p = 1,.., P$(P is the number of periods). Under the binomial distribution, instead of $\ln(r_{ap})$ is $\ln(r_{ap}/(1 - r_{ap}))$.*

Graf 7: Zobecněné lineární modely: odhadovaná míra úmrtnosti na ICHS, muži, ČR, 1980–2004
Generalised linear model: estimated mortality rate from ischaemic heart disease, men, Czech Republic, 1980–2004



6.2 KVALITA MODELU

Aby na datech odvozený model byl validní, musí vykazovat dobré statistické vlastnosti. Jejich podrobný popis není cílem tohoto článku. Pro data z tabulky 1 je výběr modelu podrobněji popsán v jiné publikaci (*Reissigová – Tomečková*, 2008). Shrňme si alespoň obecné principy modelování dat (*Pekár – Brabec*, 2009). K rozhodnutí, kterou proměnnou (věk, období, kohorta) zařadit do modelu, nám pomáhá Waldův test a test poměrem věrohodnosti. Celková vhodnost regresního modelu se posuzuje na základě hodnot deviance, zobecněných koeficientů determinace či Pearsonova chí-kvadrát testu. To jsou takzvané míry dobré shody, které kvantifikují, jak kvalitně navržený model aproximuje experimentální data. Vhodnost modelu se vyhodnocuje i graficky, a to zobrazením reziduí (např. deviačních, Pearsonových). Rezidua vypovídají o vlivu jednotlivých pozorování na kvalitu modelu. Vedle reziduí se používají i jiné diagnostické nástroje k odhalování vlivných pozorování, které mohou stát za nefunkčností modelu (např. Cookova vzdálenost). Složitost modelu se kvantifikuje informačními kritérii (např. Akaikeho, Bayesovským), které nám pomáhají odvodit model s adekvátním počtem proměnných. Není totiž pravda, že čím více proměnných bude v modelu, tím je model lepší. I zde totiž platí, že v jednoduchosti je krása. Přeparametrizovaný model je v praxi nepoužitelný, neboť má malou vypovídající hodnotu. Nakonec nezbývá než dodat, že je nutné i ověřit, zda Poissonův model nevykazuje nadměrný anebo naopak nedostatečný rozptyl, např. Cameronovým-Trivediovým testem. Typickou vlastností Poissonova rozdělení totiž je, že se rozptyl rovná střední hodnotě. Jestliže rovnost neplatí, lze přejít od Poissonova modelu například ke kvazi-Poissonovskému modelu (v případě lineárního vztahu mezi rozptylem a střední hodnotou) anebo k negativně binomickému modelu (v případě kvadratického vztahu).

6.3 SOFTWARE

Statistickou analýzu APC je možné provést volně dostupným programovacím softwarem R, který nabízí celou řadu specifických statistických aplikací k modelování APC (*R Development Core Team*, 2012). Pomocí něho je možné například aplikovat zobecněné aditivní modely, a to buď prostřednictvím knihovny *Epi* pro statistickou analýzu v epidemiologii, anebo spuštěním programů *Nordpred*, které byly vytvořeny pro predikci trendů incidence rakoviny v Norsku (*Møller et al.*, 2002). Pro software R byly vyvinuty i programy pro odhady parametrů modelu metodou intrinsického odhadu (*Yang – Land*, 2013). Vedle softwaru R jsou volně dostupné i další dva softwary. Software BAMP (Bayesian Age-period-cohort Modeling and Prediction) se zaměřuje na bayesovské odhady (*Schmid – Held*, 2007) a software MIAMOD/PIAMOD (Mortality and Incidence Analysis MODel/Prevalence and Incidence Analysis MODel) na zobecněné aditivní modely (*De Angelis et al.*, 1994; *Verdecchia et al.*, 2002). Z komerčních programů, do nichž byly speciální metody modelování APC implementovány, jmenujme například program STATA (*Rutherford et al.*, 2012; *Sasieni*, 2012) nebo SAS (*Yang – Land*, 2006). Výstupy prezentované v tomto článku byly provedeny v softwaru SAS verze 9.4, R verze 2.15.2 a Microsoft Office Excel.

7. ZÁVĚR

Modely typu věk-období-kohorta se zjišťuje, do jaké míry je vývoj zkoumaných populačních ukazatelů (např. úmrtnosti, rozvodovosti) ovlivněn věkem osob, obdobím, ve kterém žijí a generací (kohortou určenou rokem narození), ke které patří. Tento přístup se používá především v případech, kdy nemáme k dispozici data o potencionálních rizikových či protektivních faktorech (např. o kouření, stresu) zkoumaných populačních ukazatelů. Na příkladech jsme si ukázali, při jakých analýzách se modely typu věk-období-kohorta využívají a jaká je jejich interpretace. Upozornili jsme na metody, které se používají k odhadu parametrů modelů a jaké jsou výhody a nevýhody těchto metod.

Modelování typu věk-období-kohorta má význam nejen při analýze historických trendů, ale využívá se i predikcích budoucího vývoje. Ty mají na jedné straně význam administrativní (např. za účelem plánování léčebných výdajů) a na druhé straně vědecký. I když predikování budoucího vývoje nebylo cílem tohoto článku, nelze ho opomenout. Zásadním předpokladem správné predikce je totiž důsledná analýza historických dat, na kterou jsme se v uvedeném článku právě zaměřili.

Literatura

- Alves, S. M. – Castiglione, D. – Oliveira, C. M. – de Sousa, B. – Pina, M. F. 2014. Age-period-cohort effects in the incidence of hip fractures: political and economic events are coincident with changes in risk. *Osteoporos International*, 25 (2), s. 711–720.
- Arbyn, M. – Van Oyen, H. – Sartor, F. – Tibaldi, F. – Molenberghs, G. 2002. Description of the influence of age, period and cohort effects on cervical cancer mortality by loglinear Poisson models (Belgium, 1955–94). *Archives of Public Health*, 60, s. 73–100.
- Bashir, S. A. – Estève, J. 2001. Projecting cancer incidence and mortality using Bayesian age-period-cohort models. *Journal of Epidemiology and Biostatistics*, 6(3), s. 287–296.
- Bell, A. – Jones, K. 2014. Another 'futile quest'? A simulation study of Yang and Land's Hierarchical Age-Period-Cohort model. *Demographic Research*, 30, s. 333–360.
- Bencko, V. – Hrach, K. – Malý, M. – Pikhart, H. – Reissigová, J. – Svačina, Š. – Tomečková, M. – Zvárová J. 2003. *Statistické metody v epidemiologii*. Biomedicínská statistika III. Svazek 1, 2. Praha: Karolinum.
- Berzuini, C. – Clayton, D. – Bernardinelli, L.1993. Bayesian inference on the Lexis diagram. *Bulletin of the International Statistical Institute*, 50, s. 149–164.
- Berzuini, C. – Clayton, D. 1994. Bayesian analysis of survival on multiple time scales. *Statistics in Medicine*, 13(8), s. 823–838.
- Bray, I. 2002. Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51, s. 151–164.
- Breen, R. – Jonsson, J. O. 2007. Explaining Change in Social Fluidity: Educational Equalization and Educational Expansion in Twentieth-Century Sweden. *American Journal of Sociology*, 112 (6), s. 201–239.
- Burcin, B. – Kučera, T. 2008. Strukturální změny úmrtnosti v Českých zemích a na Slovensku mezi roky 1991 a 2006. *Demografie*, 3, s.173–185. Dostupné z: <http://www.czso.cz/csu/2012edicniplan.nsf/t/100032D43A/$File/demografie_3_2008.pdf>.
- Carstensen, B. 2007. Age-Period-Cohort models for the Lexis diagram. *Statistics in Medicine*, 26, s. 3018–3045.
- Caselli, G. – Cappocaccia, R. 1989. Age, period, cohort and early mortality: an analysis of adult mortality in Italy. *Population Studies*, 43(1), s. 133–153.
- Clayton. D. – Schifflers E. 1987. Models for temporal variation in cancer rates. II: Age-period-cohort models. *Statistics in Medicine*, 6, s. 449–481.
- Clark, A. K. – Eisenstein, M. A. 2013. Interpersonal trust: An age-period-cohort analysis revisited. *Social Science Research*, 42, s. 361–375.
- De Angelis, G. – De Angelis, R. – Frova, L. – Verdecchia, A. 1994. Miamod: a Computer Package to Estimate Chronic Disease Morbidity Using Mortality and Survival Data. *Computer methods and programs in biomedicine*, 44(2), s. 99–107.
- Elias, V. L. – Fullerton, A. S. – Simpson, J. M. 2013. Long-Term Changes in Attitudes Toward Premarital Sex in the United States: Reexamining the Role of Cohort Replacement. *Journal of sex research*.
- Fienberg, S. E. – Mason, W. M. 1985. Specification and Implementation of Age, Period, and Cohort Models. In Mason, W. M. – Fienberg, S. E. (eds.): *Cohort Analysis in Social Research*. New York: Springer-Verlag.
- Finnish Cancer Registry. 2009. *Cancer in Finland 2006 and 2007*. Helsinki: Cancer Society of Finland Publication No. 76.
- Fu, W. J. 2000. Ridge Estimator in Singular Design with Application to Age-Period-Cohort Analysis of Disease Rates. *Communications in Statistics—Theory and Method*, 29, s. 263–78.
- Fu, W. J. – Land, K. C. – Yang, Y. 2011. On the Intrinsic Estimator and Constrained Estimators in Age-Period-Cohort Models. *Sociological Methods* & *Research*, 40 (3), s. 453–466.
- Gelnarová, E. – Neuvirtová, L. – Svobodník, A. – Komolíková, L. – Daneš, J. – Kovajsová, M. – Bartoňková, H. – Mužík, J. – Koptíková, J. – Dušek, L. 2007. Využití Národního onkologického registru pro modelování vlivu screeningových programů v cílové populaci: age-period-cohort modely. *Klinická onkologie*, 20, Sup.1/2007, s. 167–175.
- Glenn, N. D. 2003. *Distinguishing age, period, and cohort effects*. In Mortimer, J. T. – Shanahan, M. J. (eds.): *Life Couse*. New York: Kluwer Academic.
- Heuer, C. 1997. Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrisc*, 53, s. 161–177.
- Hobcraft, J. – Menken, J. – Preston, S. 1982. Age, period, and cohort effects in demography: a review. *Population Index*, 48 (1), s. 4–43.
- Holford, T. R. 1983. The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39, s. 311–324.
- Holford, T. R. 1991. Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annual Review Public Health*, 12, s.425–457.

- Jiang, B. – Carriere K. C. 2013. Age-period-cohort models using smoothing splines: a generalized additive model approach. *Statistics in Medicine*, 33, s. 595–606.
- Kahn, J. R. – Mason, W. M. 1987. Political alienation, cohort size, and the Easterlin hypothesis. *American Sociological Review*, 52(2), s. 155–69.
- Katrňák, T. 2009. Kohortní analýza jako alternativa panelového výzkumu. *Data a výzkum – SDA Info*, 3 (1), s. 53–74.
- Keyes, K. M. – Li, G. 2010. A multi-phase method for estimating cohort effects in age-period contingency table data. *Annals of Epidemiology*, 20(10), s. 779–785.
- Keyes, K. M. – Miech, R. 2013. Age, period, and cohort effects in heavy episodic drinking in the U.S. from 1985–2009. *Drug and Alcohol Dependence*, 132, s. 140–148.
- Keyes, K. M. – Nicholson, R. – Kinley, J. – Raposo, S. – Stein, M. B. – Goldner, E. M. – Sareen, J. 2014. Age, Period, and Cohort Effects in Psychological Distress in the United States and Canada. *American journal of epidemiology*, 179(10), s. 1216–1227.
- Kupper, L. L. – Janis, J. M. – Karmous, A. – Greenberg, B.G. 1985. Statistical age-period-cohort analysis: a review and critique. *Journal of Chronic Diseases*, 38(10), s. 811–830.
- Luo, L. 2013. Assessing validity and application scope of the intrinsic estimator approach to the age-period-cohort problem. *Demograhpy*, 50, s. 1945–1967.
- Mason, K. O. – Mason, W. M. – Winsborough, H. H. – Poole, W. K. 1973. Some Methodological Issues in Cohort Analysis of Archival Data. *American Sociological Review*, 38(2), s. 242–258.
- Møller, B. – Fekjaer, H. – Hakulinen, T. – Tryggvadottir, L. – Storm, H. H. – Talback, M. – Haldorsen, T. 2002. Prediction of cancer incidence in the Nordic countries up to the year 2020. *European Journal of Cancer Prevention*, 11(Suppl 1), S1–S96.
- Nakamura, T. 1986. Bayesian Cohort Models for General Cohort Table Analysis. *Annals of the Institute of Statistical Mathematics*, 38, s. 353–370.
- O'Brien, R. M. 2000. Age period cohort characteristic models, *Social Science Research*, 29, s. 123–139.
- O'Brien, R. M. 2011. Intrinsic Estimators as Constrained Estimators in Age-Period-Cohort Accounting Models. *Sociological Methods & Research*, 40 (3), s. 467–470.
- O'Brien R. M. 2014. *Age-Period-Cohort Models: Approaches and Analyses with Aggregate Data.* Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences.
- Pavalko, E. K. – Gong, F. – Long J. S. 2007. Women's Work, Cohort Change, and Health. *Journal of Health and Social Behavior,* 48(4), s. 352–368.
- Pekár, S. – Brabec, M. 2009. *Moderní analýza biologických dat 1, Zobecněné lineární modely v prostředí R.* Praha: Scientia.
- R Development Core Team. 2012. R: *A language and environment for statistical computing, reference index version 2.15.2.* R Foundation for Statistical Computing. Vienna, Austria.
- Reissigová, J. – Tomečková, M. 2008. Ischemická choroba srdeční u mužů v České republice, 1980 2004. *European Journal for Biomedical Informatics*, 4, s. 12–16. Dostupné z: <http://www.ejbi.org/en/ejbi/article/85-cs-ischemicka-choroba-srdecni-u-muzu-v-ceske-republice-1980-2004.html>.
- Rutherford, M. J. – Thompson, J. R. – Lambert, P. C. 2012. Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines. *The international journal of biostatistics*, 8(1), s. 33.
- Rychtaříková, J. – Řehák, J. – Caselli, G. – Meslé, F. – Vallin, J. 1994. *Analysis of mortality in the Czech Republic: Causal models of mortality changes in generations and the international comparative analysis.* The Central European University, Final Report on Grant no 879, Category G.
- Rychtaříková, J. 2004. The case of the Czech Republic. Determinants of the Recent Favourable Turnover in Mortality. *Demographic Research, Special Collection 2, Determinants of Diverging Trends in Mortality*, S2–5, s. 105–137. Dostupné z: <http://demographic-research.org/special/2/5/default.htm>.
- Rychtaříková, J. 2005. Education and survival in the Czech Republic. *Acta Universitatis Carolinae Geographica*, 1–2, s. 123–137. Dostupné z: <https://web.natur.cuni.cz/ksgrrsek/acta/2005/AUC_2005_40_1-2_rychtarikova_vzdelani_a_delka.pdf>.
- Řeháková, B. 2008. Kontrasty v logistické regresi. *Sociologický časopis/Czech Sociological Review*, 44 (4), s. 745–765.
- Sasieni, P. D. 2012. Age-period-cohort models in Stata. *Stata Journal,* 12(1), s. 45–60.
- Schmid, V. J. – Held, L. 2007. Bayesian age-period-cohort modeling and prediction – BAMP. *Journal of Statistical Software*, 21 (8).

- Schwadel P. 2011. Age, Period, and Cohort Effects on Religious Activities and Beliefs. *Social Science Research*, 40, s. 181–192.
- Smith, H. L. 2008. Advances in age-period-cohort analysis: Introduction. *Special issue: Age–period–cohort models revisited. Sociological Methods & Research*, 36(3), s. 287–296.
- Verdecchia, A. – De Angelis, G. – Capocaccia, R. 2002. Estimation and Projections of Cancer Prevalence From Cancer Registry Data. *Statistics in Medicine*, 21, s. 3511–26.
- Wilmoth, J. R. 2006. Age-Period-Cohort Models in Demography. *Demography: Analysis and Synthesis*, vol. 1, s. 227–236.
- Winship, C. – Harding, D. J. 2008. A mechanism-based approach to the identification of age-period-cohort models. *Sociological Methods & Research*, 36 (3), s. 362–401.
- Yang, Y. – Fu, W. J. – Land, K. C. 2004. A Methodological Comparison of Age-Period-Cohort Models: Intrinsic Estimator and Conventional Generalized Linear Models. *Sociological Methodology*, 34, s. 75–110.
- Yang, Y. – Fu, W. J. – Land, K. C. 2006. A Mixed Models Approach to Age-Period-Cohort Analysis of Repeated Cross-Section Surveys: Trends in Verbal Test Scores. *Sociological Methodology*, 36, s. 75–97.
- Yang, Y. – Land, K. C. 2006. A mixed models approach to the age-period-cohort analysis of repeated cross-section surveys, with an application to data on trends in verbal test scores. *Sociological Methodology,* 36(1), s. 75–97.
- Yang, Y. – Schulhofer-Wohl, S. – Fu, W. J. – Land, K. C. 2008. The Intrinsic Estimator for Age-Period-Cohort Analysis: What It Is and How to Use It. American Journal of Sociology, 113(6), s. 1697–1736.
- Yang, Y. – Land, K. C. 2013. *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Chapman & Hall/CRC Interdisciplinary Statistics.

Poděkování:

Článek vznikl s podporou na dlouhodobý koncepční rozvoj výzkumné organizace RVO:67985807 a s finanční podporou grantu Grantové agentury ČR pro projekt č. P404-12-0883.

JINDRA REISSIGOVÁ

vystudovala Matematicko-fyzikální fakultu Karlovy univerzity v Praze, obor matematická statistika a teorie pravděpodobnosti. V současné době pracuje jako biostatistička v Ústavu informatiky AV ČR. Podílí se na statistickém vyhodnocování epidemiologických studií, popisu statistických analýz a interpretace výsledků. Je spoluautorkou monografie *Statistické metody v epidemiologii* (Karolinum 2003).

JITKA RYCHTAŘÍKOVÁ

je profesorkou demografie na katedře demografie a geodemografie Přírodovědecké fakulty Univerzity Karlovy v Praze a předsedkyní České demografické společnosti. Věnuje se zejména demografickým analýzám populačního vývoje České republiky se zaměřením na současné změny a v mezinárodním pohledu. Je autorkou a spoluautorkou řady odborných publikací u nás i v zahraničí, z nichž mezi poslední patří: Les défis actuels de la démographie tchèque (*Revue d'Études Comparatives Est-Ouest, 2009*), Population Aging: A Common Challenge for Europe (*Geographische Rundschau, 2010*), Impact of parental ages and other characteristics at childbearing on congenital anomalies: Results for the Czech Republic, 2000–2007 (*Demographic Research 2013*).

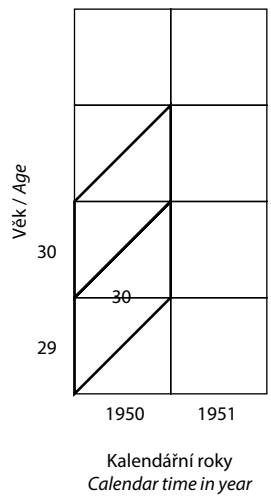
SUMMARY

The aim of the article is to examine the age-period-cohort models that have been used to evaluate the trends of various demographic, sociological and epidemiological indicators (e.g. mortality, fertility) since the 1970s

and continue to evolve. The authors explain an identification problem (the linear dependency between the age, period and cohort) and discuss possible solutions. The main age-period-cohort modelling approaches are summarised: the constrained generalised linear models, general additive models, age-period-cohort characteristic models, intrinsic estimation, the hierarchical generalised linear model and Bayes estimates. The advantages and disadvantages of individual methods are described and their use is illustrated on the basis of examples of Czech male mortality. The age-period-cohort modelling guidelines and freely available software are described. The article could help scientists better understand how such models work and the interpretation of their results.

Příloha / *Annex*

Lexisův diagram | Lexis diagram



Období / *Period*	Generace *Birth cohort*	Dosažený věk *Current age*	Zemřelí / *Deaths*	Počet mužů k 1.1. *Number of men as of 1 January*	q_x
1950	1920	30	180	68 461	0,002629
1950	1919	31	118	54 607	0,002161
1950	1918	32	95	31 891	0,002979
1950	1917	33	89	32 729	0,002719
1950	1916	34	126	36 542	0,003448
1950	1915	35	155	49 707	0,003118
1950	1914	36	211	67 617	0,003121
1950	1913	37	222	71 257	0,003115
1950	1912	38	215	72 318	0,002973

Zemřelí jsou ve druhém hlavním souboru událostí. Počet mužů k 1.1. daného roku začíná dokončeným věkem 29 let, avšak pro označení věku ukazatele (pravděpodobnost úmrtí) byl použit dosažený věk, tj. 30 let.
30 (dosažený věk) = 1950 (období) – 1920 (rok narození)
Pravděpodobnost úmrtí q_x byla počítána jako 180/68461 = 0,002629.

Pravděpodobnosti úmrtí v transversálním pohledu byly uspořádány pro graf 1:
Probability of death is organised in a cross-sectional perspective in Figure 1:
Dosažený věk / *Current age:* 30, 31,.....80 let / *years*
Období v jednotlivých letech / *Period*: 1950, 1951,....2013

Dosažený věk ***Current age***	**Kalendářní roky / *Calendar time in years***						
	1950	1951	1952	1953	1954	1955	1956
30	0,002629	0,002493	0,002268	0,002198	0,001956	0,001889	0,001762
31	0,002161	0,002453	0,001935	0,001565	0,001765	0,001654	0,001704
32	0,002979	0,002368	0,002251	0,002107	0,001985	0,001976	0,002013
33	0,002719	0,002713	0,002060	0,002206	0,002015	0,002168	0,001831
34	0,003448	0,003095	0,002589	0,002404	0,002368	0,002156	0,001943
35	0,003118	0,003091	0,002854	0,002272	0,002352	0,002319	0,002175
36	0,003121	0,002924	0,002984	0,002578	0,002495	0,001958	0,002236
37	0,003115	0,002844	0,003029	0,002762	0,002427	0,002154	0,002220

Pravděpodobnosti úmrtí v longitudinálním pohledu byly uspořádány pro graf 2:
Probability of death is organised in a cohort perspective current age in Figure 2:
Dosažený věk / *Current age*: 30, 31,.....80 let / *years*
Jednotlivé generace / *Birth cohort*: 1870, 1871,....1983

Dosažený věk ***Current age***	**Generace / *Birth cohort***						
	1977	1978	1979	1980	1981	1982	1983
30	0,001033	0,000932	0,000933	0,000742	0,000789	0,000669	0,000774
31	0,000898	0,000929	0,000809	0,000932	0,000961	0,000746	.
32	0,000812	0,000980	0,000951	0,000950	0,001064	.	.
33	0,000936	0,000937	0,001077	0,000840	.	.	.
34	0,000957	0,001088	0,000998	.	.	.	.
35	0,001062	0,001293	.	.	.	.	.
36	0,001041	.	.	.	.	.	.