MATEMATICKÉ METODY V MODELU ANALÝZY PŘEŽITÍ

Jana Kubanová

Katedra matematiky, FES, Univerzita Pardubice

Mathematical models are used in many scientific researches not only in natural sciences but in social sciences. One of the basic demographic models is shown in this contribution – the mortality model. The reproduction of population is formed by two basic events – nativity and mortality. On base of survival function, the mortality intensity can be expressed as the quotient of probability density function and survival function. To estimate the mortality intensity for senior age (more 60 years) is difficult, the shape of trend function is following: $I(t) = e^{kt} \cdot c + z$. *The second problem, that was solved, is how to calculate the number of people living to certain age, when the mortality intensity is known. By the solution of differential equation we obtain the recipe* $S(x) = S(0) \cdot e^{-\int_0^x I(t)dt}$. *The only problem is the shape of the function I(t). It could be difficult for senior age people. Solution for mortality intensity (I(t)) this category is shown in the contribution.*

Poznatky z matematiky, pravděpodobnosti a statistiky nacházejí široké uplatnění v mnoha vědních oborech a zejména v technických a ekonomických disciplínách jsou výzkum a zpracování poznatků bez matematického aparátu jen obtížně představitelné.

Velmi názorné je využití matematických metod při výzkumu reprodukce lidské populace. Procesy, vztahy a jevy související s reprodukcí populace je možné vyjádřit pomocí demografických modelů. Tyto modely obsahují matematické formule vyjadřující závislosti jevu (např. plodnosti, sňatečnosti, úmrtnosti) na čase. Základním problémem je pak výpočet různých nepodmíněných i podmíněných pravděpodobností pro dobu trvání určitého jevu.

Jedním ze základních a zřejmě také z hlediska analýzy nejtěžších demografických pojmů je intenzita úmrtnosti. Intenzita úmrtnosti je klíčovým pojmem ve stochasticky pojaté analýze přežití a je analogií vyjádření pravděpodobnosti úmrtí pomocí spojité náhodné veličiny. Tato veličina vyjadřuje pravděpodobnost úmrtí (p_x) osoby ve věku x-let před dovršením x+1 roku. Délku dalšího života osoby ve věku x let definujeme jako spojitou náhodnou veličinu, kterou označíme Y. Distribuční funkci této náhodné veličiny označíme F_Y a platí: $F_Y(t) = P(Y < t)$

Tato distribuční funkce vyjadřuje pravděpodobnost, že délka dalšího života Y je kratší než t zvolených časových období.

Pro sledování délky trvání jevu je významný doplněk této distribuční funkce F_Y, který se nazývá funkce přežití a značí se S. Pak platí:

$$S(t) = 1 - F_Y(t) = 1 - P(Y < t) = P(Y \geq t).$$

Funkce přežití vyjadřuje pravděpodobnost, že trvání dalšího života Y je t časových jednotek nebo delší. Odhadem funkce přežití z empirických dat je podíl počtu osob, u kterých trvání života překročilo t jednotek času k celkovému počtu žijících.

Důležitou funkcí v tomto modelu analýzy přežití je intenzita úmrtnosti. Intenzita úmrtnosti odpovídá funkci, která má v teorii spolehlivosti název riziková funkce. Intenzitu úmrtnosti můžeme vyjádřit jako podíl hustoty pravděpodobnosti náhodné veličiny Y a funkce přežití.

$$I(t) = \frac{f_Y(t)}{S(t)} = \frac{dF_y(t)}{dt\,S(t)} = \frac{d(1 - S(t))}{dt\,S(t)} = -\frac{S'(t)}{S(t)}$$

Intenzita úmrtnosti je podle svého vyjádřené vlastně pravděpodobnost ukončení života v čase t za podmínky, že doba života byla až do času t. Tato funkce tedy vyjadřuje pravděpodobnost úmrtí v závislosti na době trvání života. V případě této funkce rizika můžeme mluvit o pozitivní závislosti, protože pravděpodobnost ukončení života během doby trvání roste.

Odhadem intenzity úmrtnosti z reálných údajů o délce života je podíl počtu osob (m), které ukončili život v k-tém období a doby expozice, to je doby, po kterou jsou členové sledované populace vystaveni riziku úmrtí (n·Δt). Potom

$$\hat{I} = \frac{m}{n \cdot \Delta t}$$

Protože zjistit skutečnou dobu expozice je velmi obtížné, je možné při splnění určitých předpokladů použít její odhad, tzv. střední stav. Odhad intenzity úmrtnosti se nazývá obecná míra úmrtnosti a časový interval se obvykle uvažuje jeden rok.

Pokud se pokusíme vyjádřit vztah intenzity úmrtnosti a věku, zjistíme, že přibližně do 60 let věku můžeme hodnotami intenzity proložit jednoduchou křivku polynomiálního charakteru. K vyrovnání této křivky je možné použít metodu klouzavých průměrů nebo specifických metod, vyvinutých právě pro tento účel.
Pro odhad intenzity úmrtnosti ve vyšším věku však výše popsaný způsob nevyhovuje, proto byly odvozeny jiné trendové funkce.
Benjamin Gompertz navrhl již v roce 1885 funkci, kterou nazval odolností vůči destrukci. Její hodnoty odpovídají převráceným hodnotám intenzity úmrtnosti. Pak platí:

$$\frac{d\left(\frac{1}{I(t)}\right)}{dt} = -k \cdot \frac{1}{I(t)} \qquad k\ldots\text{konst., } k > 0$$

$$\int \frac{d\left(\frac{1}{I(t)}\right)}{\frac{1}{I(t)}} = -k \cdot \int dt$$

$$\ln\left|\frac{1}{I(t)}\right| = -kt + c_1$$

$$\frac{1}{I(t)} = e^{-kt} \cdot e^{c_1}$$

$$I(t) = e^{kt} \cdot c \qquad 0 < c < 1$$

Tato trendová funkce byla později zdokonalena W.M. Makehamem, který vyjádřil pomocí aditivně přidaného parametru z náhodná úmrtí, nesouvisející s věkem.
Funkce I(t) má pak tvar:

$$I(t) = e^{kt} \cdot c + z \qquad (1)$$

Dlouhodobá pozorování ukazují, že funkce (1) slouží k vyrovnání časové řady intenzit úmrtnosti pro vyšší věky (nad 60 let) velmi dobře a že extrapolace získané pomocí této funkce odpovídají skutečnosti.
Funkce $I(t) = e^{kt} \cdot c + z$ není z hlediska parametrů lineární, patří do skupiny funkcí s modifikovaným exponenciálním trendem. K odhadu jejích parametrů není proto možné použít metodu nejmenších čtverců, odhady se musí provádět některou z metod nelineární regrese, například metodou částečných součtů, metodou dílčích průměrů nebo metodou vybraných bodů. Popis těchto metod je možné nalézt v některých učebnicích matematické statistiky.

Z empirických údajů je zpravidla poměrně snadné odhadnout intenzitu úmrtnosti pro danou populaci. Nyní budeme řešit opačný úkol, jak z daných intenzit úmrtnosti vypočítat počty osob dožívajících se daného věku.
Budeme řešit jednoduchou diferenciální rovnici:

$$I(t) = -\frac{dS(t)}{S(t)dt} \qquad t \in \langle 0, \tau)$$

Integrací a separací proměnných dostaneme:

$$\int \frac{dS(t)}{S(t)} = -\int I(t)dt$$

$$\ln|S(t)| = -\int I(t)dt + c_1$$

$$S(t) = e^{-\int I(t)dt} \cdot c$$

Nahradíme-li při integraci neurčitý integrál určitým pro $t \in (0, x)$, kde proměnná x vyjadřuje věk, dostáváme:

$$\left[\ln|S(t)|\right]_0^x = -\int_0^x I(t)dt$$

Protože funkce S(t) nabývá jen nezáporných hodnot, můžeme absolutní hodnotu vynechat.

$$\ln S(x) - \ln S(0) = -\int_0^x I(t)dt$$

$$S(x) = S(0) \cdot e^{-\int_0^x I(t)dt} \qquad (2)$$

Chceme-li odhadnout počet osob, dožívajících se věku x (x > 60), dosadíme do vztahu (2) za I(t) Gompertz-Makehamovu funkci (1) a dostaneme:

$$S(x) = S(60) \cdot e^{-\int_{60}^x (e^{kt} \cdot c + z)dt} = S(60) \cdot e^{-\frac{c}{k}(e^{kx} - e^{60k}) + z(60-x)} = S(60) \cdot e^{-\frac{c}{k}e^{kx}} \cdot e^{\frac{c}{k}e^{60k}} \cdot e^{60z} \cdot e^{-xz}$$

Položíme: $K = S(60) \cdot e^{\frac{c}{k}e^{60k}} \cdot e^{60z}$

$$A = e^{-z}$$

$$B = e^{-\frac{c}{k}}$$

$$C = e^{k}$$

Výsledná funkce má tvar: $S(x) = K \cdot A^x \cdot B^{C^x}$. Na základě této trendové funkce jsme schopni odhadnou počet osob, dožívajících se věku x let.

Pomocí funkce I(t) můžeme vyčíslit veškeré charakteristiky i veličiny, které se objevují v úmrtnostních tabulkách.

Pravděpodobnost úmrtí x-leté osoby před dovršením věku x + n let můžeme vypočítat podle vztahu:

$$p_x = 1 - e^{-\int_x^{x+n} I(t)dt}$$

Střední hodnota náhodné veličiny Y se nazývá střední délka života a pro výpočet platí vztah:

$$E(Y) = \int_0^\infty t \cdot \left(1 - e^{-\int_x^{x+n} I(t)dt}\right) dt$$

Podobně disperzi D náhodné veličiny Y vypočítáme podle vztahu:

$$D(Y) = \int_0^\infty (t - E(Y))^2 \cdot \left(1 - e^{-\int_x^{x+n} I(t)dt}\right) dt$$

Pro malé soubory a cenzurovaná pozorování se jeví jako vhodnější Kaplan-Meierův model přežití. Kaplan-Meierovy odhady rizikové funkce se počítají pro každý časový interval mezi dvěma výskyty sledovaného jevu (v našem případě úmrtí osoby).

Pokud bychom předpokládali, že pravděpodobnost úmrtí je stejná pro všechny prvky daného souboru (čili pro všechny osoby – což zpravidla obecně není), situace se podstatně zjednoduší. Počet úmrtí bude náhodnou veličinou X s binomickým rozdělením pravděpodobností s parametry p a n. Disperze takové náhodné veličiny je n·p·q. Pokud platí, že n·p·q >9, můžeme tuto náhodnou veličinu aproximovat veličinou s normálním rozdělením pravděpodobností s parametry n·p a $\sqrt{n \cdot p \cdot q}$. Potom je možné na základě náhodného výběru vypočítat intervaly spolehlivosti pro střední hodnotu náhodné veličiny X i náhodné veličiny X/n. S využitím standardních poznatků z matematické statistiky je dále možné testovat významnost rozdílů daných demografických ukazatelů.

Analýza přežití je součástí velkých statistických balíků, například SPSS má samostatný modul, věnovaný této problematice, řadu funkcí lze nalézt v S-PLUS, v nových verzích STATHRAPHICu , v Unistatu i v SASu. Široký výběr klasických i moderních metod poskytuje statistický systém NCSS 2000 dříve známý v DOSovské verzi pod názvem SOLO). Součástí je i Kaplan-Meierův neparametrický odhad funkce přežití a rizikové funkce. Modul Distribution Fitting vyrovnává a odhaduje parametry sedmi nejčastěji používaných rozdělení pravděpodobnosti

Literatura:

[1] Benjamin,B.,Pollard,J.H.: The Analysis of Mortality and Other Acturial Statistics. Butterworth-Heinemamm, Oxford 1980, 1993

[2] Browers,Jr,NL a kol.: Acturial Mathematics. The Soc. of Actuaries, Itasca, 1986

[3] Chajdiak,J.,Rublíková,E.,Gudába,M.: Štatistické metódy v praxi, Statis, Bratislava 1998

[4] Koschin,F.:Vybrané demografické modely. VŠE Praha 1995.

[5] Linda,B.,Kubanová,J.: Porovnání vybraných metod odhadu parametrů Gompertzovy křivky. In: *Sborník VI. mezinárodní konference" Kvantitatívne metódy v ekonómii a podnikání"* FHI EU Bratislava 1999.

[6] Linda,B.: Některé aspekty zdravotního stavu populace v České republice. In*: Sborník 7. demografické konference,* Trenčianske Teplice 1999, s.101-102

[7] Marek, l.: Porovnání procedur pro časové řady v softwareových produktech, Výpočtová štatistika, Bratislava 1998

[8] Pacáková,V.:Špecifiká štatistickej analýzy trvania nezamestnanosti. In: *Ekonomické rozhľady 2, 1997.* EU Bratislava 1997.

[9] Pacáková,V.: Štatistika pre poistnů prax. Statis, Bratislava 1999

[10] Pastor,K.: Aktuárska demografia jako študijný předmět. In.:*Sborník 7. demografické konference,* Trenčianske Teplice 1999

[11] Řezanková,H.: Metody pro získávání znalostí z dat, Štatistické metódy v praxi, Bratislava 1998

Kontaktní adresa:

PaedDr. Jana Kubanová, CSc.
KM FES Univerzita Pardubice
Studentská 84, 532 10 Pardubice
☎ 040 6036020
e-mail: Jana.Kubanova@upce.cz

Recenzoval: doc.RNDr.Bohdan Linda,CSc., Katedra matematiky, FES, UPa