MODELOVANIE ROZDELENÍ VYUŽITÍM KNIŽNICE *FITUR* V JAZYKU R

Michal Páleš, Silvia Komara

Abstrakt

Je známe, že jazyk R ponúka viacero postupov a metód pre odhad parametrov rozdelení a takisto testov dobrej zhody na overenie vhodného typu rozdelenia pre analyzované údaje. V tomto príspevku predstavíme pomerne novú a menej známu interaktívnu aplikáciu, ktorú obsahuje knižnica *fitur* a ktorá poskytuje veľmi prehľadný „Shiny výstup" pre overenie vhodného typu najpoužívanejších spojitých rozdelení pomocou najznámejších testov dobrej zhody. Tieto analýzy sú doplnené aj výstupmi najpoužívanejších grafických testov. Jednoduchosť použitia tejto aplikácie ju predurčuje na použitie tak v akademickej a vedeckej sfére ako aj v praxi.

Kľúčové slová

rozdelenie pravdepodobnosti, testy dobrej zhody, jazyk R, fitur, Shiny

1 ÚVOD

Ak má aktuár k dispozícii údaje o počte a výške poistných plnení z predchádzajúcich období, môže ich použiť na vytvorenie modelu – náhodnej premennej, ktorá je vhodným nástrojom na odhad hodnôt sledovaných premenných pre budúce obdobie. Postup pri konštrukcii modelu potom pozostáva z týchto krokov:

1. Najskôr sa navrhne vhodný typ rozdelenia. Rozhodnutie o jeho voľbe zohľadňuje predchádzajúce skúsenosti analytikov, ale ako cenný podklad možno využiť najmä **grafické znázornenie** dostupných údajov pomocou histogramu.

2. **Odhadnú sa parametre rozdelenia**. K dispozícii sú viaceré metódy, pričom ich výber závisí od množstva údajov, od navrhnutého typu rozdelenia, ale aj od toho, či okrem aktuálneho výberu existujú aj ďalšie relevantné informácie, ktoré možno zahrnúť do analýzy. Ak sú takéto informácie k dispozícii, možno použiť bayesovský odhad, ktorý poskytuje kvalitnejší výsledok. V prípade, že popri aktuálnom výbere neexistuje dodatočná informácia, používame metódu maximálnej vierohodnosti, metódu momentov a metódu kvantilov.

3. Pomocou **testov dobrej zhody** (*Goodness of fit tests*) overíme vhodnosť konkrétneho vybraného rozdelenia. Medzi najpoužívanejšie testy dobrej zhody patria Pearsonov chí-kvadrát test dobrej zhody, Kolmogorovov-Smirnovov test, Shapiro-Wilkov test, Kuiperov test, Cramér von-Misesov test, Watsonov test, Anderson-Darlingov test atď.

   Testujeme nulovú hypotézu:

   H_0: **Náhodná premenná X má rozdelenie s hustotou $f(x)$ (resp. výber pochádza z navrhnutého pravdepodobnostného rozdelenia)**

oproti alternatívnej hypotéze:

$H1$: **Náhodná premenná X nemá rozdelenie s hustotou $f(x)$ (výber nepochádza z navrhnutého rozdelenia)**.

Princíp všetkých testov spočíva v porovnaní rozdelenia premennej v empirickom súbore s pravdepodobnostným rozdelením navrhnutej premennej. Používa sa pritom funkcia hustoty alebo distribučná funkcia. Niektoré testy dobrej zhody sú použiteľné pre ľubovoľné teoretické rozdelenie. Existujú totiž aj testy, ktorými sa overuje len normalita rozdelenia znaku (napríklad Shapiro – Wilkov test), pričom v aktuárskom modelovaní sa však zaoberáme najmä pravostranne zošikmenými rozdeleniami. Pre viac informácií pozri [1].

Tieto analýzy obvykle realizujeme v rôznych softvéroch (jazyk R, jazyk Python, SAS, SPSS, VOSE ModelRisk, Statgraphics a pod.). Ďalej si ukážeme niektoré atribúty využitia menej známej knižnice v jazyku R, ktorou je knižnica *fitur*. Táto knižnica **poskytuje interaktívnu aplikáciu** s využitím rozhrania *Shiny* a iných atribútov jazyka R. Pre viac informácii pozri tiež [2] – [4].

2 VSTUPNÉ ÚDAJE A ZADANIE

K dispozícii máme údaje o výške poistného plnenia pri konkrétnom poistnom produkte za predchádzajúci rok v eurách (obrázok 1).

280	355	356	200	215	241	254	131	430	350
97	386	320	399	220	114	133	365	297	120
116	392	189	210	323	330	132	265	300	135
149	213	450	210	356	100	260	271	144	354
149	410	196	188	240	100	260	130	300	410
150	168	500	211	342	100	256	284	311	132
332	450	146	160	165	110	265	289	118	420
350	175	196	180	408	250	265	335	320	400

Obr. 1, zdroj: [1]

Overíme predpoklad, či údaje môžu pochádzať napríklad z exponenciálneho alebo lognormálneho rozdelenia. Použijeme pritom grafické testy a dostupné testy dobrej zhody aplikácie z knižnice *fitur*, ktorú vyvinul Thomas Roh. Tieto výsledky porovnáme so získanými výsledkami v iných softvérových riešenia uvedených v [1].

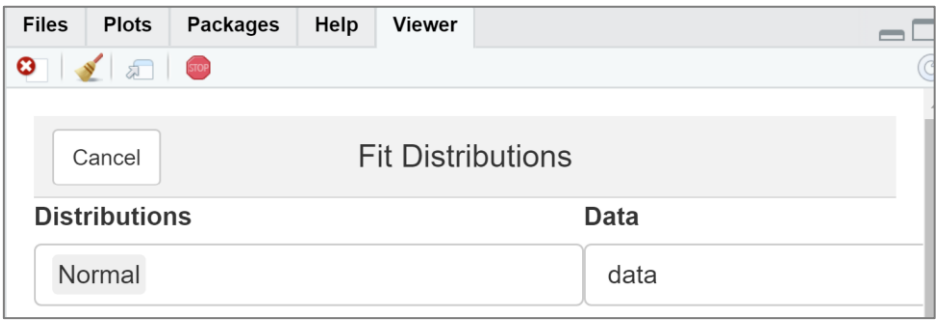
3 RIEŠENIE A ZÁVER

Pri analýze postupujeme postupne od načítania údajov až po zobrazenie výstupu pomocou nasledovného kódu jazyka R:

```
library(fitur)
library(actuar)
data<-scan()
fitur::fit_dist_addin()
```

Analýzu realizujeme v RStudiu, pričom si všímame okno IV. kvadrantu (obrázok 2), kde klikneme na ikonu *Otvoriť v novom okne* a aplikácia sa nám otvorí na celú obrazovku v predvolenom internetovom prehliadači (obrázok 3). Obratom sa zobrazia výsledky analýzy (tabuľka s výsledkami testov dobrej zhody a tri grafické výstupy). Tu môžeme interaktívne voliť:

- údaje (automaticky ponúka názvy všetkých uložených objektov s údajmi v prostredí)
- rozdelenie/a pravdepodobnosti, ktoré chceme testovať (z dostupných rozdelení, ktoré ponúka aplikácia),
- počet tried (*bins*) histogramu (prednastavená hodnota je 30),
- zoradenie výsledkov charakteristík ponúkaných troch testov dobrej zhody, resp. ich *p*-hodnôt.



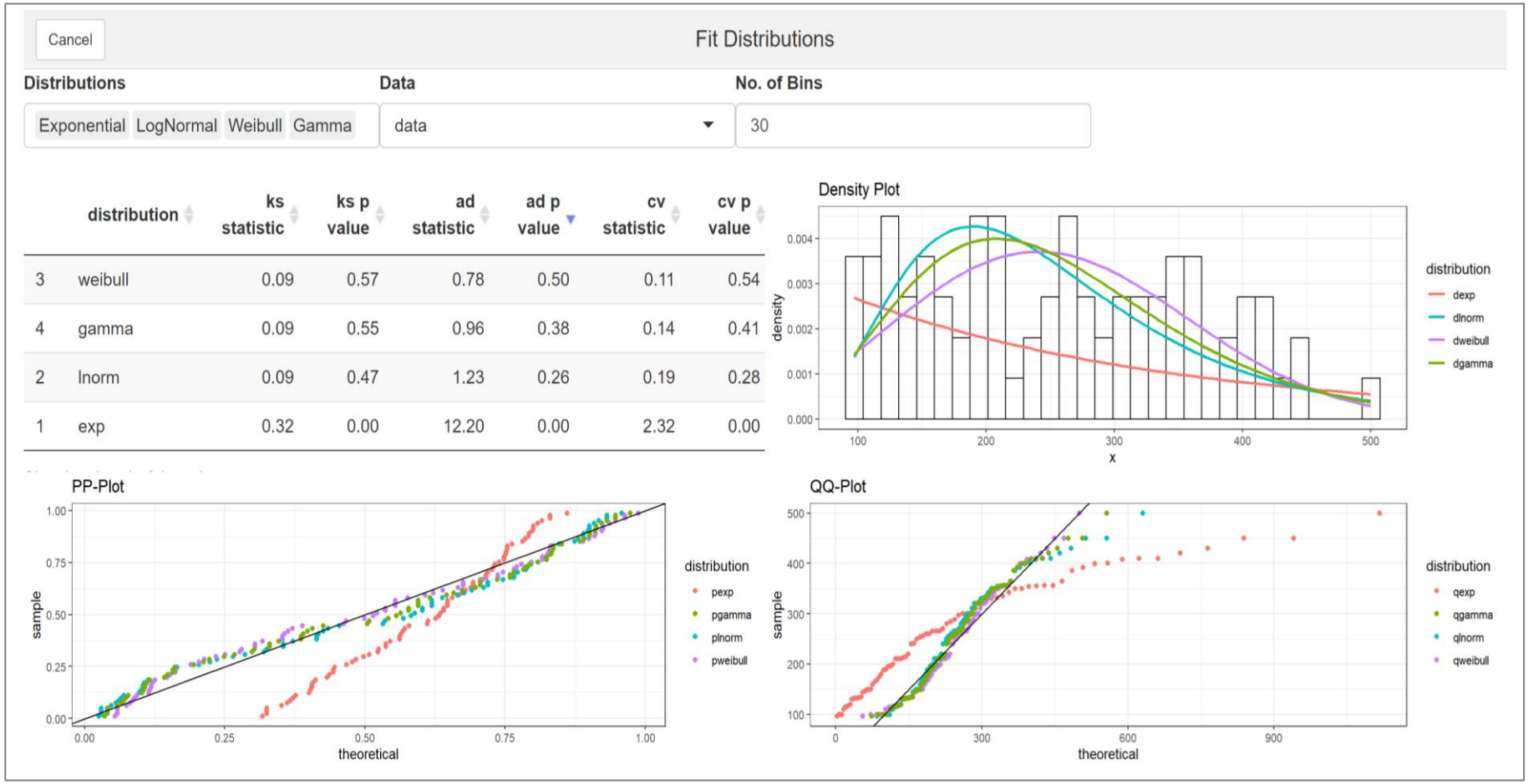
Obr. 2, zdroj: vlastný

Je potrebné upozorniť, že rozsiahle dátové súbory môžu byť náročnejšie na spracovanie výsledkov a čas zobrazenia sa môže predĺžiť. Pre ukončenie aplikácie stlačíme tlačidlo *Cancel* (obrázok 3) alebo *STOP* (obrázok 2).

Pre odhad parametrov príslušných rozdelení (pozn. ktoré v podstate boli využité v danej analýze aplikácie) môžeme rovnako využiť knižnicu *fitur* nasledovne (tu pre lognormálne rozdelenie) nasledovne:

```
fitted <- fit_univariate(data, 'lnorm', type = "continuous")
fitted$parameters # meanlog = 5.4445985, sdlog = 0.4458946
```

Porovnaním výstupov uvedenými v [1] sa môžeme presvedčiť, že v softvéroch Statgraphics Centurion a SAS Enterprise Guide sme získali rovnaké výsledky Kolmogorovho-Smirnovho testu, Cramérovho-von Misesosovho testu ako aj Andersonovho-Darlingovho testu. Na základe vyčíslených *p*-hodnôt vzťahujúcich sa na overovanie zhody empirického rozdelenia s exponenciálnym rozdelením nulovú hypotézu zamietame, a teda nemôžeme predpokladať, že výška poistných plnení sa riadi exponenciálnym rozdelením. Naopak, pre všetky tri uvedené testy vzťahujúce sa na overovanie zhody empirického rozdelenia s lognormálnym rozdelením nulovú hypotézu nezamietame. V [1] predpokladáme, že výška poistných plnení sa riadi lognormálnym rozdelením s príslušnými parametrami (uvedenými vyššie). Avšak nakoľko sme v tejto aplikácií uvažovali aj s ďalšími rozdeleniami, vidíme, že tieto môžu byť pre analyzované údaje ešte lepším modelom. Ďalšie úvahy ponechávame na čitateľovi. Aplikácia by mohla byť ešte rozšírená o Paretovo rozdelenie.



Obr. 3, zdroj: vlastný

POUŽITÉ ZDROJE

[1] KOTLEBOVÁ, E. – LUKÁČIK, M. – PÁLEŠ, M. – ŠOLTÉS, E.. Aktuárska štatistika. Bratislava: Letra Edu, 2020.

[2] PÁLEŠ, M. Jazyk R pre aktuárov. Bratislava: Letra Edu, 2019.

[3] ROH, T. fitur: Fit Univariate Distributions. R package version 0.6.2., 2021. <https://cran.r-project.org/web/packages/fitur/index.html>

[4] <https://www.youtube.com/watch?v=srsTC9SXajw>

PROJEKT

VEGA č. 1/0431/22 – *Implementácia inovatívnych prístupov modelovania rizík v procese ich riadenia v interných modeloch poisťovní v kontexte s požiadavkami direktívy Solvency II*

VEGA č. 1/0561/21 – *Vplyv krízy COVID-19 na demografiu podnikov a zamestnanosť v SR a EÚ*

KONTAKTNÉ ÚDAJE

Páleš, Michal, doc. Ing., PhD., Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave, Dolnozemská cesta 1, 852 35 Bratislava, e-mail: michal.pales@euba.sk

Komara, Silvia, Ing., PhD., Katedra štatistiky, Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave, Dolnozemská cesta 1, 852 35 Bratislava, e-mail: silvia.komara@euba.sk