



**MATEMATICKO-FYZIKÁLNÍ
FAKULTA**
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Barbora Viskupová

Pravděpodobnostní předpověď v modelech exponenciálního vyrovnávání

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Šárka Hudecová, Ph.D.

Studijní program: Matematika (N1101)

Studijní obor: MFAPM (1103T008)

Praha 2020

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V dne

Podpis autora

Chcela by som sa poďakovať svojej školiteľke RNDr. Šárke Hudecovej, Ph.D. za námet, ochotu, odborné rady a pripomienky a za čas, ktorý mi venovala pri písaní tejto práce.

Název práce: Pravděpodobnostní předpověď v modelech exponenciálního vyrovnávání

Autor: Bc. Barbora Viskupová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Šárka Hudecová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Táto diplomová práca sa zaoberá využitím štatistických stavových modelov exponenciálneho vyrovnávania pri odhadovaní podmieneného pravdepodobnostného rozdelenia budúcich hodnôt časových radov. Jeho znalosť umožňuje počítať nielen bodové, ale aj intervalové predpovede. Práca popisuje metódy exponenciálneho vyrovnávania a zasadzuje ich do kontextu stavových modelov. Venuje sa analytickým postupom a simulačným metódam, využívaným pri výpočte intervalových predpovedí, konkrétne uvažuje simulácie založené na predpoklade normálneho rozdelenia, vychádzajúce z bootstrap metódy alebo na základe odhadnutého parametrického modelu. Obsahuje ukážku ich aplikácie na simulované aj na reálne dáta a porovnanie ich výsledkov.

Klíčová slova: časové rady, exponenciálne vyrovnávanie, stavové modely, pravdepodobnostné predpovede

Title: Probability forecast in exponential smoothing models

Author: Bc. Barbora Viskupová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Šárka Hudecová, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis deals with the use of statistical state space models of exponential smoothing for estimating the conditional probability distribution of future values of time series. This knowledge allows calculation of interval predictions, not only point forecasts. Methods of exponential smoothing are described and set into the context of state space models. Analytical and simulation methods used in the calculation of interval predictions are presented, in particular simulations based on assumption of normality, bootstrap method or estimated parametric model. The methods are applied to simulated as well as real data and their results are compared.

Keywords: time series, exponential smoothing, state space models, probability forecast

Obsah

Úvod	2
1 Základné pojmy	4
2 Exponenciálne vyrovňovanie	6
2.1 Jednoduché exponenciálne vyrovňovanie	6
2.2 Dvojité exponenciálne vyrovňovanie	9
2.3 Holtova metóda	11
2.3.1 Metódy s tlmeným trendom	12
2.4 Holtova-Wintersova metóda	14
2.4.1 Holtova-Wintersova aditívna metóda	15
2.4.2 Tlmená Holtova-Wintersova metóda	15
2.5 Klasifikácia metód exponenciálneho vyrovňovania	16
3 Stavové modely	19
3.1 Všeobecné stavové modely	19
3.2 Lineárne stavové modely	20
3.2.1 Odhad parametrov	21
3.2.2 Výber modelu	23
4 Pravdepodobnostné rozdelenie predpovedí a predikčné intervaly	24
4.1 Analytické odvodenie	24
4.1.1 Stredná hodnota rozdelenia predpovede	25
4.1.2 Rozptyl rozdelenia predpovede	25
4.1.3 Predikčné intervaly	26
4.2 Simulované rozdelenie predpovedí a predikčné intervaly	28
5 Praktická časť	31
5.1 Simulované dáta	31
5.1.1 Normálne rozdelenie chybových členov	31
5.1.2 t-rozdelenie chybových členov	37
5.1.3 Chyby z posunutého exponenciálneho rozdelenia	43
5.1.4 Zhrnutie	49
5.2 Reálne dáta	50
Záver	56
Zoznam použitej literatúry	58
Zoznam obrázkov	59
Zoznam tabuliek	60

Úvod

V podnikaní a priemysle sa často stretávame s potrebou predpovedať budúce hodnoty časových radov. Napríklad pri riadení zásob podniku je požiadavka pravidelne predpovedať dopyt pre veľké množstvo produktov, aby mohli byť uspokojivo napĺňané potreby zákazníkov. Exponenciálne vyrovnávanie predstavuje triedu metód, ktoré sú schopné takéto predpovede produkovať s ohľadom na trend alebo sezónnosť dostupných minulých dát.

Metódy exponenciálneho vyrovnávania sú známe zhruba od 50-tych rokov dvadsiateho storočia a dodnes sú jedným z najpopulárnejších predpovedných nástrojov v podnikaní (Hyndman a kol., 2008). Pôvodne boli atraktívne najmä vďaka ich nízkemu nároku na množstvo skladovaných dát, čo však so súčasnými technológiami už nepredstavuje takú veľkú výhodu oproti iným metódam. Dnes možno za primárny dôvod ich častej aplikácie považovať intuitívnosť a zrozumiteľnosť rovníc, ktoré využívajú na predpovedanie budúcich hodnôt, rovnako ako jednoduchosť ich aplikácie.

Pôvodne boli metódy exponenciálneho vyrovnávania navrhnuté ad-hoc. Medzi prvými sa touto myšlienkou zaoberal Robert G. Brown okolo roku 1944, keď ju použil vo výpočtovom zariadení na sledovanie rýchlosti a uhla používanom pri streľbe na ponorky (Hyndman a kol., 2008). Ďalej svoju prácu v tejto oblasti rozšíril aj na časové rady s diskretným časom a pridal zložky pre trend a sezónnosť, čím sa zaoberá v svojej prvej knihe (Brown, 1959).

Formuláciou štatistických modelov, vychádzajúcich z predpovedných metód exponenciálneho vyrovnávania, sa ľudia začali zaoberať až v 90-tych rokoch minulého storočia. Prvý krát sa objavuje zrejme v (Ord a kol., 1997). Výhodou tohto prístupu je najmä možnosť počítať nielen bodové, ale aj intervalové predpovede, ale tiež možnosť automatizácie výberu vhodného modelu či využívania metódy maximálnej vierohodnosti pre odhad jeho parametrov. V dôsledku toho všetkého bol štatistický rámec metód exponenciálneho vyrovnávania výrazne rozvíjaný v prvom desaťročí dvadsiateho storočia (Hyndman a kol., 2008).

Cieľom tejto práce je, z dôvodu širokého uplatnenia tohto prístupu v praxi, popísať metódy exponenciálneho vyrovnávania, rovnako ako aj štatistické stavové modely, pre ktoré sú základom, prezentovať ich využitie pri odhadovaní pravdepodobnostného rozdelenia budúcich hodnôt a jednotlivé metódy aplikovať na konkrétne príklady.

Celá práca je rozdelená na teoretickú a praktickú časť. V teoretickej časti najskôr v prvej kapitole uvedieme základné pojmy, s ktorými sa môžeme stretnúť pri analýze časových radov. Druhá kapitola popisuje metódy z triedy exponenciálneho vyrovnávania a tretia zase ich zovšeobecnenie do stavových modelov. Štvrtá kapitola sa venuje odvodeniu a odhadu pravdepodobnostného rozdelenia budúcich hodnôt, podmieneného pozorovanou minulosťou.

Druhá časť práce je praktická. V piatej kapitole najskôr aplikujeme popísané metódy a modely na simulované dáta a ich výsledky pre predpovede porovnáme so skutočnými hodnotami odvodenými na základe znalosti dáta generujúceho procesu. Ďalej dané metódy aplikujeme aj na reálne dáta a ich bodové predpovede porovnáme so skutočnými pozorovanými hodnotami.

Vlastným prínosom v rámci tejto práce je detailnejšie odvodenie niektorých

vzorcov v teoretickej časti a ilustrácia použitia popísaných metód na príkladoch. Naša práca tiež prepája štatistický prístup k metódam exponenciálneho vyrovnávania s prístupom, ktorý je vyučovaný na MFF. Navyše v praktickej časti obsahuje aj vlastnú simulačnú štúdiu a aplikáciu metód na reálne dáta.

1. Základné pojmy

Časový rad vzniká pozorovaním nejakej veličiny v čase, pričom môže byť pozorovaná spojito, alebo diskrétno v pravidelných intervaloch. V tejto práci sa budeme venovať *časovým radom s diskrétnym časom*, t.j. pod pojmom *časový rad* budeme rozumieť ľubovoľnú postupnosť náhodných veličín $\{y_t\}$, usporiadaných v čase (Cipra, 2008). Ďalej vždy uvažujeme, že máme k dispozícii pozorovania pre čas $t = 1, 2, \dots, T$.

Cielom analýzy časových radov je nájdenie modelu, ktorý generuje postupnosti podobné analyzovanému radu. Takýto model nám okrem iného umožní napríklad konštruovať predpovede pre budúce obdobia.

Jedným z prístupov k analýze časových radov je ich *dekompozícia* na špecifické zložky, ktorými sú napríklad trendová (Tr), sezónna (S), cyklická (C) či reziduálna (E). V rámci dekompozície rozlišujeme aditívne a multiplikatívne modely, prípadne ich kombinácie.

Pri *aditívnej dekompozícii* predpokladáme, že časový rad je možné zapísať v tvare

$$y_t = Tr_t + S_t + C_t + E_t. \quad (1.1)$$

Multiplikatívny model má zase tvar

$$y_t = Tr_t \cdot S_t \cdot C_t \cdot E_t.$$

Môžeme sa však stretnúť aj s ich kombináciou, napríklad v tvare

$$y_t = Tr_t \cdot S_t \cdot C_t + E_t.$$

Vo všetkých prípadoch sa zvyčajne trendová, sezónna aj cyklická zložka považujú za deterministické funkcie času, zatiaľ čo reziduálna zložka je náhodná.

Multiplikatívne modely sú oproti aditívnym komplikovanejšie a ich aplikácia v praxi môže viesť k rôznym numerickým problémom (Hyndman a kol., 2008). Aj preto sa v rámci tejto práce budeme venovať len aditívnym modelom tvaru (1.1), pričom nebudeme uvažovať cyklickú zložku C_t . Všetky popísané metódy sú však po určitej modifikácii aplikovateľné aj za predpokladu, že niektoré zo zložiek sú v multiplikatívnom vzťahu k ostatným.

Ak hovoríme o *vyrovnávaní časového radu*, ide o odstránenie jeho náhodnej zložky. Následne dostávame *vyrovnané hodnoty* časového radu, označované \hat{y}_t , kde $t = 1, 2, \dots, T$. V prípade, že z časového radu eliminujeme sezónnu zložku, hovoríme o *sezónnom očistení*.

Predpovede budúcich, ešte nepozorovaných hodnôt časového radu o τ krokov dopredu, založené na pozorovaniach do času t , budeme označovať ako $\hat{y}_{t+\tau}(t)$. Ide o bodové predpovede, ktoré sú výstupom algoritmu alebo predpovednej metódy.

Exponenciálne vyrovnávanie, ktorému sa v tejto práci venujeme, je jednou z tried týchto predpovedných metód.

Z matematického hľadiska sú však zrejme zaujímavejšie štatistické modely, ktorých výstupom je stochastický, dáta generujúci proces. Ten nám umožňuje odhadovať celé pravdepodobnostné rozdelenie, podmienené pozorovanou minulosťou, pre budúce hodnoty. Bodovú predpoveď potom získame ako odhad strednej hodnoty tohto rozdelenia a navyše môžeme počítat aj predikčné intervaly.

$(1 - \alpha) \cdot 100$ -percentný *predikčný interval* je interval, ktorý pokrýva budúcu, ešte nepozorovanú hodnotu časového radu s pravdepodobnosťou $1 - \alpha$, $\alpha \in (0,1)$.

Pri špecifikácii štatistického modelu časového radu musíme určiť predpokladané rozdelenie jeho reziduálnej zložky. V literatúre sa často predpokladá, že chyby sú realizáciou gaussovského bieleho šumu (Hyndman a kol., 2008). V praktickej časti našej práce však pracujeme aj s časovými radmi, u ktorých predpokladáme t -rozdelenie chybových zložiek, alebo že chyby sú z posunutého exponenciálneho rozdelenia.

Biely šum $\{\epsilon_t, t = 1, \dots, T\}$ je postupnosť nekorelovaných náhodných veličín s nulovou strednou hodnotou a konečným kladným rozptylom σ^2 . Hovoríme, že proces $\{\epsilon_t, t = 1, \dots, T\}$ je *gaussovský biely šum*, ak sú všetky jeho konečnorozmerné rozdelenia normálne (Prášková, 2016). Potom pre ϵ_t , $t = 1, \dots, T$, nutne platí, že sú nezávislé a rovnako rozdelené.

2. Exponenciálne vyrovnávanie

Pri spracovávaní tejto kapitoly sme vychádzali najmä z (Cipra, 2008), metódy s tlmeným trendom a posledná sekcia kapitoly sú spracované s využitím (Hyndman a Athanasopoulos, 2018) a (Hyndman a kol., 2008). Vlastným prínosom v tejto kapitole je detailnejšie rozpísanie niektorých odvození a ilustrácia aplikácie popisovaných metód na konkrétnych príkladoch¹.

Exponenciálne vyrovnávanie je jedna z metód eliminácie trendu v časových radoch a patrí medzi takzvané adaptívne prístupy k trendovej zložke. Tie sú charakteristické svojou schopnosťou pracovať so systematickými zložkami, ktoré v čase menia svoj globálny charakter, a teda ich nemožno popísať pomocou matematických kriviek s pevnými parametrami. Stále však musíme predpokladať, že vyrovnávanie pomocou takýchto kriviek je možné aspoň v krátkych úsekoch časového radu. Ich parametre sú potom zvyčajne pre jednotlivé úseky radu odlišné.

V prípade jednoduchého exponenciálneho vyrovnávania dostávame vyrovnanú hodnotu \hat{y}_t tak, že pozorované hodnoty vážime do minulosti exponenciálne klesajúcimi váhami. Pre hodnoty \hat{y}_t pritom platí, že minimalizujú výrazy typu

$$(y_t - \hat{y}_t)^2 + (y_{t-1} - \hat{y}_{t-1})^2\beta + (y_{t-2} - \hat{y}_{t-2})^2\beta^2 + \dots, \quad (2.1)$$

kde $\beta \in (0,1)$ je dopredu zvolená takzvaná diskontná konštanta. Týmto spôsobom priradujeme pozorovaniam vzdialenejším do minulosti stále nižšie váhy.

Na začiatok predpokladajme, že pracujeme s časovým radom tvaru $y_t = Tr_t + E_t$, kde Tr_t reprezentuje trendovú a E_t reziduálnu zložku.

2.1 Jednoduché exponenciálne vyrovnávanie

Jednoduché exponenciálne vyrovnávanie sa používa v prípade, že trend v analyzovanom časovom rade môžeme považovať za lokálne konštantný, tj. $Tr_t = \beta_0$.

Cieľom je nájsť odhad b_0 parametra β_0 , ale keďže exponenciálne vyrovnávanie je adaptívnym prístupom k časovým radom, tento odhad bude závisieť na čase t , v ktorom ho vykonáme. Označme ho preto radšej $b_0(t)$. Tento odhad predstavuje jednak odhadnutú hodnotu trendu v čase t , no taktiež je vyrovnanou hodnotou \hat{y}_t analyzovaného časového radu. Podľa vzťahu (2.1) potom $b_0(t)$ dostávame minimalizáciou výrazu

$$\sum_{j=1}^{\infty} (y_{t-j} - \beta_0)^2 \beta^j \quad (2.2)$$

vzhľadom na β_0 , diskontná konštanta β je zvolená dopredu z intervalu $(0,1)$.

Môžeme si všimnúť, že výraz (2.2) je nekonečným súčtom, napriek tomu, že v praxi vždy pracujeme iba s konečnými počtami pozorovaní y_t . Táto formulácia nám však poslúži na jednoduchšie odvozenie vzorcov, ktoré už pracujú iba s konečným počtom pozorovaných hodnôt.

Minimalizujme teda výraz (2.2) a položíme jeho prvú deriváciu podľa β_0 rovnú nule:

$$\sum_{j=0}^{\infty} -2(y_{t-j} - \beta_0)\beta^j \stackrel{!}{=} 0.$$

¹Dáta pochádzajú zo stránky profesora Hyndmana, ale v nami použitých zdrojoch, ktorých je autorom, s nimi nepracuje.

Z toho a z konvexity minimalizovanej funkcie dostávame:

$$\begin{aligned}\sum_{j=0}^{\infty} y_{t-j} \beta^j &= b_0(t) \sum_{j=0}^{\infty} \beta^j, \\ \sum_{j=0}^{\infty} y_{t-j} \beta^j &= \frac{b_0(t)}{1-\beta}, \\ \hat{y}_t = b_0(t) &= (1-\beta) \sum_{j=0}^{\infty} \beta^j y_{t-j}.\end{aligned}$$

Tento vzorec však stále obsahuje nekonečný súčet a preto sa pre použitie v praxi prevádza na rekurentný tvar

$$\begin{aligned}\hat{y}_t &= (1-\beta)y_t + (1-\beta) \sum_{j=1}^{\infty} \beta^j y_{t-j} = (1-\beta)y_t + \beta(1-\beta) \sum_{j=0}^{\infty} \beta^j y_{t-j-1} = \\ &= (1-\beta)y_t + \beta\hat{y}_{t-1}.\end{aligned}$$

Označme $\alpha = 1 - \beta$, $\alpha \in (0,1)$, α sa nazýva vyrovnávacia konštanta. Potom dostávame tvar:

$$\hat{y}_t = \alpha y_t + (1-\alpha)\hat{y}_{t-1}. \quad (2.3)$$

Vyrovnaná hodnota \hat{y}_t je teda konvexnou kombináciou (váženým priemerom) pozorovanej hodnoty y_t a predchádzajúcej vyrovnanej hodnoty \hat{y}_{t-1} .

Zo vzťahu (2.3) môžeme priamo vidieť niektoré z výhod exponenciálneho vyrovnávania. Jednak je to jednoduchosť výpočtov, taktiež úspornosť metódy, čo sa týka množstva skladovaných dát, keďže využíva len predchádzajúcu vyrovnanú hodnotu, a v neposlednom rade možnosť riadiť adaptívnosť metódy. Zvolením väčšej hodnoty α , v literatúre sa uvádza napríklad $\alpha = 0.3$ (Cipra, 2008), metóda reaguje na zmeny v charaktere dát rýchlejšie, no znižuje sa jej schopnosť vyhladzovania. Naopak pri menšom α , napríklad $\alpha = 0.1$ (Cipra, 2008), vyrovnávacia schopnosť metódy silnie.

Obrázok 2.1 ilustruje vyrovnanie časového radu pomocou metódy jednoduchého exponenciálneho vyrovnávania pre rôzne hodnoty vyrovnávacej konštanty α . Ide o časový rad ročných úhrnov zrážok v palcoch pre Londýn v rokoch 1813 až 1912². Zobrazená je krivka vyrovnaných hodnôt pre spomínané hodnoty $\alpha = 0.1$ a $\alpha = 0.3$ a tiež pre softvérom odhadnutú hodnotu $\alpha = 0.02$. Môžeme si všimnúť, že pre hodnoty 0.02 a 0.1 je krivka naozaj hladšia, pre hodnotu 0.3 jej správanie bližšie kopíruje skutočný nárast a pokles množstva zrážok.

Pre predpovede o $\tau > 0$ krokov dopredu pomocou metódy jednoduchého exponenciálneho vyrovnávania platí:

$$\hat{y}_{t+\tau}(t) = \hat{y}_t. \quad (2.4)$$

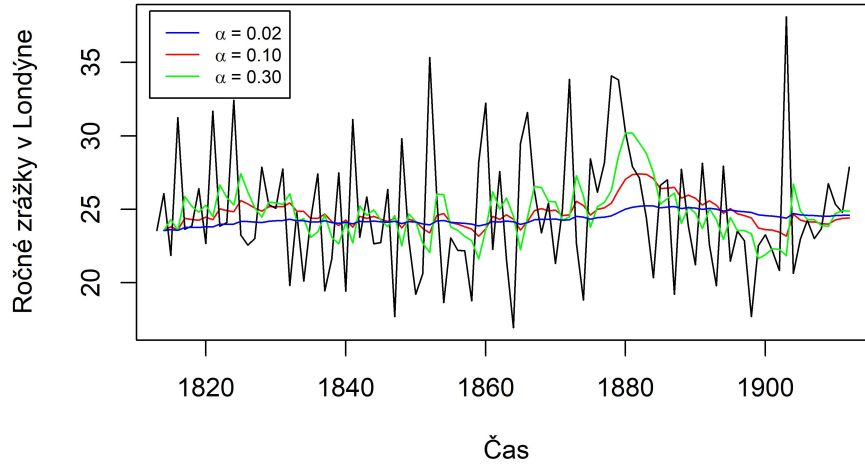
Predpovede tejto metódy sú teda konštantné a rovné poslednej vyrovnanej hodnote. Preto je prirodzené očakávať, že dlhodobejšie predpovede, t.j. hodnoty $\hat{y}_{t+\tau}(t)$ pre väčšie τ , nebudú veľmi presné.

S využitím vzťahov (2.3) a (2.4) dostávame ešte tretí tvar vzorca pre vyrovnanú hodnotu časového radu:

$$\hat{y}_t = \hat{y}_{t-1} + \alpha(y_t - \hat{y}_{t-1}) = \hat{y}_{t-1} + \alpha(y_t - \hat{y}_t(t-1)) = \hat{y}_{t-1} + \alpha e_t. \quad (2.5)$$

²zdroj: <http://www.robjhyndman.com/tsdldata/hurst/precip1.dat>

Vyrovnanie pre rôzne hodnoty α



Obr. 2.1: Vyrovnanie metódou jednoduchého exponenciálneho vyrovňovania pre rôzne hodnoty vyrovňovacej konštanty.

Vzorec (2.5) sa nazýva aj chybový (Cipra, 2008). Totiž v momente ako dostaneme pozorovanie y_t , vieme určiť chybu predpovede $e_t = y_t - \hat{y}_t(t-1)$ o jeden krok dopredu skonštruovanej v čase $t-1$. Vzorec (2.5) potom opravuje predchádzajúcu vyrovnanú hodnotu o redukovanú chybu predpovede.

Ako už bolo spomenuté vyššie, pre použitie tejto metódy v praxi sa využíva rekurentný vzorec (2.3). Pred jeho aplikáciou však musíme najskôr zvoliť počiatočnú hodnotu \hat{y}_0 a vyrovňovaciu konštantu α .

Za počiatočnú hodnotu \hat{y}_0 môžeme vziať aritmetický priemer niekoľkých počiatočných hodnôt y_t , alebo priamo prvé pozorovanie y_1 .

Vyrovňovaciu konštantu α volíme zvyčajne z intervalu $(0, 0.3]$ (Cipra, 2008) jedným z nasledujúcich spôsobov:

1. α zvolíme ako fixnú hodnotu $\alpha = 0.1$ alebo $\alpha = 0.2$;
2. porovnaním stredných vekov váh jednoduchých klzavých priemerov a exponenciálneho vyrovňovania, t.j. riešením rovnice

$$\begin{aligned} \sum_{k=0}^{2m} k \cdot \frac{1}{2m+1} &= \sum_{k=0}^{\infty} k \cdot (1-\beta) \cdot \beta^k, \\ \frac{1}{2m+1} \cdot \frac{2m(2m+1)}{2} &= (1-\beta) \cdot \beta \cdot \sum_{k=1}^{\infty} k \cdot \beta^{k-1}, \\ m &= (1-\beta) \cdot \beta \cdot \left(\sum_{k=1}^{\infty} \beta^k \right)', \\ m &= (1-\beta) \cdot \beta \cdot \frac{1}{(1-\beta)^2}, \\ m &= \frac{1-\alpha}{\alpha} \Rightarrow \alpha = \frac{1}{m+1}, \end{aligned}$$

kde $2m - 1$ by bola adekvátna dĺžka kľzavých priemerov;

3. výberom α zo siete tak, aby v danom rade poskytovala najlepšie predpovede v zmysle minimálnej hodnoty súčtu štvorcových chýb

$$SSE = \sum_{t=1}^T (y_t - \hat{y}_t)^2.$$

Poznámka 2.1.1. V príklade s ročnými zrážkami v Londýne, na ktorom sme ilustrovali vyrovnávanie časového radu metódou jednoduchého exponenciálneho vyrovnania pre rôzne hodnoty parametra α , je hodnota $\hat{\alpha} = 0.02$ odhadnutá softvérom vybraná práve spôsobom popísanom v bode 3.

V tretej kapitole predstavíme aj modernejší prístup k odhadu parametrov metód exponenciálneho vyrovnávania, ktorý vychádza z ich formulácie prostredníctvom štatistického modelu. Ten nám umožňuje parametre odhadovať aj s využitím metódy maximálnej vierohodnosti. Ako však ukážeme, v prípade že pre chyby časového radu predpokladáme normálne rozdelenie, tento odhad je ekvivalentný práve s bodom 3., teda s minimalizáciou súčtu štvorcových chýb.

2.2 Dvojité exponenciálne vyrovnávanie

Dvojité exponenciálne vyrovnávanie, niekedy označované aj ako Brownova metóda, sa používa v prípade, že môžeme trend vo vyrovnávanom časovom rade považovať za lokálne lineárny, tj. v krátkych úsekoch radu platí:

$$Tr_{t-j} = \beta_0 + \beta_1 \cdot (-j).$$

Označme odhady parametrov β_0 a β_1 vypočítané v čase t ako $b_0(t), b_1(t)$. Dostávame ich minimalizáciou výrazu

$$\sum_{j=0}^{\infty} [y_{t-j} - (\beta_0 + \beta_1 \cdot (-j))]^2 \beta^j \quad (2.6)$$

vzhľadom na β_0, β_1 , pričom β je opäť dopredu zvolená diskontná konštanta z intervalu $(0,1)$.

Derivujme výraz (2.6) a položme jeho parciálne derivácie rovné nule:

$$\begin{aligned} \frac{\partial}{\partial \beta_0} &= \sum_{j=0}^{\infty} \beta^j y_{t-j} - \beta_0 \cdot \sum_{j=0}^{\infty} \beta^j + \beta_1 \cdot \sum_{j=0}^{\infty} j \beta^j \stackrel{!}{=} 0, \\ \frac{\partial}{\partial \beta_1} &= \sum_{j=0}^{\infty} j \beta^j y_{t-j} - \beta_0 \cdot \sum_{j=0}^{\infty} j \beta^j + \beta_1 \cdot \sum_{j=0}^{\infty} j^2 \beta^j \stackrel{!}{=} 0. \end{aligned}$$

Tieto rovnice sa dajú zjednodušiť s využitím nasledujúcich vzorcov:

$$\sum_{j=0}^{\infty} \beta^j = \frac{1}{1-\beta}, \quad \sum_{j=0}^{\infty} j \beta^j = \frac{\beta}{(1-\beta)^2}, \quad \sum_{j=0}^{\infty} j^2 \beta^j = \frac{\beta(1+\beta)}{(1-\beta)^3}.$$

Ich aplikáciou dostávame sústavu rovníc

$$b_0(t) - \frac{\beta}{1-\beta} \cdot b_1(t) = (1-\beta) \cdot \sum_{j=0}^{\infty} \beta^j y_{t-j}, \quad (2.7)$$

$$\beta \cdot b_0(t) - \frac{\beta(1+\beta)}{1-\beta} \cdot b_1(t) = (1-\beta)^2 \cdot \sum_{j=0}^{\infty} j \beta^j y_{t-j}. \quad (2.8)$$

Označme

$$S_t = (1-\beta) \cdot \sum_{j=0}^{\infty} \beta^j y_{t-j}. \quad (2.9)$$

S_t sa nazýva jednoduchá vyrovnávací štatistika. Môžeme si všimnúť, že zodpovedá hodnote daného časového radu v čase t vyrovnaného metódou jednoduchého exponenciálneho vyrovnávania. Analogicky s (2.3) teda platí

$$S_t = \alpha y_t + (1-\alpha) S_{t-1}, \quad (2.10)$$

kde opäť $\alpha = 1 - \beta$.

Podobne definujeme dvojité vyrovnávací štatistiku $S_t^{[2]}$

$$S_t^{[2]} = (1-\beta) \cdot \sum_{j=0}^{\infty} \beta^j S_{t-j}.$$

Analogicky s (2.10) máme vzťah

$$S_t^{[2]} = \alpha S_t + (1-\alpha) S_{t-1}^{[2]}. \quad (2.11)$$

Pomocou nich môžeme sústavu (2.7), (2.8) prepísať do tvaru

$$\begin{aligned} b_0(t) - \frac{\beta}{1-\beta} \cdot b_1(t) &= S_t, \\ \beta \cdot b_0(t) - \frac{\beta(1+\beta)}{1-\beta} \cdot b_1(t) &= S_t^{[2]} - (1-\beta) \cdot S_t, \end{aligned}$$

z ktorého dostávame hľadané odhady ako

$$b_0(t) = 2S_t - S_t^{[2]}, \quad (2.12)$$

$$b_1(t) = \frac{1-\beta}{\beta} \cdot (S_t - S_t^{[2]}) = \frac{\alpha}{1-\alpha} \cdot (S_t - S_t^{[2]}). \quad (2.13)$$

Predpoveď o τ ($\tau > 0$) krokov dopredu sa potom v čase t počíta podľa vzťahu

$$\hat{y}_{t+\tau}(t) = b_0(t) + b_1(t) \cdot \tau = \left(2 + \frac{\alpha \cdot \tau}{1-\alpha}\right) S_t - \left(1 + \frac{\alpha \cdot \tau}{1-\alpha}\right) S_t^{[2]}.$$

Pre $\tau = 0$ dostávame vyrovnanú hodnotu daného časového radu

$$\hat{y}_t = 2S_t - S_t^{[2]}.$$

Hodnoty vyrovnávacích štatistík $S_t, S_t^{[2]}$ pritom počítame pomocou rekurentných vzťahov (2.10), (2.11).

Realizácia metódy dvojitého exponenciálneho vyrovnávania v praxi opäť vyžaduje voľbu počiatočných hodnôt. V tomto prípade musíme nastaviť počiatočné hodnoty vyrovnávacích štatistík S_t a $S_t^{[2]}$ a hodnotu vyrovnávacej konštanty α .

Počiatočné hodnoty jednoduchej a dvojitej vyrovnávacej štatistiky sa zvyčajne určia zo vzťahov (2.12) a (2.13) pre $t = 0$, pričom odhady $b_0(0)$ a $b_1(0)$ kladieme rovné bežným regresným odhadom parametrov β_0, β_1 vypočítaným tak, že preložíme priamku nejakým krátkym počiatočným úsekom radu.

Voľbu vyrovnávacej konštanty α môžeme robiť rovnako ako v prípade jednoduchého exponenciálneho vyrovnávania. Môžeme si zvoliť fixnú hodnotu $\alpha = 0.2$ alebo $\alpha = 0.3$, porovnať stredný vek váh jednoduchých kľzavých priemerov a váh dvojitého exponenciálneho vyrovnávania, alebo vybrať α zo siete. Porovnávaním stredného veku váh dostaneme

$$\alpha = \sqrt{\frac{1}{m+1}},$$

kde $2m + 1$ je dĺžka kľzavých priemerov, ktorá by bola pre daný časový rad adekvátne (Cipra, 2008).

2.3 Holtova metóda

Holtova metóda je zovšeobecnením dvojitého exponenciálneho vyrovnávania, teda predpokladá časový rad s lokálne lineárnym trendom. Používa dve vyrovnávacie konštanty, konštantu α na vyrovnanie úrovne L_t časového radu a γ na vyrovnanie smernice T_t . Obe konštanty sa volia z intervalu $(0,1)$. Platí:

$$\begin{aligned} L_t &= \alpha y_t + (1 - \alpha)(L_{t-1} + T_{t-1}), \\ T_t &= \gamma(L_t - L_{t-1}) + (1 - \gamma)T_{t-1}, \\ \hat{y}_t &= L_t, \\ \hat{y}_{t+\tau}(t) &= L_t + T_t \cdot \tau, \quad (\tau > 0). \end{aligned}$$

Ako počiatočné hodnoty sa odporúča zvoliť $L_0 = y_1$ a $T_0 = y_2 - y_1$, dá sa ich však voliť rôznymi spôsobmi, viď (Cipra, 2008) a ďalšie kapitoly.

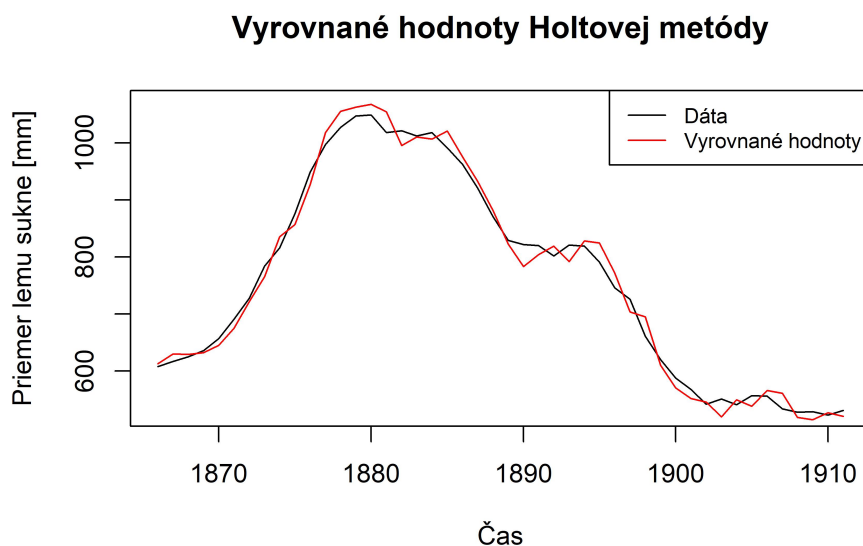
Z prvej rovnice vidíme, že úroveň L_t časového radu je váženým priemerom úrovne v čase t vyjadrenej hodnotou y_t a úrovne pre čas t odhadnutej v čase $t - 1$. Smernica T_t je vážený priemer odhadu smernice v čase t , vypočítanej z hodnôt L_t a L_{t-1} , a predchádzajúcej smernice v čase $t - 1$.

Pri Holtovej metóde predpokladáme lokálne lineárny trend. Hodnoty L_t a T_t potom môžeme chápať ako odhady absolútneho člena, resp. smernice, priamky popisujúcej tento trend v čase t , vypočítané na základe hodnôt y_1, \dots, y_t .

Holtova metóda bola pôvodne navrhnutá ad-hoc, iba na základe logických úvah. Až neskôr sa podarilo dokázať, že dvojité exponenciálne vyrovnávanie s vyrovnávacou konštantou α je jej špeciálnym prípadom (Cipra, 2008), ak za vyrovnávacie konštanty $\alpha_{Holt}, \gamma_{Holt}$ Holtovej metódy zvolíme

$$\alpha_{Holt} = \alpha(2 - \alpha), \quad \gamma_{Holt} = \frac{\alpha}{2 - \alpha}.$$

Vyrovnanie časového radu Holtovou metódou je ilustrované na pozorovaniach priemeru lemu dámskych sukni v milimetroch medzi rokmi 1866 až 1911³ na obrázku 2.2. Parametre vystupujúce v rovniciach vyššie boli odhadnuté softvérom ako $\hat{\alpha} = 0.8436$, $\hat{\gamma} = 0.8436$. Môžeme si všimnúť, že vyrovnané hodnoty celkom dobre kopírujú správanie časového radu, avšak zmena trendu z rastúceho na klesajúci a naopak sa vždy prejaví s malým oneskorením.



Obr. 2.2: Vyrovnanie časového radu pre priemer lemu dámskych sukni medzi rokmi 1866 až 1911.

Predpovede budúcich hodnôt $\hat{y}_{t+\tau}(t)$ generované Holtovou metódou vykazujú konštantný, do nekonečna rastúci (v prípade že $T_t > 0$) alebo do nekonečna klesajúci (pre $T_t < 0$) trend. Preto v prípade, že konštruujeme predpovede do vzdialenejšej budúcnosti, môžeme použitím tejto metódy dospieť k príliš vysokej alebo príliš nízkej hodnote. Tento fakt bol motiváciou k zavedeniu ďalšieho parametra, ktorý postupne v budúcnosti trend utlmí. Prvý krát bol predstavený v (Gardner Jr a McKenzie, 1985).

2.3.1 Metódy s tlmeným trendom

Vyrovňavacie rovnice pre úroveň časového radu L_t a smernicu T_t majú v tomto prípade tvar:

$$L_t = \alpha y_t + (1 - \alpha)(L_{t-1} + \phi T_{t-1}), \quad (2.14)$$

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)\phi T_{t-1}, \quad (2.15)$$

kde $0 < \phi < 1$. Vyrovnanú hodnotu \hat{y}_t a predpoveď $\hat{y}_{t+\tau}(t)$ o $\tau > 0$ krokov dopredu skonštruovanú v čase t potom dostaneme ako:

$$\begin{aligned} \hat{y}_t &= L_t, \\ \hat{y}_{t+\tau}(t) &= L_t + (\phi + \phi^2 + \dots + \phi^\tau)T_t. \end{aligned}$$

³zdroj: <http://www.robjhyndman.com/tsdldata/roberts/skirts.dat>

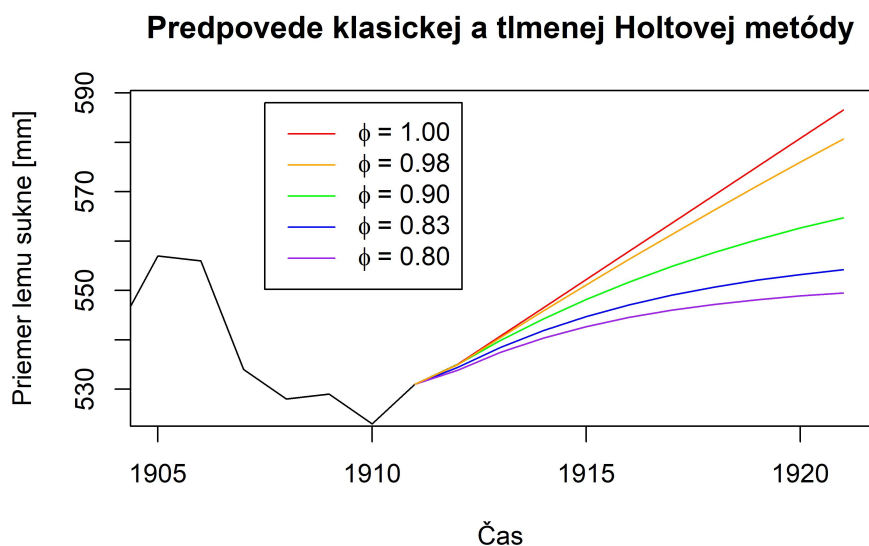
Ak položíme $\phi = 1$, dostávame Holtovu metódu. Pre hodnoty ϕ medzi 0 a 1, platí:

$$\hat{y}_{t+\tau}(t) = L_t + (\phi + \phi^2 + \dots + \phi^\tau)T_t = L_t + \frac{\phi \cdot (\phi^\tau - 1)}{\phi - 1} \cdot T_t. \quad (2.16)$$

Parameter ϕ tlmí trend tak, že v nejakom budúcom čase sa blíži ku konštante. Z rovnice (2.16) vidíme, že pre τ blížiacie sa k nekonečnu, predpoveď konverguje k $L_T + \frac{\phi}{1-\phi} \cdot T_T$ pre ľubovoľné ϕ z intervalu $(0,1)$.

Parameter ϕ sa zvyčajne obmedzuje na hodnoty v rozmedzí 0.8 až 0.98 (Hyndman a Athanasopoulos, 2018). Pre menšie hodnoty je tlmiaci efekt príliš výrazný, pre hodnoty blízke 1 sa naopak stráca jeho význam, keďže dostávame predpovede veľmi blízke tým z netlmeného modelu.

Rozdiel medzi predpoveďami klasickej Holtovej metódy a jej verzie s tlmným trendom je zobrazený na obrázku 2.3. Metódy sme opäť aplikovali na časový rad priemeru lemu dámskych sukni medzi rokmi 1866 a 1911, a vypočítali sme predpovede na 10 nasledujúcich rokov.



Obr. 2.3: Porovnanie predpovedí klasickej a tlmenej Holtovej metódy.

Červenou sú zobrazené predpovede pre $\phi = 1$, teda pre klasickú Holtovu metódu. Žltá a fialová predstavujú predpovede pre odporúčané hraničné hodnoty tlmiaceho parametra ϕ . Metóda zobrazená modrou farbou má hodnotu parametra $\phi = 0.83$, čo je hodnota odhadnutá softvérom (metódy odhadu budú popísané v ďalších častiach). Pre porovnanie, predpoveď klasickej Holtovej metódy o 10 rokov dopredu má hodnotu $\hat{y}_{1911+10}(1911) = 586.57$, zatiaľ čo tlmená metóda s odhadnutou hodnotou parametra ϕ túto hodnotu predpovedala ako $\hat{y}_{1911+10}(1911) = 554.20$.

2.4 Holtova-Wintersova metóda

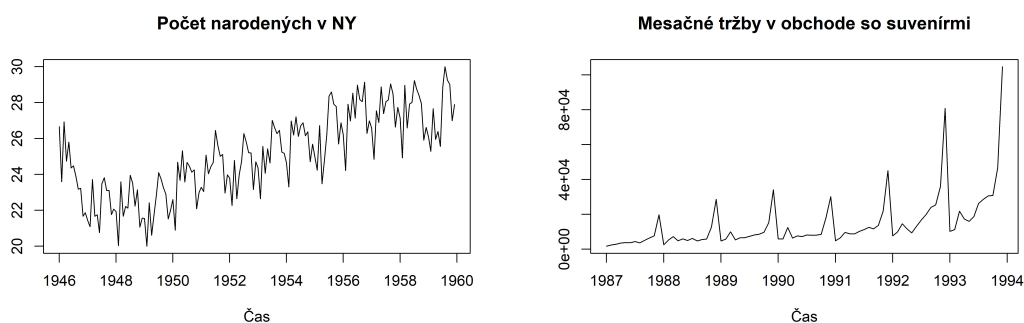
Doteraz sme predpokladali, že vyrovnávaný časový rad pozostáva iba z trendovej a reziduálnej zložky. V tejto časti popíšeme rozšírenie Holtovej metódy, Holtovu-Wintersovu metódu, (Holt, 2004)⁴ a (Winters, 1960), pomocou ktorej budeme schopní vyrovnávať časové rady obsahujúce aj sezónnu zložku S_t .

Holtova-Wintersova metóda pozostáva z troch vyrovnávacích rovníc. Tak ako pri Holtovej metóde máme jednu pre vyrovnanie úrovne a jednu pre vyrovnanie smernice, pribudne nám však ďalšia pre vyrovnávanie sezónnosti. Zodpovedajúce vyrovnávacie konštanty označíme v poradí α , γ a δ . Frekvenciu sezónnosti budeme označovať písmenom m . Potom napríklad pre mesačné dáta máme $m = 12$, pre štvrťročné $m = 4$, atď.

Existujú dve varianty Holtovej-Wintersovej metódy, ktoré sa líšia podľa toho, či uvažujeme aditívnu alebo multiplikatívnu sezónnosť, teda či predpokladáme, že vyrovnávaný časový rad má tvar

$$y_t = Tr_t + S_t + E_t \quad \text{alebo} \quad y_t = Tr_t \cdot S_t + E_t.$$

Aditívna sezónnosť sa zvyčajne používa v prípade, že sezónne vplyvy sú približne konštantné, teda keď zmena sezónnosti nekorešponduje so zmenou trendu. V prípade, že sú sezónne zmeny proporcionálne zmenám úrovne časového radu, preferuje sa naopak multiplikatívna sezónnosť (Cipra, 2008). Porovnanie časových radov s aditívnou⁵ (vľavo) a multiplikatívnou⁶ sezónnosťou je na obrázku 2.4.



Obr. 2.4: Časový rad s aditívnou (vľavo) a multiplikatívnou sezónnosťou.

Pri aditívnej metóde časový rad sezónne očistujeme odčítaním sezónnej zložky a platí, že za rok by sa mali hodnoty sezónnosti nasčítat približne na nulu. Pri multiplikatívnej metóde vykonávame sezónne očistenie delením sezónnou zložkou a sčítaním hodnôt sezónnosti cez jeden rok by sme mali dostať hodnotu m (Cipra, 2008).

V tejto práci sa však zaoberáme iba časovými radmi tvaru (1.1), a teda uvažujeme iba aditívnu sezónnu zložku.

⁴Originálna verzia je správa Úradu pre námorný výskum (Office of Naval Research, ONR) z roku 1957, až do roku 2004 však nebola vydaná.

⁵zdroj: <https://robjhyndman.com/tsdldata/data/fancy.dat>

⁶zdroj: <https://robjhyndman.com/tsdldata/data/nybirths.dat>

2.4.1 Holtova-Wintersova aditívna metóda

Na vyrovnanie úrovne L_t , smernice T_t a sezónnosti S_t používame v aditívnej metóde rovnice:

$$L_t = \alpha(y_t - S_{t-m}) + (1 - \alpha)(L_{t-1} + T_{t-1}), \quad (2.17)$$

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)T_{t-1}, \quad (2.18)$$

$$S_t = \delta(y_t - L_{t-1} - T_{t-1}) + (1 - \delta)S_{t-m}. \quad (2.19)$$

Z prvej rovnice vidíme, že úroveň L_t je váženým priemerom sezónne očisteného pozorovania ($y_t - S_{t-m}$) a nesezónnej predpovede pre čas t vypočítanej v čase $t-1$. Rovnica smernice je identická ako v Holtovej metóde. Čo sa týka rovnice pre sezónnu zložku, ide o vážený priemer súčasnej hodnoty sezónnosti a minuloročnej hodnoty z tej istej sezóny.

Posledná rovnica (2.19) sa často uvádza aj v tvare (Hyndman a Athanasopoulos, 2018)

$$S_t = \delta^*(y_t - L_t) + (1 - \delta^*)S_{t-m}.$$

Ak do nej dosadíme L_t z rovnice (2.17), dostávame

$$S_t = \delta^*(1 - \alpha)(y_t - L_{t-1} - T_{t-1}) + [1 - \delta^*(1 - \alpha)]S_{t-m},$$

čo je ekvivalentné s tvarom (2.19), ak za δ zvolíme $\delta^*(1 - \alpha)$. Z obmedzenia $0 < \delta^* < 1$ potom dostávame podmienku $0 < \delta < 1 - \alpha$.

Vyrovnané hodnoty a predpovede s použitím tejto metódy počítame ako:

$$\begin{aligned} \hat{y}_t &= L_t + S_t, \\ \hat{y}_{t+\tau}(t) &= L_t + T_t \cdot \tau + S_{t+\tau-m}, \quad \tau = 1, 2, \dots, m, \\ &= L_t + T_t \cdot \tau + S_{t+\tau-2m}, \quad \tau = m + 1, m + 2, \dots, 2m, \\ &\vdots \end{aligned}$$

V tomto prípade musíme rozlišovať vzorce pre predpovede v rámci rôznych rokov. Na ich výpočet totiž chceme vždy použiť hodnoty sezónnych zložiek z posledného pozorovaného roku.

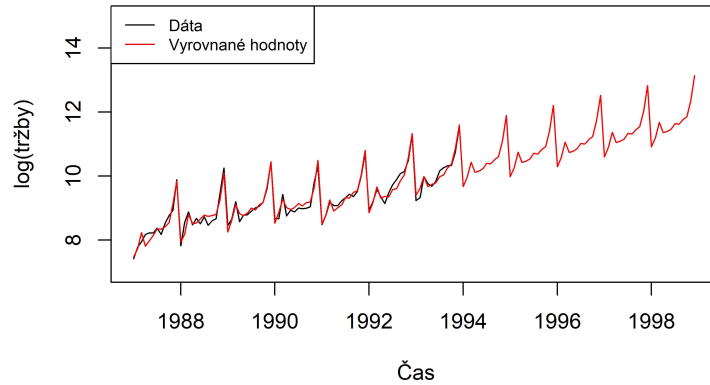
Príklad vyrovnaní časového radu použitím Holtovej-Wintersovej metódy je na obrázku 2.5. Metódu sme aplikovali na rad logaritmických mesačných tržieb obchodu so suvenírmí v plážovom rezorte v Austrálii od januára 1987 do decembra 1993. Tieto dáta sme dostali logaritmickou transformáciou pozorovaní na obrázku 2.4 vpravo.

Môžeme si všimnúť, že vyrovnané hodnoty naozaj zachytili sezónne vplyvy v danom časovom rade. Tie boli taktiež prenesené aj do predpovedí budúcich hodnôt na najbližších 5 rokov.

2.4.2 Tlmená Holtova-Wintersova metóda

V predchádzajúcich častiach popísanú metódu tlmenia trendu je možné aplikovať aj na Holtovu-Wintersovu metódu, či už jej aditívnu alebo multiplikatívnu verziu.

Predpovede Holtovej-Wintersovej metódy



Obr. 2.5: Vyrovanie a predpoveď časového radu logaritmickej mesačných tržieb v obchode so suvenírmi počas rokov 1987 až 1993.

Pri aditívnej metóde majú vyrovnávacie rovnice pre úroveň, trend a sezónnosť tvar:

$$\begin{aligned} L_t &= \alpha (y_t - S_{t-m}) + (1 - \alpha) (L_{t-1} + \phi T_{t-1}), \\ T_t &= \gamma (L_t - L_{t-1}) + (1 - \gamma) \phi T_{t-1}, \\ S_t &= \delta (y_t - L_{t-1} - T_{t-1}) + (1 - \delta) S_{t-m}, \end{aligned}$$

kde znovu parametre volíme $0 < \alpha, \gamma, \delta < 1$. Vyrovnané hodnoty a predpovede budúcich hodnôt o $\tau > 0$ krokov dopredu počítame podľa vzorcov:

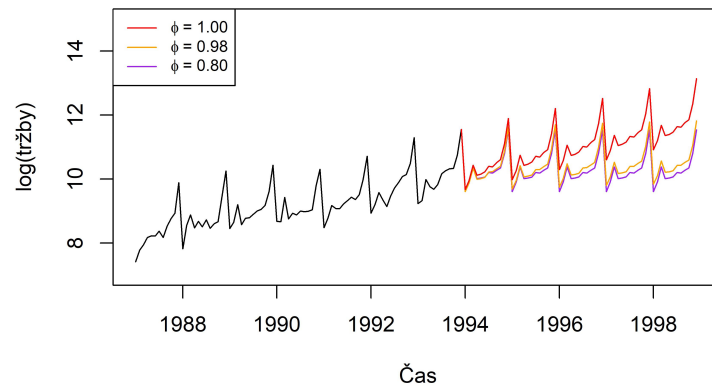
$$\begin{aligned} \hat{y}_t &= L_t + S_t, \\ \hat{y}_{t+\tau}(t) &= L_t + (\phi + \phi^2 + \dots + \phi^\tau) T_t + S_{t+\tau-m}, \quad \tau = 1, 2, \dots, m, \\ &= L_t + (\phi + \phi^2 + \dots + \phi^\tau) T_t + S_{t+\tau-2m}, \quad \tau = m + 1, \dots, 2m, \\ &\vdots \end{aligned}$$

Na obrázku 2.6 je zobrazené porovnanie predpovedí Holtovej-Wintersovej metódy pre časový rad logaritmickej mesačných tržieb v obchode so suvenírmi z predchádzajúceho príkladu pre rôzne hodnoty tlmiaceho parametra ϕ . Predpovede pre $\phi = 1$ zobrazené červenou predstavujú predpovede klasickej Holtovej-Wintersovej metódy, žltou a fialovou sú vykreslené predpovede pre odporúčané hraničné hodnoty tlmiaceho parametra.

2.5 Klasifikácia metód exponenciálneho vyrovnávania

Doteraz prezentované predpovedné metódy z triedy exponenciálneho vyrovnávania nie sú všetky, ktoré existujú. Kombináciami rôznych typov trendu a sezónnosti je možné vytvoriť až pätnásť odlišných metód. Zvyčajne sa označujú dvojicou písmen (T, S) , kde $T \in \{N, A, A_d, M, M_d\}$ je uvažovaný trend, pričom

Predpovede Holtovej-Wintersovej metódy



Obr. 2.6: Porovnanie predpovedí klasickej a tlmenej Holtovej-Wintersovej metódy.

N je žiaden, A aditívny a M multiplikatívny trend. Index d označuje tlmenný aditívny alebo tlmenný multiplikatívny trend. Ak označíme $Tr_{t+\tau}$ predpovedanú budúcu hodnotu trendu o τ krokov dopredu, v našom značení z predchádzajúcich častí jednotlivé typy trendu znamenajú nasledovné:

$$\begin{aligned} N : Tr_{t+\tau} &= L_t, \\ A : Tr_{t+\tau} &= L_t + T_t \cdot \tau, \\ A_d : Tr_{t+\tau} &= L_t + (\phi + \phi^2 + \dots + \phi^\tau) \cdot T_t, \\ M : Tr_{t+\tau} &= L_t \cdot T_t^\tau, \\ M_d : Tr_{t+\tau} &= L_t \cdot T_t^{(\phi + \phi^2 + \dots + \phi^\tau)}. \end{aligned}$$

Poznámka 2.5.1. Zrejme by bolo intuitívnejšie nazývať trend $Tr_{t+\tau} = L_t$ konštantným a nie žiadnym, $Tr_{t+\tau} = L_t + T_t \tau$ lineárnym, nie aditívnym, a podobne, avšak v tomto prípade ide o zaužívanú konvenciu, preto budeme aj ďalej používať zavedené označenie.

S v dvojici (T, S) , $S \in \{N, A, M\}$ je typ sezónnosti, kde N , A a M opäť znamenajú žiadnu, aditívnu alebo multiplikatívnu sezónnosť.

Niektoré metódy prezentované v predchádzajúcich častiach sa využitím tejto klasifikácie dajú označovať nasledovne:

- (N, N) je jednoduché exponenciálne vyrovnanie,
- (A, N) je Holtova metóda,
- (A_d, N) je Holtova metóda s tlmenným trendom,
- (A, A) je Holtova-Wintersova aditívna metóda,
- (A, M) je Holtova-Wintersova multiplikatívna metóda.

K dvojici (T, S) môžeme navyše uvažovať ešte typ chybovej zložky. Ku každej z pätnástich metód pre rôzne typy trendu a sezónnosti existuje alternatíva s aditívnou alebo multiplikatívnou chybovou zložkou. Celkovo potom máme 30

modelov exponenciálneho vyrovnávania, špecifikovaných trojicou (E,T,S) , kde $E \in \{A,M\}$ reprezentuje uvažovaný typ chybovej zložky, pričom A znamená opäť aditívnu, M multiplikatívnu chybu, a T a S sú stále uvažovaný typ trendu a sezónnosti.

Tvar vyrovnávacích rovníc pre úroveň, smernicu a sezónnosť, rovnako ako rovnice pre vyrovnané hodnoty a bodové predpovede každej z týchto metód možno nájsť napríklad v (Hyndman a Athanasopoulos (2018), tabuľka 7.7).

3. Stavové modely

V tejto kapitole sa budeme zaoberať štatistickými modelmi, ktoré sú základom pre popísané metódy exponenciálneho vyrovnávania. Tieto modely nám umožnia konštruovať nielen bodové, ale aj intervalové predpovede, a študovať pravdepodobnostné rozdelenie budúcich hodnôt časového radu. Vychádzať pritom budeme hlavne z (Hyndman a kol., 2008) a (Hyndman a Athanasopoulos, 2018).

Ďalej prezentovaný prístup sa objavil v 90-tych rokoch minulého storočia a prvýkrát bol formulovaný zrejme v (Ord a kol., 1997). Metódy exponenciálneho vyrovnávania boli totiž pôvodne formulované ad-hoc a nemali štatistický základ.

3.1 Všeobecné stavové modely

Nech y_t je aj naďalej pozorovanie v čase t , a \mathbf{x}_t je takzvaný vektor stavov, ktorý obsahuje nepozorované hodnoty úrovne, trendu a sezónnosti daného časového radu v čase t , $\mathbf{x}_t = (L_t, T_t, S_t, S_{t-1}, \dots, S_{t-m+1})^T$. Stavový model potom vieme zapísať ako

$$y_t = w(\mathbf{x}_{t-1}) + r(\mathbf{x}_{t-1})\epsilon_t, \quad (3.1)$$

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}) + g(\mathbf{x}_{t-1})\epsilon_t, \quad (3.2)$$

kde predpokladáme, že $\{\epsilon_t\}$ je gausovský biely šum s rozptylom σ^2 . Rovnica (3.1) sa nazýva rovnicou pozorovania, rovnica (3.2) je stavová rovnica.

S takto špecifikovaným modelom môžeme zistiť, napríklad, aká je podmienená stredná hodnota budúceho pozorovania, za predpokladu že poznáme minulé hodnoty. Toto označujeme ako $\mu_{t+\tau|t} = E(y_{t+\tau}|\mathbf{x}_t)$. Pre mnoho modelov budú tieto podmienené stredné hodnoty identické s bodovými predpoveďami metód, ktoré reprezentujú, tj. $\mu_{t+\tau|t} = \hat{y}_{t+\tau}(t)$. Vo všeobecnosti to však nebude platiť pre $\tau \geq 2$ v modeloch s multiplikatívnym trendom alebo multiplikatívnou sezónnosťou (Hyndman a kol., 2008). Ďalej budeme pre jednoduchosť $\mu_{t+1|t}$ značiť len μ_{t+1} . Pre model s aditívnou chybovou zložkou je $r(\mathbf{x}_{t-1}) = 1$, aby platilo $y_t = \mu_t + \epsilon_t$.

Každý z tridsiatich modelov ETS(\cdot, \cdot, \cdot) sa dá zapísať v tvare (3.1), (3.2). Príslušné rovnice pre μ_t , L_t , T_t a S_t sa dajú nájsť napríklad v (Hyndman a kol. (2008), tabuľky 2.2 a 2.3). Ďalej niektoré z nich podrobnejšie popíšeme aj v tejto práci.

Niektoré z kombinácií chyby, trendu a sezónnosti môžu viesť k numerickým problémom. Napríklad modely, ktoré vyžadujú delenie niektorou zo stavových zložiek, môžu obsahovať delenie nulou alebo hodnotou veľmi blízkou nule. Konkrétne medzi nestabilné modely patria tri modely s multiplikatívnou sezónnosťou a to ETS(A, N, M), ETS(A, A, M) a ETS(A, A_d, M), ktoré sa zvyčajne ani neuvažujú pri výbere najvhodnejšieho modelu (Hyndman a Athanasopoulos, 2018).

Modely s multiplikatívnym trendom sú zase užitočné v prípade, že dáta sú striktné kladné, ale nie sú stabilné, ak pozorovania obsahujú nuly alebo negatívne hodnoty (Hyndman a Athanasopoulos, 2018). Preto v prípade, že pracujeme s dátami, ktoré nie sú kladné, sa aplikuje len šesť čisto aditívnych modelov. V ďalších častiach práce sa zameriame práve na tieto modely a budeme teda uva-

žovať iba nasledujúce kombinácie typu chyby, trendu a sezónnosti: ETS(A,N,N), ETS(A,A,N), ETS(A,A_d,N), ETS(A,N,A), ETS(A,A,A), ETS(A,A_d,A).

3.2 Lineárne stavové modely

V prípade, že uvažujeme len šesť čisto aditívnych modelov, môžeme rovnice (3.1), (3.2) písať v tvare

$$y_t = \mathbf{w}'\mathbf{x}_{t-1} + \epsilon_t, \quad (3.3)$$

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{g}\epsilon_t, \quad (3.4)$$

kde \mathbf{w} a \mathbf{g} sú vektory a \mathbf{F} je matica. Oba vektory aj matica sú pevné. Člen $\mathbf{w}'\mathbf{x}_{t-1}$ v rovnici (3.3) reprezentuje vplyv minulosti na hodnotu y_t , ϵ_t predstavuje jej nepredpovedateľnú, náhodnú časť. Predpokladáme, že ϵ_t je gaussovský biely šum s rozptylom σ^2 . Rovnica (3.4) je rekurentným vzťahom pre vývoj stavového vektora v čase. Člen $\mathbf{F}\mathbf{x}_{t-1}$ je opäť efekt minulosti na súčasný stav \mathbf{x}_t , $\mathbf{g}\epsilon_t$ je náhodná zmena v \mathbf{x}_t .

Poznámka 3.2.1. Takto formulovaný stavový model pripomína Kalmanov filter, popísaný napríklad v (Cipra, 2008). Hlavný rozdiel medzi týmito modelmi spočíva v tom, že Kalmanov filter predpokladá, že náhodné chyby, ktoré vystupujú v rovnici pozorovania a v stavovej rovnici, pochádzajú z normálnych rozdelení s rôznymi rozptylmi a že sú nekorelované. Stavové modely exponenciálneho vyrovnávania však uvažujú jednu náhodnú chybu v čase t , ktorá ovplyvňuje pozorovanie aj stavový vektor v tomto čase.

Ilustrujme myšlienku takéhoto štatistického modelu na príklade klasickej Holtovej metódy aj Holtovej metódy s tlmeným trendom.

Príklad 3.2.1 (Stavový model pre Holtovu metódu). Nech $\mu_t = \hat{y}_t = L_{t-1} + T_{t-1}$ označuje jedнокrokovú predpoveď hodnoty y_t za predpokladu, že poznáme hodnoty všetkých parametrov do času $t-1$. Parametrami v tomto prípade rozumieme hodnoty úrovne L_t a trendu T_t . Ďalej nech $\epsilon_t = y_t - \mu_t$ je chyba tejto predpovede v čase t . Potom vieme, že

$$y_t = L_{t-1} + T_{t-1} + \epsilon_t, \quad (3.5)$$

a dosadením do vyrovnávacích rovníc z predchádzajúcej kapitoly dostaneme

$$\begin{aligned} L_t &= \alpha y_t + (1 - \alpha)(L_{t-1} + T_{t-1}) = \\ &= \alpha(L_{t-1} + T_{t-1} + \epsilon_t) + (1 - \alpha)(L_{t-1} + T_{t-1}) = \\ &= L_{t-1} + T_{t-1} + \alpha\epsilon_t, \\ T_t &= \gamma(L_t - L_{t-1}) + (1 - \gamma)T_{t-1} = \\ &= \gamma(L_{t-1} + T_{t-1} + \alpha\epsilon_t - L_{t-1}) + (1 - \gamma)T_{t-1} = \\ &= T_{t-1} + \alpha\gamma\epsilon_t. \end{aligned}$$

Pre jednoduchosť označme $\beta = \alpha\gamma$, $0 < \beta < 1$ a môžeme písať

$$L_t = L_{t-1} + T_{t-1} + \alpha\epsilon_t, \quad (3.6)$$

$$T_t = T_{t-1} + \beta\epsilon_t. \quad (3.7)$$

Potom položíme $\mathbf{x}_t = (L_t, T_t)'$ a rovnice (3.5)-(3.7) môžeme prepísať v tvare

$$\begin{aligned} y_t &= (1, 1) \mathbf{x}_{t-1} + \epsilon_t, \\ \mathbf{x}_t &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \mathbf{x}_{t-1} + \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \epsilon_t. \end{aligned}$$

Vo všeobecnosti pracujeme s predpokladom, že postupnosť $\{\epsilon_t\}$ je gausovský biely šum s rozptylom σ^2 .

Príklad 3.2.2 (Stavový model pre Holtovu metódu s tlmeným trendom). Tak ako v prechádzajúcom príklade, opäť označme μ_t jednokrokovú predpoveď hodnoty y_t z času $t - 1$, teda uvažujme, že $\mu_t = \hat{y}_t = L_{t-1} + \phi T_{t-1}$. Nech $\epsilon_t = y_t - \mu_t$ a teda

$$y_t = L_{t-1} + \phi T_{t-1} + \epsilon_t. \quad (3.8)$$

Analogicky ako pri klasickej Holtovej metóde, dosadením do vyrovnávacích rovníc dostávame

$$L_t = L_{t-1} + \phi T_{t-1} + \alpha \epsilon_t, \quad (3.9)$$

$$T_t = \phi T_{t-1} + \alpha \gamma \epsilon_t, \quad (3.10)$$

a ak položíme $\beta = \alpha \gamma, 0 < \beta < 1$ môžeme rovnice (3.8)-(3.10) zapísať maticovo ako

$$\begin{aligned} y_t &= (1, \phi) \mathbf{x}_{t-1} + \epsilon_t, \\ \mathbf{x}_t &= \begin{pmatrix} 1 & \phi \\ 0 & \phi \end{pmatrix} \mathbf{x}_{t-1} + \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \epsilon_t. \end{aligned}$$

V tabuľke 3.1 sú tvary vektorov $\mathbf{x}_t, \mathbf{w}, \mathbf{g}$ a matice \mathbf{F} pre všetkých šesť aditívnych modelov. Doteraz sme používali vyrovnávaciu konštantu α pre vyrovnanie úrovne, γ pre vyrovnanie trendu a δ pre vyrovnanie sezónnosti. V tabuľke 3.1 sme však zaviedli označenie konzistentné s (Hyndman a kol., 2008) a (Hyndman a Athanasopoulos, 2018), nakoľko rovnaké označenie parametrov využíva aj v piatej kapitole používaný softvér R. Význam vyrovnávacej konštanty α zostáva nezmenený, β označuje súčin $\alpha \gamma$, rovnako ako v príkladoch 3.2.1 a 3.2.2, a parameter δ pre vyrovnanie sezónnej zložky, ktorý sme uvažovali pri Holtovej-Wintersovej metóde, sme v tejto tabuľke preznačili na γ .

3.2.1 Odhad parametrov

V stavových modeloch sú počiatkové stavy a parametre zvyčajne neznáme a musíme ich odhadovať na základe pozorovaných dát. V tejto práci však budeme pracovať so zjednodušujúcim predpokladom, že dáta generujúci proces začal v čase 1, a teda budeme uvažovať vektor počiatkových stavov \mathbf{x}_0 za fixný a budeme s ním pracovať ako s vektorom parametrov. V prípade že by sme uvažovali, že dáta generujúci proces prebiehal už dlhé časové obdobie pred prvým pozorovaním, vektor počiatkových stavov by sme museli brať ako náhodný, nemali by sme však k dispozícii jeho pozorovania (Hyndman a kol., 2008).

Jednou z metód, ktoré sa dajú na účel odhadovania parametrov použiť, je metóda maximálnej vierohodnosti. Vierohodnostná funkcia závisí na vektore θ

Model	\mathbf{x}_t	\mathbf{w}	\mathbf{g}	\mathbf{F}
(A,N,N)	L_t	1	α	1
(A,A,N)	$(L_t, T_t)'$	$(1,1)'$	$(\alpha, \beta)'$	$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$
(A, A_d, N)	$(L_t, T_t)'$	$(1, \phi)'$	$(\alpha, \beta)'$	$\begin{pmatrix} 1 & \phi \\ 0 & \phi \end{pmatrix}$
(A, N, A)	$(L_t, S_t, S_{t-1}, \dots, S_{t-(m-1)})'$	$(1, \underbrace{0, \dots, 0}_{m-1}, 1)'$	$(\alpha, \gamma, \underbrace{0, \dots, 0}_{m-1})'$	$\begin{pmatrix} 1 & \mathbf{0}'_{m-1} & 0 \\ 0 & \mathbf{0}'_{m-1} & 1 \\ \mathbf{0}_{m-1} & \mathbf{I}_{m-1} & \mathbf{0}_{m-1} \end{pmatrix}$
(A, A, A)	$(L_t, T_t, S_t, S_{t-1}, \dots, S_{t-(m-1)})'$	$(1, 1, \underbrace{0, \dots, 0}_{m-1}, 1)'$	$(\alpha, \beta, \gamma, \underbrace{0, \dots, 0}_{m-1})'$	$\begin{pmatrix} 1 & 1 & \mathbf{0}'_{m-1} & 0 \\ 0 & 1 & \mathbf{0}'_{m-1} & 0 \\ 0 & 0 & \mathbf{0}'_{m-1} & 1 \\ \mathbf{0}_{m-1} & \mathbf{0}_{m-1} & \mathbf{I}_{m-1} & \mathbf{0}_{m-1} \end{pmatrix}$
(A, A_d, A)	$(L_t, T_t, S_t, S_{t-1}, \dots, S_{t-(m-1)})'$	$(1, \phi, \underbrace{0, \dots, 0}_{m-1}, 1)'$	$(\alpha, \beta, \gamma, \underbrace{0, \dots, 0}_{m-1})'$	$\begin{pmatrix} 1 & \phi & \mathbf{0}'_{m-1} & 0 \\ 0 & \phi & \mathbf{0}'_{m-1} & 0 \\ 0 & 0 & \mathbf{0}'_{m-1} & 1 \\ \mathbf{0}_{m-1} & \mathbf{0}_{m-1} & \mathbf{I}_{m-1} & \mathbf{0}_{m-1} \end{pmatrix}$

Tabuľka 3.1: Špecifikácia aditívnych modelov (ETS).

parametrov akými sú napríklad vyrovnávacie parametre, v predchádzajúcich častiach označené ako α, β a γ , alebo tlmiaci faktor ϕ , ktoré v modeli (3.1), (3.2) vystupujú vo vektoroch \mathbf{g} a \mathbf{w} . Ďalej je funkciou rozptylu chybových členov σ^2 a vektora \mathbf{x}_0 počiatočných stavov.

Označme \mathbf{y} vektor pozorovaných hodnôt, tj. $\mathbf{y} = (y_1, \dots, y_T)'$. Potom vierohodnostná funkcia $\mathcal{L}(\mathbf{y}; \theta, \sigma^2, \mathbf{x}_0)$ má tvar

$$\mathcal{L}(\mathbf{y}; \theta, \sigma^2, \mathbf{x}_0) = \prod_{t=1}^T f_{y_t}(y_t | y_1, \dots, y_{t-1}, \theta, \sigma^2, \mathbf{x}_0) = \prod_{t=1}^T f_{y_t}(y_t | \theta, \sigma^2, \mathbf{x}_{t-1}) = \prod_{t=1}^T f_\epsilon(\epsilon_t),$$

kde funkcia f_{y_t} označuje hustotu rozdelenia hodnôt y_t , a f_ϵ je hustota chybových členov. Vzťah medzi y_t a ϵ_t je pritom určený rovnicou (3.3) ako

$$\epsilon_t = y_t - \mathbf{w}'\mathbf{x}_{t-1},$$

kde $\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{g}\epsilon_t$. Hodnoty ϵ_t , $t = 1, \dots, T$, teda vieme vyjadriť pomocou známych hodnôt \mathbf{y} a pomocou odhadovaných parametrov.

Ak predpokladáme, že ϵ_t má normálne rozdelenie so strednou hodnotou nula a rozptylom σ^2 , tak ďalej môžeme písať

$$\mathcal{L}(\mathbf{y}; \theta, \sigma^2, \mathbf{x}_0) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left(-\frac{1}{2} \sum_{t=1}^T \frac{\epsilon_t^2}{\sigma^2}\right) \quad (3.11)$$

a logaritmickej vierohodnostnej funkcie je

$$\ell(\mathbf{y}; \theta, \sigma^2, \mathbf{x}_0) = \log \mathcal{L}(\mathbf{y}; \theta, \sigma^2, \mathbf{x}_0) = -\frac{T}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{t=1}^T \frac{\epsilon_t^2}{\sigma^2}.$$

Tú môžeme zderivovať podľa σ^2 a deriváciu položiť rovnú nule, aby sme dostali odhad $\hat{\sigma}^2$

$$\frac{\partial \ell(\mathbf{y}; \theta, \sigma^2, \mathbf{x}_0)}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2} \sum_{t=1}^T \frac{\epsilon_t^2}{(\sigma^2)^2} \stackrel{!}{=} 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum_{t=1}^T \epsilon_t^2}{T}.$$

Po dosadení tohto odhadu do rovnice (3.11) dostaneme profilovú vierohodnosť

$$\mathcal{L}^*(\mathbf{y}; \theta, \mathbf{x}_0) = (2\pi\hat{\sigma}^2)^{-\frac{T}{2}} \exp\left(-\frac{T}{2\sum_{t=1}^T \epsilon_t^2} \cdot \sum_{t=1}^T \epsilon_t^2\right) = (2\pi e\hat{\sigma}^2)^{-\frac{T}{2}}$$

a logaritmickú vierohodnosť

$$\ell^*(\mathbf{y}; \theta, \mathbf{x}_0) = -\frac{T}{2} \log(2\pi e\hat{\sigma}^2) = -\frac{T}{2} \log\left(\frac{2\pi e \sum_{t=1}^T \epsilon_t^2}{T}\right) = -\frac{T}{2} \log\left(\sum_{t=1}^T \epsilon_t^2\right) + c_T,$$

kde $c_T = -\frac{T}{2} \log\left(\frac{2\pi e}{T}\right)$ je konštanta, ktorá nezávisí od θ ani \mathbf{x}_0 . Odhady parametrov metódou maximálnej vierohodnosti potom dostávame minimalizáciou výrazu $\frac{T}{2} \cdot \log\left(\sum_{t=1}^T \epsilon_t^2\right)$. To je ekvivalentné s minimalizáciou $SSE = \sum_{t=1}^T (y_t - \hat{y}_t)^2 = \sum_{t=1}^T e_t^2$, čo sme uvádzali aj v druhej kapitole ako jednu z metód výberu hodnoty parametra α pri jednoduchom a dvojitom exponenciálnom vyrovnávaní. Tento odhad teda má zmysel uvažovať aj v prípade, že chyby nemajú normálne rozdelenie, v takom prípade však nebude maximálne vierohodným odhadom.

3.2.2 Výber modelu

Jedna z metód porovnávania modelov za účelom výberu toho najlepšieho je porovnávanie presnosti predpovedí, ktoré dané modely poskytujú. Za mieru presnosti predpovede pritom môžeme vziať súčet štvorcových chýb $SSE = \sum_{t=1}^T (y_t - \hat{y}_t)^2 = \sum_{t=1}^T e_t^2$, strednú absolútnu chybu $MAE = \sum_{t=1}^T |e_t|$, alebo mnohé iné, uvedené napríklad v (Cipra (2008), strany 243 – 247).

Ďalšou možnosťou je minimalizácia hodnoty Akaikeho informačného kritéria modelu, ďalej AIC. Predpokladajme, že máme štatistický model tvaru (3.3), (3.4), a aj ďalej označujme θ vektor jeho parametrov, ktorými sú vyrovnávacie konštanty a prípadne tlmiaci faktor. Označme počet parametrov modelu, teda súčet dĺžok vektora θ a vektora \mathbf{x}_0 , ako k . Potom hodnota Akaikeho informačného kritéria modelu je

$$AIC = 2k - 2 \log \mathcal{L}(\mathbf{y}; \hat{\theta}, \hat{\mathbf{x}}_0), \quad (3.12)$$

kde $\mathcal{L}(\mathbf{y}; \theta, \mathbf{x}_0)$ je vierohodnostná funkcia modelu, $\mathbf{y} = (y_1, \dots, y_T)'$ a $\hat{\theta}, \hat{\mathbf{x}}_0$ sú odhady vektora parametrov a vektora počiatočných stavov vypočítané metódou maximálnej vierohodnosti. Za najvhodnejší spomedzi porovnávaných modelov považujeme ten, ktorý má najnižšiu hodnotu AIC.

Tento prístup k výberu najvhodnejšieho modelu pomocou členu $2k$ penalizuje modely, pri ktorých musíme odhadovať väčšie množstvo parametrov. Ak však model s veľkým počtom parametrov dobre vystihuje dáta, hodnota jeho logaritmickkej vierohodnosti bude vysoká, čím sa hodnota Akaikeho kritéria zase zníži.

Ako sme ukázali v časti venovanej odhadu parametrov modelu, v prípade že pre chyby modelu uvažujeme normálne rozdelenie, maximalizácia vierohodnosti je ekvivalentná s minimalizáciou súčtu štvorcových chýb SSE . Potom môžeme povedať, že pre normálne rozdelené chyby je aj minimalizácia Akaikeho informačného kritéria len určitou modifikáciou prístupu minimalizácie SSE . Má však tú výhodu, že zohľadňuje počet parametrov, ktoré musíme pre daný model odhadnúť.

4. Pravdepodobnostné rozdelenie predpovedí a predikčné intervaly

V druhej kapitole tejto práce, zaoberajúcej sa popisom jednotlivých metód exponenciálneho vyrovnávania, sme uviedli algoritmy pre výpočet bodových predpovedí daných metód. Vzťahy pre predpovede zvyšných metód sa dajú nájsť napríklad v (Hyndman a Athanasopoulos (2018), tabuľka 7.7). V praxi však môže byť užitočné poznať aj pravdepodobnostné rozdelenie predpovedí a predikčné intervaly pre jednotlivé modely popísané v kapitole 3. V tejto časti práce sa budeme venovať tomu, ako ich vypočítať pre šesť aditívnych modelov exponenciálneho vyrovnávania. Budeme pri tom vychádzať primárne z (Hyndman a kol., 2008) a (Bergmeir a kol., 2016).

Predpokladajme, že poznáme model a jeho parametre, vrátane počiatočného stavového vektora \mathbf{x}_0 . Potom poznáme aj vektor \mathbf{x}_T , teda hodnoty stavov na konci pozorovaného obdobia, keďže platí

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{g}\epsilon_t,$$

kde ϵ_t môžeme počítať ako

$$\epsilon_t = y_t - \mathbf{w}'\mathbf{x}_{t-1}.$$

Potom môžeme hľadať podmienené pravdepodobnostné rozdelenie budúcej hodnoty daného časového radu $y_{T+\tau}$ za podmienky \mathbf{x}_T . Strednú hodnotu tohto rozdelenia budeme označovať $\mu_{T+\tau|T} = \mathbf{E}(y_{T+\tau}|\mathbf{x}_T)$. Príslušný rozptyl označujeme $v_{T+\tau|T} = \text{var}(y_{T+\tau}|\mathbf{x}_T)$.

Pravdepodobnostné rozdelenie predpovedí môžeme získať rôznymi spôsobmi. Najčastejším prístupom, ktorý sa dá aplikovať pre ľubovoľný model časového radu, je simulácia veľkého množstva možných budúcich trajektórií z odhadnutého modelu a následný odhad rozdelenia z nasimulovaných hodnôt. Ide o vcelku jednoduchý prístup, pomocou ktorého však dostaneme len numerické predikčné intervaly, a teda ho nemôžeme použiť na algebraickú analýzu rozdelenia predpovedí.

Alternatívou je analytické odvodenie tohto rozdelenia a predikčných intervalov. Pre čisto aditívne modely, ktorými sa v tejto práci zaoberáme, ide o relatívne jednoduchý výpočet, avšak pre niektoré z modelov exponenciálneho vyrovnávania neexistuje analytické vyjadrenie pre strednú hodnotu a rozptyl predpovedí. Ide napríklad o modely s multiplikatívnou chybovou zložkou, multiplikatívnym trendom a žiadnou alebo multiplikatívnou sezónnosťou (Hyndman a kol., 2008).

Za nevýhodu tohto prístupu však možno považovať fakt, že pri odvodení predikčných intervalov sa silne opiera o predpoklad normality chybových zložiek. Nemusí teda dobre fungovať v prípade, že chyby v analyzovanom časovom rade nie sú normálne rozdelené. Ako ukážeme v príkladoch v piatej kapitole, simulačné metódy s využitím bootstrapu sa dokážu zaoberať aj bez splnenia tohto predpokladu.

4.1 Analytické odvodenie

V tejto časti odvodíme algebraické vzorce pre podmienenú strednú hodnotu, podmienený rozptyl a predikčné intervaly pre budúce hodnoty. Strednú hodnotu a

rozptyl vyjadríme pomocou vektorov \mathbf{w} , \mathbf{g} a matice \mathbf{F} vystupujúcich v rovniciach

$$y_t = \mathbf{w}'\mathbf{x}_{t-1} + \epsilon_t, \quad (4.1)$$

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{g}\epsilon_t, \quad (4.2)$$

popisujúcich lineárny model exponenciálneho vyrovnávania. Pritom uvažujeme, že ϵ_t je realizáciou gaussovského bieleho šumu. Tvary \mathbf{w} , \mathbf{g} a \mathbf{F} pre konkrétne modely sú v tabulke 3.1 na strane 22.

4.1.1 Stredná hodnota rozdelenia predpovede

Nech $m_{T+\tau|T} = \mathbf{E}(\mathbf{x}_{T+\tau}|\mathbf{x}_T)$. Potom $m_{T|T} = \mathbf{x}_T$ a pre $\tau > 0$ platí

$$\begin{aligned} m_{T+\tau|T} &= \mathbf{E}(\mathbf{x}_{T+\tau}|\mathbf{x}_T) = \\ &= \mathbf{E}(\mathbf{F}\mathbf{x}_{T+\tau-1} + \mathbf{g}\epsilon_{T+\tau}|\mathbf{x}_T) = \\ &= \mathbf{F}\mathbf{E}(\mathbf{x}_{T+\tau-1}|\mathbf{x}_T) = \\ &= \mathbf{F}m_{T+\tau-1|T} = \mathbf{F}^2m_{T+\tau-2|T} = \dots = \mathbf{F}^\tau m_{T|T} = \mathbf{F}^\tau \mathbf{x}_T, \end{aligned}$$

kde tretia rovnosť platí, keďže predpokladáme, že chyby ϵ_t sú realizáciou gaussovského bieleho šumu a nezávisia na hodnotách \mathbf{x}_s pre $t > s$, teda $\mathbf{E}(\epsilon_{T+\tau}|\mathbf{x}_T) = \mathbf{E}(\epsilon_{T+\tau}) = 0$ pre $\tau \geq 1$. Z toho ďalej máme

$$\begin{aligned} \mu_{T+\tau|T} &= \mathbf{E}(y_{T+\tau}|\mathbf{x}_T) = \\ &= \mathbf{E}(\mathbf{w}'\mathbf{x}_{T+\tau-1} + \epsilon_{T+\tau}|\mathbf{x}_T) = \\ &= \mathbf{w}'m_{T+\tau-1|T} = \\ &= \mathbf{w}'\mathbf{F}^{\tau-1}\mathbf{x}_T, \end{aligned}$$

kde sme v tretej rovnosti opäť využili predpoklad, že $\mathbf{E}(\epsilon_{T+\tau}|\mathbf{x}_T) = 0$, pre $\tau \geq 1$, a v poslednej rovnosti sme len za $m_{T+\tau-1|T}$ dosadili predchádzajúci výsledok.

4.1.2 Rozptyl rozdelenia predpovede

Označme rozptylovú maticu predpovedaného vektora stavov ako $\mathbf{V}_{T+\tau|T} = \text{Var}(\mathbf{x}_{T+\tau}|\mathbf{x}_T)$. Potom $\mathbf{V}_{T|T} = \mathbf{0}$, kde $\mathbf{0}$ označuje nulovú maticu. Ďalej

$$\begin{aligned} \mathbf{V}_{T+\tau|T} &= \text{Var}(\mathbf{F}\mathbf{x}_{T+\tau-1} + \mathbf{g}\epsilon_{T+\tau}|\mathbf{x}_T) = \\ &= \mathbf{F}\text{Var}(\mathbf{x}_{T+\tau-1}|\mathbf{x}_T)\mathbf{F}' + \mathbf{g}\mathbf{g}'\sigma^2 = \\ &= \mathbf{F}\mathbf{V}_{T+\tau-1|T}\mathbf{F}' + \mathbf{g}\mathbf{g}'\sigma^2 = \\ &= \mathbf{F}(\mathbf{F}\mathbf{V}_{T+\tau-2|T}\mathbf{F}' + \mathbf{g}\mathbf{g}'\sigma^2)\mathbf{F}' + \mathbf{g}\mathbf{g}'\sigma^2 = \\ &= \sigma^2 \sum_{j=0}^{\tau-1} \mathbf{F}^j \mathbf{g}\mathbf{g}' (\mathbf{F}')^j, \end{aligned}$$

kde v prvej rovnosti sme využili nezávislosť \mathbf{x}_T a $\epsilon_{T+\tau}$ pre $\tau \geq 1$, ktorá platí keďže ϵ_t je realizácia gaussovského bieleho šumu s rozptylom σ^2 . Potom pre podmienený rozptyl platí

$$\begin{aligned} v_{T+\tau|T} &= \text{var}(y_{T+\tau}|\mathbf{x}_T) = \\ &= \text{var}(\mathbf{w}'\mathbf{x}_{T+\tau-1} + \epsilon_{T+\tau}|\mathbf{x}_T) = \\ &= \mathbf{w}'\mathbf{V}_{T+\tau-1|T}\mathbf{w} + \sigma^2 = \begin{cases} \sigma^2, & \text{ak } \tau = 1, \\ \sigma^2 \left[1 + \sum_{j=1}^{\tau-1} c_j^2\right], & \text{ak } \tau \geq 2, \end{cases} \end{aligned}$$

kde $c_j = \mathbf{w}'\mathbf{F}^{j-1}\mathbf{g}$. V tretej rovnosti sme znovu využili predpoklad, že ϵ_t sú z gaussovského bieleho šumu s rozptylom σ^2 a teda že \mathbf{x}_T a $\epsilon_{T+\tau}$ sú pre $\tau \geq 1$ nezávislé. Posledná rovnosť platí, keďže pre $\tau = 1$ dostávame

$$\mathbf{w}'\mathbf{V}_{T+\tau-1|T}\mathbf{w} + \sigma^2 = \mathbf{w}'\underbrace{\mathbf{V}_{T|T}}_{=\mathbf{0}}\mathbf{w} + \sigma^2 = \sigma^2,$$

zatiaľ čo pre $\tau \geq 2$ máme

$$\begin{aligned} \mathbf{w}'\mathbf{V}_{T+\tau-1|T}\mathbf{w} + \sigma^2 &= \sigma^2 \sum_{j=0}^{\tau-2} \mathbf{w}'\mathbf{F}^j \mathbf{g} \mathbf{g}' (\mathbf{F}')^j \mathbf{w} + \sigma^2 = \\ &= \sigma^2 \left[1 + \sum_{j=0}^{\tau-2} \mathbf{w}'\mathbf{F}^j \mathbf{g} \mathbf{g}' (\mathbf{F}')^j \mathbf{w} \right] = \\ &= \sigma^2 \left[1 + \sum_{j=1}^{\tau-1} (\mathbf{w}'\mathbf{F}^{j-1}\mathbf{g})^2 \right] = \sigma^2 \left[1 + \sum_{j=1}^{\tau-1} c_j^2 \right]. \end{aligned}$$

Poznámka 4.1.1. Vzorce pre strednú hodnotu aj rozptyl sme odvodili za predpokladu, že ϵ_t je realizácia gaussovského bieleho šumu. Výsledky sú však rovnaké pre ľubovoľné rozdelenie ϵ_t , stačí aby bol splnený predpoklad ich nezávislosti, nulovej strednej hodnoty a konštantného rozptylu.

4.1.3 Predikčné intervaly

V aditívnych lineárnych modeloch, kde predpokladáme $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, bude aj minulosťou podmienené rozdelenie predpovedí normálne so strednou hodnotou $\mu_{T+\tau|T}$ a rozptylom $v_{T+\tau|T}$. Potom môžeme $(1 - \alpha) \cdot 100$ -percentné predikčné intervaly počítat klasicky s využitím strednej hodnoty a rozptylu. Ich hranice sú teda

$$\mu_{T+\tau|T} \mp u_{1-\alpha/2} \sqrt{\hat{v}_{T+\tau|T}},$$

kde u_q je q -ty kvantil štandardného normálneho rozdelenia $\mathcal{N}(0,1)$. Vo vyjadrení rozptylu $v_{T+\tau|T}$ vystupuje rozptyl chybových členov σ^2 , za ktorý pri výpočte dosadíme odhad metódou maximálnej vierohodnosti

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \hat{\epsilon}_t^2,$$

ktorý sme odvodili v tretej kapitole, pričom $\hat{\epsilon}_t = y_t - \hat{\mu}_{t|t-1} = y_t - \hat{\mathbf{w}}'\hat{\mathbf{F}}\mathbf{x}_{t-1}$. Vyjadrenie $v_{T+\tau|T}$ obsahuje tiež členy $c_j = \mathbf{w}'\mathbf{F}^{j-1}\mathbf{g}$, za ktoré dosádzame odhady $\hat{c}_j = \hat{\mathbf{w}}'\hat{\mathbf{F}}^{j-1}\hat{\mathbf{g}}$, pričom odhady parametrov vystupujúcich v $\hat{\mathbf{w}}$, $\hat{\mathbf{g}}$ a $\hat{\mathbf{F}}$ dostávame taktiež metódou maximálnej vierohodnosti popísanej v tretej kapitole.

Príklad 4.1.1. Ilustrujme analytický výpočet na modeli ETS(A,N,N), ktorý predstavuje metódu jednoduchého exponenciálneho vyrovnávania. Podľa tabuľky 3.1 pre tento model máme

$$\mathbf{x}_t = L_t, \quad \mathbf{w} = 1, \quad \mathbf{g} = \alpha \quad \text{a} \quad \mathbf{F} = 1.$$

Rovnice (4.1),(4.2) teda majú tvar

$$\begin{aligned}y_t &= L_{t-1} + \epsilon_t, \\L_t &= L_{t-1} + \alpha\epsilon_t,\end{aligned}$$

a môžeme počítat

$$\begin{aligned}\mu_{T+\tau|T} &= \mathbf{E}(y_{T+\tau}|L_T) = \mathbf{E}(L_{T+\tau-1} + \epsilon_{T+\tau}|L_T) = \\&= \mathbf{E}(L_{T+\tau-1}|L_T) + \underbrace{\mathbf{E}(\epsilon_{T+\tau}|L_T)}_{=0} = \\&= \mathbf{E}(L_{T+\tau-2} + \alpha\epsilon_{T+\tau-1}|L_T) = \dots = \\&= \mathbf{E}(L_T + \alpha\epsilon_{T+1}|L_T) = \mathbf{E}(L_T|L_T) = L_T, \\v_{T+\tau|T} &= \text{var}(y_{T+\tau}|L_T) = \text{var}(L_{T+\tau-1} + \epsilon_{T+\tau}|L_T) = \\&= \text{var}(L_{T+\tau-1}|L_T) + \sigma^2 = \\&= \begin{cases} \sigma^2, & \text{ak } \tau = 1, \\ \text{var}(L_{T+\tau-2} + \alpha\epsilon_{T+\tau-1}|L_T) + \sigma^2, & \text{ak } \tau \geq 2. \end{cases}\end{aligned}$$

Potom pre $\tau \geq 2$ dostávame

$$\begin{aligned}v_{T+\tau|T} &= \text{var}(L_{T+\tau-2} + \alpha\epsilon_{T+\tau-1}|L_T) + \sigma^2 = \\&= \text{var}(L_{T+\tau-2}|L_T) + \alpha^2\sigma^2 + \sigma^2 = \\&= \text{var}(L_{T+\tau-2}|L_T) + (1 + \alpha^2)\sigma^2 = \\&= \begin{cases} (1 + \alpha^2)\sigma^2, & \text{ak } \tau = 2, \\ \text{var}(L_{T+\tau-3}|L_T) + (1 + \alpha^2 + \alpha^2)\sigma^2, & \text{ak } \tau \geq 3, \end{cases} \\&\vdots \\v_{T+\tau|T} &= \begin{cases} \sigma^2, & \text{ak } \tau = 1, \\ (1 + (\tau - 1)\alpha^2)\sigma^2, & \text{ak } \tau \geq 2. \end{cases}\end{aligned}$$

Dosadením do odvodených vzorcov by sme dostali $\mu_{T+\tau|T} = 1 \cdot 1^{\tau-1}L_T = L_T$ a

$$v_{T+\tau|T} = \begin{cases} \sigma^2, & \text{ak } \tau = 1, \\ \sigma^2 \left(1 + \sum_{j=1}^{\tau-1} (1 \cdot 1^{j-1}\alpha^2)\right) = \sigma^2 (1 + (\tau - 1)\alpha^2), & \text{ak } \tau \geq 2, \end{cases}$$

teda zhodné výsledky.

Hranice $(1-p) \cdot 100$ -percentného predikčného intervalu pre predpovede modelu ETS(A,N,N) potom dostávame ako

$$L_T \mp u_{1-p/2} \sqrt{(1 + (\tau - 1)\hat{\alpha}^2) \hat{\sigma}^2},$$

kde $u_{1-p/2}$ je $(1 - \frac{p}{2})$ -kvantil štandardného normálneho rozdelenia.

Poznámka 4.1.2. V (Cipra, 2008) má predikčný interval pre jednoduché exponenciálne vyrovňovanie tvar

$$\left(\hat{y}_{T+\tau}(t) \mp u_{1-p/2} \cdot d_\tau \cdot MAE\right),$$

kde

d_τ je konštanta vo výške $\sqrt{\frac{\pi}{2}}$,

MAE je stredná absolútna chyba, t.j. $MAE = \frac{1}{T} \sum_{t=1}^T |y_t - \hat{y}_t(t-1)|$.

Keďže pre normálne rozdelenú náhodnú veličinu $X \sim \mathcal{N}(\mu, \sigma^2)$ platí

$$E|X - \mu| = \sqrt{\frac{2}{\pi}}\sigma,$$

tak namiesto $\sqrt{\frac{1}{n} \sum_{i=1}^T (X_i - \bar{X})^2}$ môžeme za odhad σ zobrať $\sqrt{\frac{\pi}{2}} \cdot \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|$, pre X_i , ktoré sú nezávislé a rovnako rozdelené, $X_i \sim \mathcal{N}(\mu, \sigma)$. Analogicky teda namiesto odhadu $\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \hat{\epsilon}_t^2$, ktorý používame vo vyjadrení predikčného intervalu v rámci tejto práce, môžeme použiť odhad prostredníctvom MAE . Predpokmeď $\hat{y}_{T+\tau}(t)$ je v tomto prípade $\hat{y}_{T+\tau}(t) = L_T$. Intervaly by teda boli ekvivalentné, avšak ten nami odvodený v príklade 4.1.1 obsahuje navyše člen $\sqrt{(1 + (\tau - 1) \hat{\alpha}^2)}$. V intervale podľa (Cipra, 2008) naopak nevystupuje žiaden člen, ktorý by závisel na τ , a teda potom pomocou daného vzorca dostávame predikčné intervaly, ktoré sú rovnako široké pre ľubovoľnú predpovednú dobu. Prirodzene však očakávame, že s väčším τ by sa mal predikčný interval rozširovať, nakoľko so vzdialenejšou budúcnosťou sa spája väčšia neistota.

4.2 Simulované rozdelenie predpovedí a predikčné intervaly

Znovu pripomeňme, že lineárny model exponenciálneho vyrovnávania má tvar (4.1), (4.2), kde ϵ_t sú realizáciou gaussovského bieleho šumu.

Predpokladajme, že máme k dispozícii odhady parametrov tohto modelu, počiatočný stavový vektor \mathbf{x}_0 , pozorovania $\{y_t\}$, $t = 1, 2, \dots, T$, a teda aj stavový vektor \mathbf{x}_T . Ak chceme simulovať možné budúce trajektórie, znamená to, že chceme vygenerovať pozorovania $\{y_t^{(i)}\}$, $t = T+1, T+2, \dots, T+\tau$. Ako počiatočný vektor stavov v tomto prípade použijeme \mathbf{x}_T z odhadnutého modelu a za hodnoty ϵ_t pre $t = T+1, T+2, \dots, T+\tau$ vezmeme náhodne vygenerované hodnoty z normálneho rozdelenia $\mathcal{N}(0, \hat{\sigma}^2)$, kde $\hat{\sigma}^2$ je odhad rozptylu chybových členov modelu. Toto zopakujeme $i = 1, 2, \dots, I$ krát, kde I je dost veľké celé číslo (napríklad 5000). Charakteristiky pravdepodobnostného rozdelenia predpovede $y_{T+\tau|T}$ potom odhadujeme zo simulovaných hodnôt $\{y_{T+\tau}^{(1)}, \dots, y_{T+\tau}^{(I)}\}$.

Predikčné intervaly môžeme odhadnúť pomocou výberových kvantilov nasi-mulovaných hodnôt časového radu. Hranice $(1 - \alpha) \cdot 100$ -percentného intervalu pre predikčný horizont τ sú potom $\frac{\alpha}{2}$ a $(1 - \frac{\alpha}{2})$ výberové kvantily hodnôt $y_{T+\tau|T}$.

Túto procedúru môžeme prípadne upraviť, ak predpokladáme pre ϵ_t iné ako normálne rozdelenie, generovaním hodnôt z tohto rozdelenia. V simulačnej štúdií v piatej kapitole uvažujeme aj modely s chybami z t -rozdelenia alebo z posunutého exponenciálneho rozdelenia, pričom v simuláciách budúcich trajektórií generujeme hodnoty ϵ_t , $t = T+1, \dots, T+\tau$, z týchto rozdelení s parametrami odhadnutými z dát.

Taktiež môžeme postupovať bez predpokladu o pravdepodobnostnom rozdelení chýb ϵ_t a použiť historické hodnoty a bootstrap metódu. V tom prípade si vypočítame historické chyby $e_t, t = 1, \dots, T$, ako

$$e_t = y_t - \hat{\mathbf{w}}' \mathbf{x}_{t-1}$$

a z týchto T hodnôt vyberáme náhodne s opakovaním čísla, ktoré použijeme vo vzorcoch (4.1), (4.2) ako budúce hodnoty ϵ_t pre $t = T + 1, \dots, T + \tau$.

V (Hyndman a Athanasopoulos, 2018) sa uvádza, že takmer všetky predikčné intervaly modelov časových radov sú príliš úzke. Dôvodom môže byť, že existujú minimálne štyri zdroje neistoty v predpovedaní s modelmi časových radov, avšak klasicky sa zohľadňuje iba neistota vyplývajúca z náhodnosti chybovej zložky. Zvyšné zdroje neistoty vyplývajú z toho, že odhadujeme parametre modelu, volíme model pre historické dáta a predpokladáme, že sa týmto modelom budú riadiť aj budúce pozorovania. Ak chceme zohľadniť aspoň neistotu z odhadu parametrov, s výnimkou niekoľkých špeciálnych prípadov, neexistujú algebraické vzorce pre výpočet predpovedí (Hyndman a kol., 2008). To však do určitej miery rieši špeciálny typ bootstrapu, v angličtine nazývaný moving block bootstrap (ďalej MBB).

Použitím MBB môžeme vygenerovať nový časový rad, ktorý bude podobný tomu pozorovanému. Ak chceme aplikovať MBB metódu, najskôr časový rad transformujeme Boxovou-Coxovou transformáciou definovanou ako

$$y_t^\lambda = \begin{cases} \log(y_t), & \text{ak } \lambda = 0, \\ \frac{y_t^\lambda - 1}{\lambda}, & \text{ak } \lambda \neq 0, \end{cases}$$

kde parameter λ odhadujeme maximalizáciou profilovej vierohodnosti, postupom poísaným napríklad v (Zvára, 2008). Potom časový rad rozložíme na trendovú, sezónnu a chybovú zložku využitím STL metódy pôvodne navrhnutej v (Cleveland a kol., 1990).

Skratka STL je z anglického Seasonal and Trend decomposition using Loess. Samotná metóda pozostáva zo sérií vyrovnávacích operácií s využitím lokálne váženej regrese – loess. Zjednodušene povedané, využitím tejto metódy najskôr určíme približnú trendovú zložku pre analyzovaný časový rad a odčítame ju od neho, aby sme dostali dáta bez trendu. Z tých potom vytvoríme časové rady pre každú sezónu zvlášť a vyrovnáme ich. Takto vyrovnané čiastkové rady dáme znovu dokopy a vzniknutý rad ešte raz vyrovnáme. Opäť odstránime trend, rad sezónne očistíme pomocou faktorov z predchádzajúcich krokov a očistený rad vyrovnáme, aby sme dostali jeho trendovú zložku. Odčítaním takto získanej trendovej a sezónnej zložky získame aj rad chýb.

Rad chýb si rozdelíme na bloky dopredu stanovenej dĺžky (napríklad v (Bergmeir a kol., 2016) používajú dĺžku bloku aspoň dva roky, aby zachytili akúkoľvek zostávajúcu sezónnosť), z ktorých potom náhodne vyberáme a vyskladáme nový rad. Následne k nemu naspäť pridáme trendovú a sezónnu zložku, invertujeme Boxovu-Coxovu transformáciu a tak dostávame nový časový rad.

Ak takto vygenerujeme I nových časových radov, kde I je dost veľké celé číslo, pre každý odhadneme vhodný model exponenciálneho vyrovnávania a následne simulujeme jednu budúcu trajektóriu, opäť dostaneme simulované hodnoty $\{y_{T+\tau}^{(1)}, \dots, y_{T+\tau}^{(I)}\}$, z ktorých môžeme odhadovať charakteristiky pravdepodobnostného rozdelenia v budúcom čase $T + \tau$.

Pri výbere modelu zvyčajne dostaneme rovnaký ETS model, avšak jeho parametre budú v jednotlivých prípadoch odhadnuté inak. Preto takto zohľadňujeme aj neistotu vyplývajúcu z ich odhadu aj z výberu modelu pre historické dáta, nielen tú vyplývajúcu z náhodnosti chybových zložiek.

Poznámka 4.2.1. Tento prístup nám navyše umožňuje počítať aj presnejšie bodové predpovede (Hyndman a Athanasopoulos, 2018). Ak máme záujem zlepšiť len bodovú predpoveď a nie predikčný interval, môžeme vypočítať bodovú predpoveď z jednotlivých modelov pre každý vygenerovaný rad a spočítať priemer týchto hodnôt. Takto dostávame presnejšiu predpoveď, ako tú z originálneho časového radu a je to rýchlejšie a jednoduchšie, ako výpočet zo simulácií. V (Bergmeir a kol., 2016) autori túto metódu odporúčajú najmä pre dáta s mesačnou sezónnosťou.

5. Praktická časť

V tejto časti našej práce sa budeme venovať aplikácii postupov popísaných v štvrtej kapitole na simulované aj reálne dáta. Využijeme pri tom softvér R (R Core Team, 2019), konkrétne funkciu *ets()*, ktorá je súčasťou balíčku *forecast* (Hyndman a Khandakar, 2008). Táto funkcia, ak neurčíme inak, automaticky vyberie najvhodnejší model exponenciálneho vyrovnávania podľa Akaikeho informačného kritéria (3.12). Môžeme však aj sami určiť model, ktorý chceme, aby táto funkcia odhadla. V tom prípade zadáme v jej špecifikácii kombináciu ETS typu chybovej zložky, trendu a sezónnosti (model = "ETS"), prípadne či chceme pracovať s tlmeným trendom (damped = TRUE) alebo nie.

Funkcia *ets()* ďalej odhadne parametre a počiatočné stavy modelu metódou maximálnej vierohodnosti, popísanou v časti 3.2.1 na strane 21, a vypočíta vyrovnané hodnoty. Pomocou funkcie *forecast.ets()* potom môžeme získať aj bodovú predpoveď vybraného modelu pre daný časový rad a jej predikčný interval spočítaný na základe analytických vzorcov uvedených v časti 4.1 v predchádzajúcej kapitole.

5.1 Simulované dáta

Najskôr budeme pracovať so simulovanými dátami. Tie budeme generovať vždy z rovnakého modelu

$$y_t = \log(t) + \epsilon_t,$$

kde však budeme uvažovať rôzne rozdelenie chybových členov ϵ_t . Konkrétne budeme postupne pracovať s normálne rozdelenými chybami, chybami s t -rozdelením a s chybami z posunutého exponenciálneho rozdelenia.

Najväčšiu pozornosť budeme venovať pravdepodobnostnému rozdeleniu budúcich hodnôt takéhoto časového radu, ktorého charakteristiky budeme odhadovať zo simulácií budúceho vývoja. Výsledky z nich porovnáme s hodnotami z analytických vzorcov z časti 4.1 a tiež s teoretickými hodnotami odvodenými na základe znalosti rozdelenia členov ϵ_t .

V každom z príkladov detailne popíšeme analýzu pre jednu konkrétnu realizáciu daného časového radu, výsledky jednotlivých metód však budeme porovnávať na základe výstupov zo sto opakovaní ich aplikácie.

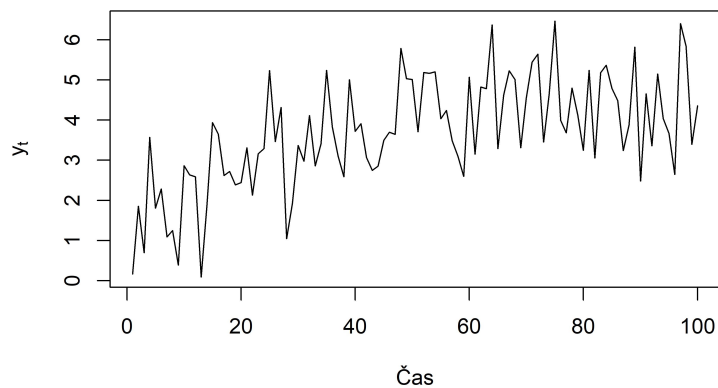
5.1.1 Normálne rozdelenie chybových členov

Predpokladajme, že máme 100 pozorovaní časového radu, ktorý je generovaný ako

$$y_t = \log(t) + \epsilon_t, \quad t = 1, 2, \dots, 100,$$

kde $\epsilon_t \sim \mathcal{N}(0,1)$. Jedna z možných realizácií je na obrázku 5.1.

Spomínaná funkcia *ets()*, ktorú sme nechali automaticky identifikovať model pre tieto simulované dáta, určila ako najlepší model ETS(A, A_d, N), teda model s aditívnou chybovou zložkou, aditívnym tlmeným trendom a žiadnou sezónnos-



Obr. 5.1: Simulované pozorovania časového radu $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$.

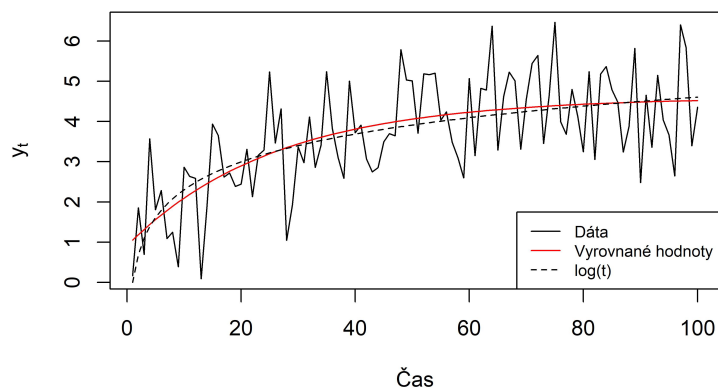
ťou. Pomocou rovníc ho môžeme zapísať nasledovne

$$y_t = (1, \phi) \cdot \begin{pmatrix} L_{t-1} \\ T_{t-1} \end{pmatrix} + \epsilon_t,$$

$$\begin{pmatrix} L_t \\ T_t \end{pmatrix} = \begin{pmatrix} 1 & \phi \\ 0 & \phi \end{pmatrix} \cdot \begin{pmatrix} L_{t-1} \\ T_{t-1} \end{pmatrix} + \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \cdot \epsilon_t.$$

Vyrovnané hodnoty časového radu $\hat{y}_t = L_t$ sú zobrazené na obrázku 5.2. Môžeme si všimnúť, že krivka zobrazujúca vyrovnané hodnoty tvarom naozaj zodpovedá logaritmickému trendu, ktorú sme použili na generovanie pozorovaní y_t .

Vyrovnané hodnoty modelu ETS(A,Ad,N)



Obr. 5.2: Pozorovania časového radu $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$, a vyrovnané hodnoty odhadnutého modelu ETS(A,Ad,N).

Odhadnuté hodnoty parametrov α a β a odhad hodnoty tlmiaceho parametra ϕ sú uvedené v tabuľke 5.1. Rovnako v nej môžeme vidieť aj softvérom odhadnutý

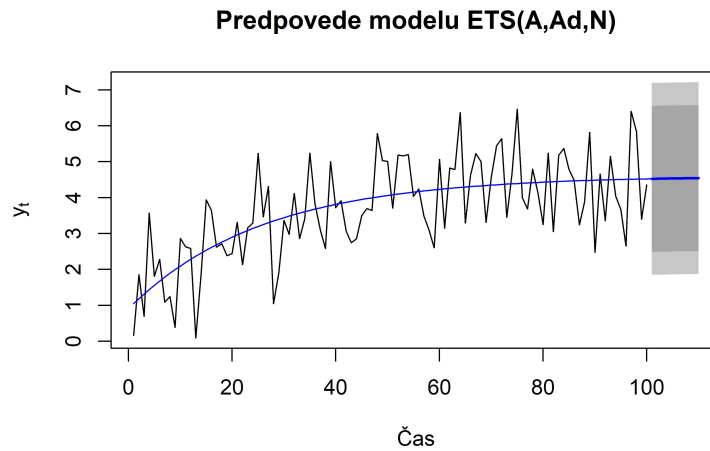
počiatočný stavový vektor $\mathbf{x}_0 = (L_0, T_0)'$ a odhad $\hat{\sigma}^2$ rozptylu chybových členov. Všetky odhady boli vypočítané metódou maximálnej vierohodnosti.

Parametre	Počiatočné stavy	Rozptyl
$\hat{\alpha}$ 0.0001	L_0 0.9103	$\hat{\sigma}^2$ 1.0752
$\hat{\beta}$ 0.0001	T_0 0.1456	
$\hat{\phi}$ 0.9620		

Tabuľka 5.1: Odhady parametrov modelu ETS(A, A_d, N) pre časový rad $y_t = \log(t) + \epsilon_t, \epsilon_t \sim \mathcal{N}(0,1)$.

Parametre α a β boli softvérom odhadnuté ako hodnoty blízke nule. Vyrovnané hodnoty $\hat{y}_t, t = 1, \dots, 100$, ktoré zodpovedajú úrovni L_t v čase t , sú potom takmer identické s hodnotou $L_{t-1} + \phi T_{t-1}$, teda s úrovňou pre čas t odhadnutou v čase $t-1$. Smernica T_t v čase t je kvôli malej hodnote parametra β približne 96 percent predchádzajúcej smernice z času $t-1$.

Bodové predpovede pre 10 nasledujúcich období a príslušné predikčné intervaly, vypočítané s využitím funkcie `forecast.ets()`, sú na obrázku 5.3. Bledosivý pás predstavuje 99-percentné predikčné intervaly, užší tmavosivý pás zobrazuje 95-percentné predikčné intervaly.



Obr. 5.3: Bodové predpovede modelu ETS(A, A_d, N) o 10 krokov dopredu a predikčné intervaly pre rad $y_t = \log(t) + \epsilon_t, \epsilon_t \sim \mathcal{N}(0,1)$.

Presné hodnoty bodových predpovedí a predikčné intervaly uvádza tabuľka 5.2. Funkcia `forecast.ets()` tieto hodnoty, ak neurčíme inak, počíta na základe analytických vzorcov, ktoré boli odvodené v štvrtej kapitole. Porovnaním s tabuľkou 5.3 si môžeme všimnúť, že bodová predpoveď v čase 110 z tabuľky 5.2 je rovná strednej hodnote rozdelenia v tomto čase. Taktiež predikčné intervaly, ktoré sme vypočítali pomocou vzorcov manuálne, sa naozaj zhodujú s výstupmi tejto funkcie.

Okrem bodových predpovedí nás však zaujímalo najmä pravdepodobnostné rozdelenie v budúcich časoch podmienené informáciou do času $T = 100$. Kon-

	Bodová predpoveď	95% pred. int.	99% pred. int.
$\hat{y}_{101}(100)$	4.5223	(2.4899, 6.5546)	(1.8513, 7.1933)
$\hat{y}_{102}(100)$	4.5250	(2.4926, 6.5573)	(1.8540, 7.1959)
$\hat{y}_{103}(100)$	4.5275	(2.4952, 6.5599)	(1.8566, 7.1985)
$\hat{y}_{104}(100)$	4.5300	(2.4977, 6.5624)	(1.8590, 7.2010)
$\hat{y}_{105}(100)$	4.5324	(2.5000, 6.5648)	(1.8614, 7.2034)
$\hat{y}_{106}(100)$	4.5347	(2.5023, 6.5670)	(1.8637, 7.2057)
$\hat{y}_{107}(100)$	4.5369	(2.5045, 6.5693)	(1.8659, 7.2079)
$\hat{y}_{108}(100)$	4.5390	(2.5067, 6.5714)	(1.8681, 7.2100)
$\hat{y}_{109}(100)$	4.5411	(2.5087, 6.5734)	(1.8701, 7.2121)
$\hat{y}_{110}(100)$	4.5430	(2.5107, 6.5754)	(1.8721, 7.2140)

Tabuľka 5.2: Hodnoty bodových predpovedí modelu $ETS(A, A_d, N)$ a príslušné predikčné intervaly pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, 1)$.

krátne sme sa zamerali na čas $t = 110$ a charakteristiky jeho pravdepodobnostného rozdelenia sme odhadovali zo simulácií spôsobom popísaným v nasledujúcich častiach.

Z výstupu funkcie $ets()$ máme k dispozícii stavový vektor na konci pozorovaného obdobia

$$\mathbf{x}_{100} = (L_{100}, T_{100}),$$

a odhady všetkých parametrov, ktoré sa nachádzajú vo vektoroch \mathbf{w} , \mathbf{g}

$$\hat{\mathbf{w}} = (1, \hat{\phi})', \quad \hat{\mathbf{g}}' = (\hat{\alpha}, \hat{\beta}),$$

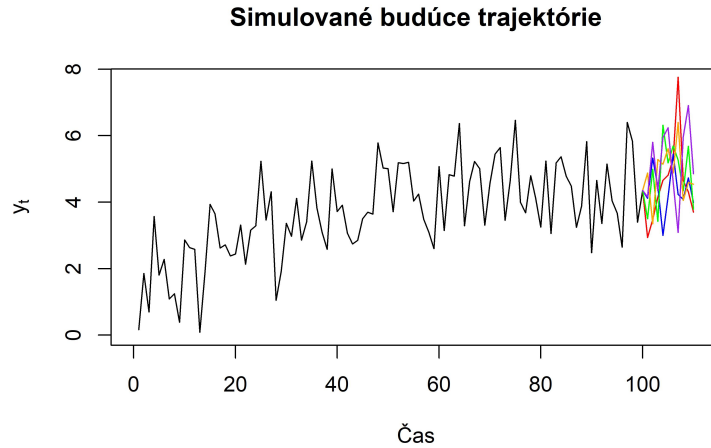
a v matici \mathbf{F}

$$\hat{\mathbf{F}} = \begin{pmatrix} 1 & \hat{\phi} \\ 0 & \hat{\phi} \end{pmatrix}.$$

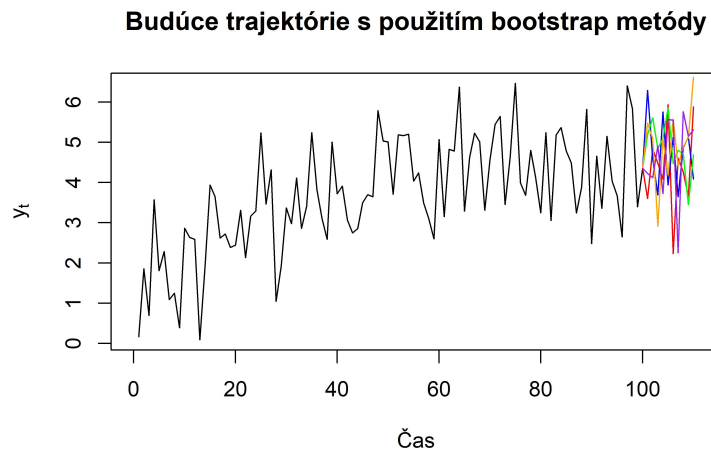
Potom môžeme podľa vzorca (4.1) vypočítať hodnotu y_{101} a s využitím (4.2) prepočítať novú hodnotu \mathbf{x}_{101} , ktorú následne použijeme vo výpočte pre y_{102} . Takto sme počítali 10 budúcich hodnôt, pričom hodnoty ϵ_t , pre $t = 101, \dots, 110$, ktoré v týchto rovniciach vystupujú, sme generovali z normálneho rozdelenia $\mathcal{N}(0, \hat{\sigma}^2)$, kde $\hat{\sigma}^2$ je odhad rozptylu chybových členov. Celkovo sme týmto spôsobom nasimulovali 5000 možných budúcich trajektórií. Na obrázku 5.4 môžeme vidieť 5 z nich.

Alternatívne, bez predpokladu o pravdepodobnostnom rozdelení chybových členov, sme simulovali ďalších 5000 trajektórií využitím bootstrap metódy. To znamená, že sme zo známych minulých chýb $e_t = y_t - \hat{\mathbf{w}}' \mathbf{x}_{t-1}$, $t = 1, \dots, 100$, vyberali náhodne s opakovaním hodnoty ϵ_t pre časy $t = 101, \dots, 110$, ktoré sme použili na výpočet hodnôt y_t , \mathbf{x}_t , pre $t = 101, \dots, 110$, podľa (4.1), (4.2). Päť takto získaných trajektórií je na obrázku 5.5.

V oboch prípadoch nás zaujímala stredná hodnota a rozptyl nasimulovaného rozdelenia v čase 110. Tie sme odhadli ako výberový priemer a výberový rozptyl z 5000 hodnôt, ktoré sme mali k dispozícii pre každú simuláciu. Predikčné intervaly pre predpoveď v tomto čase sme zo simulácií odhadovali ako $\left(\frac{\alpha}{2}\right)$ a $\left(1 - \frac{\alpha}{2}\right)$ výberové kvantily nasimulovaných hodnôt, pričom sme uvažovali $\alpha = 0.05$, resp. $\alpha = 0.01$.



Obr. 5.4: Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$, s chybami generovanými z $\mathcal{N}(0, \hat{\sigma}^2)$.



Obr. 5.5: Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$, s využitím metódy bootstrap.

Strednú hodnotu, rozptyl a predikčné intervaly sme pre porovnanie vypočítali tiež s použitím analytických vzorcov odvodených v časti 4.1 v predchádzajúcej kapitole.

Keďže v našom prípade sú hodnoty y_t nezávislé a vieme, že sa riadia modelom $y_t = \log(t) + \epsilon_t$, poznáme aj ich skutočnú strednú hodnotu a rozptyl a môžeme vypočítať aj teoretické intervaly, ktoré budú pokrývať budúcu hodnotu s požadovanou pravdepodobnosťou, ďalej označované ako skutočné predikčné intervaly.

Pre strednú hodnotu a rozptyl v čase 110 platí

$$\begin{aligned}
 \mu_{110|100} &= E[y_{110}|\mathbf{x}_{100}, y_{100}, \dots, y_1] = \\
 &= E[\log(110) + \epsilon_t|\mathbf{x}_{100}, y_{100}, \dots, y_1] = \\
 &= \log(110) + E[\epsilon_t|\mathbf{x}_{100}, y_{100}, \dots, y_1] \\
 &= \log(110) \doteq 4.7005, \\
 \\
 \text{var}_{110|100} &= \text{var}[y_{110}|\mathbf{x}_{100}, y_{100}, \dots, y_1] = \\
 &= \text{var}[\log(110) + \epsilon_t|\mathbf{x}_{100}, y_{100}, \dots, y_1] = \\
 &= \text{var}[\epsilon_t|\mathbf{x}_{100}, y_{100}, \dots, y_1] = 1.
 \end{aligned}$$

Skutočný 95, resp. 99-percentný predikčný interval dostaneme ako $\log(110) \mp u_{0,975}$, resp. $\log(110) \mp u_{0,995}$, kde u_p je p -kvantil štandardného normálneho rozdelenia.

Porovnanie výsledkov oboch simulačných prístupov s analyticky vypočítanými a so skutočnými hodnotami je uvedené v tabuľke 5.3.

	$\mu_{110 100}$	$\text{var}_{110 100}$	95% pred. int.	99% pred. int.
Normálne chyby	4.5403	1.0722	(2.4836, 6.5465)	(1.8579, 7.1549)
Bootstrap	4.5184	1.0304	(2.5366, 6.6149)	(2.2398, 6.6742)
Analytické vzorce	4.5430	1.0752	(2.5107, 6.5754)	(1.8721, 7.2140)
Skutočná hodnota	4.7005	1.0000	(2.7405, 6.6604)	(2.1247, 7.2763)

Tabuľka 5.3: Porovnanie výsledkov pre pravdepodobnostné predpovede radu $y_t = \log(t) + \epsilon_t, \epsilon_t \sim \mathcal{N}(0,1)$ o 10 krokov dopredu.

Môžeme si všimnúť, že v tomto prípade sú odhady strednej hodnoty a rozptylu vypočítané zo simulácií za predpokladu normálneho rozdelenia chybových členov bližšie k tým, ktoré sme dostali analytickým odvodením, ako odhady z bootstrap simulácií. Predikčný interval sme však bootstrapom získali presnejší, dokonca ide o interval, ktorý je najbližšie k tomu skutočnému. Taktiež vidíme, že pomocou analytických vzorcov dostaneme strednú hodnotu rovnú bodovej predpovedi funkcie `forecast.ets()` a jej predikčný interval je totožný s analytickým.

Ak však chceme posúdiť, ktorá metóda poskytuje najlepšie výsledky v porovnaní s teoretickými hodnotami, nestačí sa pozrieť na jednu realizáciu modelu $y_t = \log(t) + \epsilon_t, \epsilon_t \sim \mathcal{N}(0,1)$. Preto sme vygenerovali 100 možných realizácií a pre každú sme odhadli strednú hodnotu, rozptyl a predikčný interval spôsobom, ktorý sme teraz podrobne popísali pre prvú z nich. Dostali sme teda 100 odhadov týchto hodnôt pre každú zo simulačných metód (s chybami generovanými z normálneho rozdelenia $\mathcal{N}(0, \hat{\sigma}^2)$ a s využitím bootstrap metódy) a pre jednotlivé realizácie sme ich vypočítali aj analyticky podľa vzorcov zo štvrtej kapitoly. Následne sme pre každú metódu vypočítali priemery zo získaných sto odhadov strednej hodnoty, rozptylu, hornej hranice 99-percentného predikčného intervalu a jeho šírky a tie sme porovnali s ich skutočnými hodnotami. Toto porovnanie uvádza tabuľka 5.4.

V takomto porovnaní vidíme, že rovnako ako v predchádzajúcej časti, simulácie za predpokladu normálneho rozdelenia chybových členov poskytnú výsledky,

	$\mu_{110 100}$	$\text{var}_{110 100}$	Horná hranica 99% pred. int.	Šírka 99% pred. int.
Normálne chyby	4.5422	1.0893	7.2137	5.3611
Bootstrap	4.6238	1.0291	7.1703	5.0744
Analytické vzorce	4.5447	1.0908	7.2271	5.3649
Skutočná hodnota	4.7005	1.0000	7.2763	5.1517

Tabuľka 5.4: Porovnanie výsledkov zo sto simulácií modelu $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$, pre rôzne metódy odhadov.

ktoré sú bližšie k tým analytickým. S využitím bootstrap metódy však dostaneme presnejší odhad strednej hodnoty, rozptylu aj šírky 99-percentného predikčného intervalu v porovnaní so skutočnými hodnotami.

Tiež si môžeme všimnúť, že použitím všetkých metód nám vyšli vcelku podobné výsledky. Toto však nebude platiť pre ďalšie príklady, kde uvažujeme iné rozdelenie chybových členov. V tabuľkách 5.8 a 5.12 v ďalších častiach môžeme vidieť, že výsledky jednotlivých metód sa najmä pri predikčných intervaloch líšia výraznejšie, ako v tomto prípade.

5.1.2 t-rozdelenie chybových členov

Uvažujme podobný časový rad ako v predchádzajúcom prípade a rovnako predpokladajme, že máme k dispozícii 100 jeho pozorovaní. Tentokrát ho však budeme generovať ako

$$y_t = \log(t) + \epsilon_t, \quad t = 1, 2, \dots, 100,$$

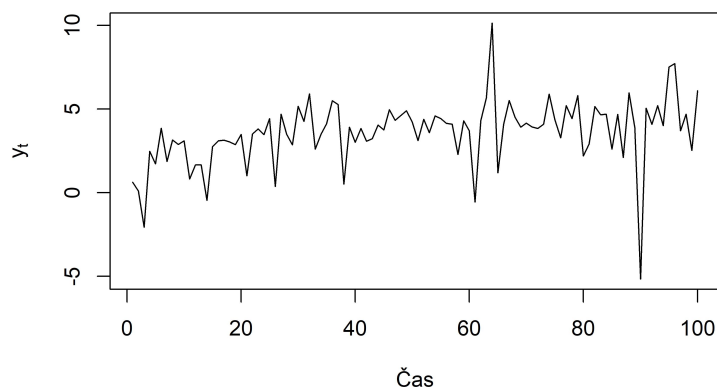
kde ϵ_t má Studentovo t -rozdelenie s tromi stupňami voľnosti, t.j. $\epsilon_t \sim t_3$. Ide o rozdelenie s ťažkými chvostami. Príklad realizácie takéhoto časového radu je na obrázku 5.6. Môžeme si všimnúť, že takto generovaný časový rad má väčšie výkyvy ako ten s chybami generovanými z normálneho rozdelenia. Je to práve v dôsledku ťažkých chvostov t -rozdelenia, kvôli ktorým majú extrémne hodnoty chýb väčšiu pravdepodobnosť výskytu.

Na tento simulovaný časový rad sme znovu aplikovali funkciu $ets()$, ktorú sme nechali samú určiť najvhodnejší model. Opäť vybrala $ETS(A, A_d, N)$, popísaný v predchádzajúcej časti, a odhadla jeho parametre, počiatočné stavy aj rozptyl chybových členov. Odhady sú uvedené v tabuľke 5.5.

Parametre	Počiatočné stavy	Rozptyl
$\hat{\alpha}$ 0.0001	L_0 -0.2979	$\hat{\sigma}^2$ 3.2510
$\hat{\beta}$ 0.0001	T_0 0.5077	
$\hat{\phi}$ 0.8942		

Tabuľka 5.5: Odhady parametrov modelu $ETS(A, A_d, N)$ pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim t_3$.

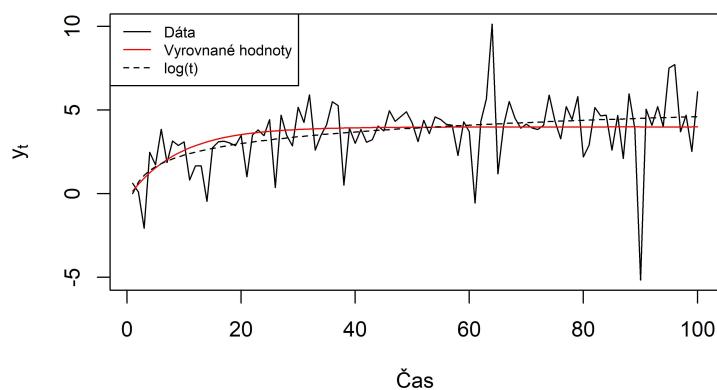
Kvôli malým hodnotám odhadov $\hat{\alpha}, \hat{\beta}$, môžeme znovu povedať, že vyrovnané hodnoty modelu, pre ktoré platí $\hat{y}_t = L_t$, $t = 1, \dots, 100$, takmer zodpovedajú od-



Obr. 5.6: Simulované pozorovania časového radu
 $y_t = \log(t) + \epsilon_t, \epsilon_t \sim t_3$.

hadu úrovne pre čas t z času $t - 1$. Môžeme ich vidieť na obrázku 5.7. Pre porovnanie je na ňom zobrazená tiež logaritmická funkcia, ktorú sme použili na generovanie hodnôt časového radu. Môžeme si všimnúť, že vyrovnané hodnoty ju celkom dobre kopírujú.

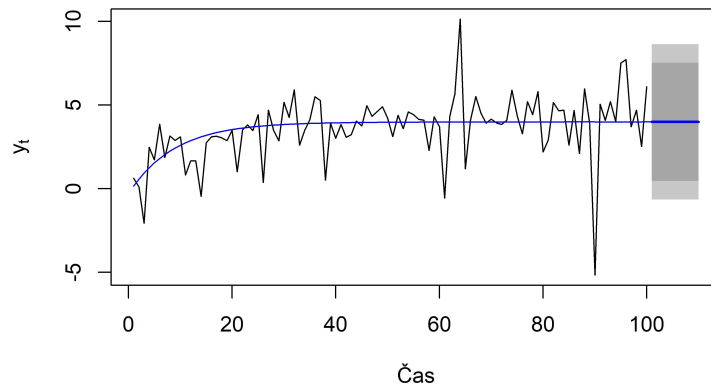
Vyrovnané hodnoty modelu ETS(A,Ad,N)



Obr. 5.7: Pozorovania časového radu $y_t = \log(t) + \epsilon_t, \epsilon_t \sim t_3$,
a vyrovnané hodnoty odhadnutého modelu ETS(A,Ad,N).

Bodové predpovede modelu pre nasledujúcich 10 období sme počítali s využitím funkcie `forecast.ets()`. Okrem modelu a dĺžky predpovedného obdobia sme nešpecifikovali žiadne ďalšie parametre, takže výstupy tejto funkcie budú opäť zhodné s výsledkami, ktoré nám poskytnú analytické vzorce odvodené v štvrtej kapitole. Predpovede spolu s bledosivými 99-percentnými a tmavosivými 95-percentnými predikčnými intervalmi sú na obrázku 5.8. Číselné hodnoty sú v tabuľke 5.6.

Predpovede modelu ETS(A,Ad,N)



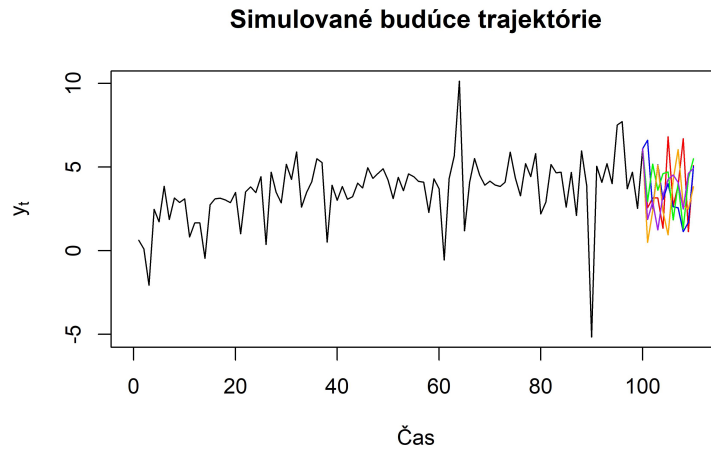
Obr. 5.8: Bodové predpovede modelu ETS(A,Ad,N) o 10 krokov dopredu a predikčné intervaly pre rad $y_t = \log(t) + \epsilon_t, \epsilon_t \sim t_3$.

	Bodová predpoveď	95% pred. int.	99% pred. int.
$\hat{y}_{101}(100)$	3.9860	(0.4520, 7.5199)	(-0.6584, 8.6303)
$\hat{y}_{102}(100)$	3.9863	(0.4523, 7.5202)	(-0.6581, 8.6307)
$\hat{y}_{103}(100)$	3.9866	(0.4526, 7.5205)	(-0.6578, 8.6310)
$\hat{y}_{104}(100)$	3.9868	(0.4529, 7.5208)	(-0.6576, 8.6312)
$\hat{y}_{105}(100)$	3.9871	(0.4531, 7.5210)	(-0.6573, 8.6315)
$\hat{y}_{106}(100)$	3.9873	(0.4533, 7.5212)	(-0.6571, 8.6317)
$\hat{y}_{107}(100)$	3.9875	(0.4535, 7.5214)	(-0.6569, 8.6319)
$\hat{y}_{108}(100)$	3.9876	(0.4537, 7.5216)	(-0.6568, 8.6320)
$\hat{y}_{109}(100)$	3.9878	(0.4538, 7.5217)	(-0.6566, 8.6322)
$\hat{y}_{110}(100)$	3.9879	(0.4540, 7.5219)	(-0.6565, 8.6323)

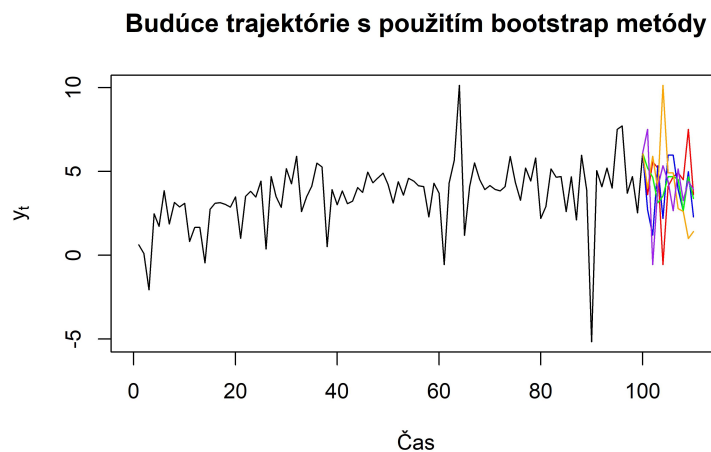
Tabuľka 5.6: Hodnoty bodových predpovedí modelu ETS(A,Ad,N) a príslušné predikčné intervaly pre časový rad $y_t = \log(t) + \epsilon_t, \epsilon_t \sim t_3$.

Aj v tomto prípade nás zaujímalo pravdepodobnostné rozdelenie v čase $t = 110$, podmienené informáciou do času $T = 100$. Jeho charakteristiky a predikčný interval sme opäť odhadovali zo simulácií. Postupovali sme pritom rovnako ako v prvom príklade. Hodnoty $\epsilon_t, t = 101, \dots, 110$, ktoré sme dosádzali do vzorcov (4.1) a (4.2), sme teda znovu generovali z normálneho rozdelenia $\mathcal{N}(0, \hat{\sigma}^2)$, alebo sme ich náhodne vyberali zo známych minulých chýb $e_t = y_t - \hat{\mathbf{w}}' \mathbf{x}_{t-1}, t = 1, \dots, 100$. V oboch prípadoch sme urobili 5000 simulácií, z ktorých 5 môžeme vidieť na obrázkoch 5.9 a 5.10.

Následne sme z 5000 simulovaných hodnôt v čase 110 odhadovali podmiennú strednú hodnotu $\mu_{110|100}$, podmienený rozptyl $\text{var}_{110|100}$ a hranice 95 a 99-percentného predikčného intervalu. Tie sme dostali ako 0.025, resp. 0.005, a 0.975, resp. 0.995 výberové kvantily 5000 nasimulovaných hodnôt v čase 110. Výsledky sú zhrnuté v tabuľke 5.7. Okrem odhadov zo simulácií tabuľka 5.7 obsahuje aj hodnoty, ktoré sme dostali manuálnym výpočtom z analytických vzorcov, a sku-



Obr. 5.9: Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim t_3$, s chybami generovanými z $\mathcal{N}(0, \hat{\sigma}^2)$.



Obr. 5.10: Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim t_3$, s využitím metódy bootstrap.

točné hodnoty, ktoré vieme vypočítať, keďže v uvažovanom príklade sú hodnoty y_t nezávislé a riadia sa známym modelom $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim t_3$.

Pre skutočnú strednú hodnotu a rozptyl platí

$$\begin{aligned}
 \mu_{110|100} &= E[y_{110} | \mathbf{x}_{100}, y_{100}, \dots, y_1] = \\
 &= E[\log(110) + \epsilon_t | \mathbf{x}_{100}, y_{100}, \dots, y_1] = \\
 &= \log(110) + E[\epsilon_t | \mathbf{x}_{100}, y_{100}, \dots, y_1] \\
 &= \log(110) \doteq 4.7005,
 \end{aligned}$$

$$\begin{aligned}
\text{var}_{110|100} &= \text{var} [y_{110} | \mathbf{x}_{100}, y_{100}, \dots, y_1] = \\
&= \text{var} [\log(110) + \epsilon_t | \mathbf{x}_{100}, y_{100}, \dots, y_1] = \\
&= \text{var} [\epsilon_t | \mathbf{x}_{100}, y_{100}, \dots, y_1] = \\
&= \frac{3}{3-2} = 3,
\end{aligned}$$

skutočný 95-percentný predikčný interval dostaneme zo vzťahu

$$y_{110} = \log(110) + \epsilon_{110},$$

kde $\epsilon_{110} \sim t_3$. Potom ak $t_3(0.025)$ a $t_3(0.975)$ sú 0.025 a 0.975-kvantily Studentovho t -rozdelenia s tromi stupňami voľnosti, platí

$$P(t_3(0.025) \leq \epsilon_{110} \leq t_3(0.975)) = 0.95$$

a pre y_{110} máme

$$P(t_3(0.025) + \log(110) \leq y_{110} \leq t_3(0.975) + \log(110)) = 0.95.$$

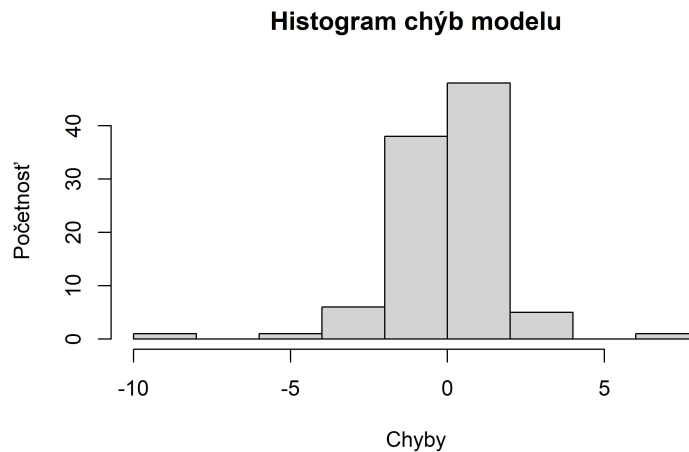
Skutočný 99-percentný predikčný interval vieme odvodiť analogicky a dostaneme $(t_3(0.005) + \log(110), t_3(0.995) + \log(110))$, kde $t_3(0.005)$ a $t_3(0.995)$ sú 0.005 a 0.995-kvantily t -rozdelenia s tromi stupňami voľnosti.

	$\mu_{110 100}$	$\text{var}_{110 100}$	95% pred. int.	99% pred. int.
Normálne chyby	3.9884	3.2904	(0.4458, 7.6023)	(-0.7072, 8.8371)
Bootstrap	3.8915	3.3267	(0.4269, 7.5180)	(-5.1732, 10.1452)
Analytické vzorce	3.9879	3.2511	(0.4540, 7.5219)	(-0.6565, 8.6323)
Skutočná hodnota	4.7005	3.0000	(1.5180, 7.8829)	(-1.1404, 10.5414)

Tabuľka 5.7: Porovnanie výsledkov pre pravdepodobnostné predpovede radu $y_t = \log(t) + \epsilon_t, \epsilon_t \sim t_3$ o 10 krokov dopredu.

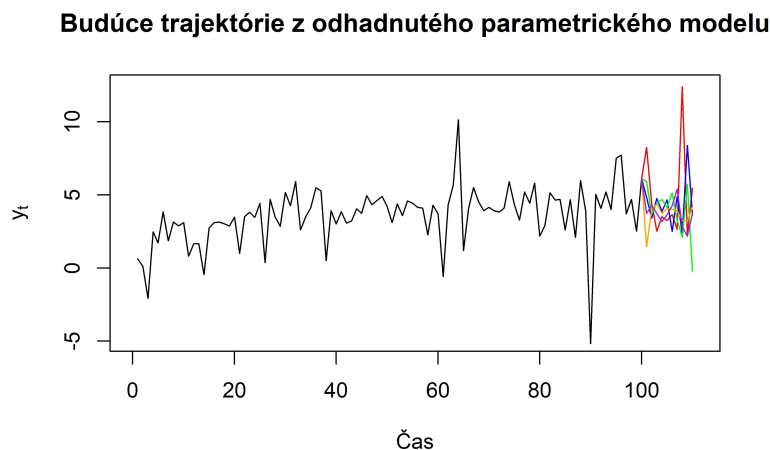
V tabuľke 5.7 vidíme, že výsledky sú opäť podobné pre všetky metódy. Výraznejšie sa líšia iba v prípade 99-percentného predikčného intervalu. Simulácie vychádzajúce z bootstrap metódy poskytli najlepší odhad jeho hornej hranice, hoci je celý interval o dosť širší, ako ten skutočný. Nepresnosť odhadov simulácií s normálnymi chybami aj analytických vzorcov by v tomto prípade mohla indikovať, že predpoklad normality z ktorého vychádzajú, nie je splnený.

Keďže sme pracovali so simulovanými dátami, vieme, že chyby majú v skutočnosti t -rozdelenie. Toto podozrenie by sme mohli získať aj na základe histogramu chybových členov, ktorý je na obrázku 5.11. Na ňom sa nám ukázali ťažké chvosty t -rozdelenia. Napriek tomu sme pomocou analytických vzorcov odvodených s predpokladom normálneho rozdelenia chybových členov stále dostali 95-percentný predikčný interval, ktorý je v porovnaní so skutočným celkom presný. Väčší rozdiel si môžeme všimnúť pri 99-percentnom predikčnom intervale. Pozreli sme sa ale aj na to, aké výsledky dostaneme v prípade, že budeme chyby do simulácií generovať z t -rozdelenia.



Obr. 5.11: Histogram chybových členov modelu $ETS(A, A_d, N)$ pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim t_3$.

Najskôr sme potrebovali určiť počet stupňov voľnosti. Ten sme použitím softvéru odhadli metódou maximálnej vierohodnosti z dát¹, ako 2.5233. Následne sme simulovali 5000 možných budúcich trajektórií pre 10 nasledujúcich období, pričom hodnoty ϵ_t , $t = 101, \dots, 110$, ktoré vystupujú v rekurzívnych vzorcoch (4.1), (4.2), sme generovali z t -rozdelenia práve s týmto počtom stupňov voľnosti. Päť takto získaných budúcich trajektórií je na obrázku 5.12.



Obr. 5.12: Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$ s chybami z odhadnutého modelu.

Z nasimulovaných hodnôt v čase 110 sme odhadli strednú hodnotu a rozptyl

¹Pomocou funkcie *optim()*, v ktorej špecifikácii sme zvolili počiatkovú hodnotu optimalizovaného parametra rovnú 5, zadali funkciu logaritmickej vierohodnosti a zvolili Brentovu metódu maximalizácie.

ako

$$\hat{\mu}_{110|100} = 3.9530, \quad \widehat{\text{var}}_{110|100} = 2.8218,$$

95-percentný predikčný interval nám vyšiel (0.7784, 7.4371) a 99-percentný predikčný interval sme dostali (−1.8948, 9.1274). Tento odhad strednej hodnoty i rozptylu je porovnateľný s odhadmi ostatných metód. Pokiaľ ide o predikčné intervaly, týmto spôsobom sme dostali užšie intervaly, ktoré takmer pokrývajú tie skutočné, a odhady ich spodných hraníc sú bližšie k skutočným spodným hraniciam, ako boli odhady simulácií s bootstrapom aj s normálnymi chybami.

Aby sme mohli posúdiť, ktorá z použitých metód poskytuje najpresnejšie výsledky, aj v tomto príklade sme vygenerovali sto rôznych realizácií modelu $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim t_3$, $t = 1, 2, \dots, 100$. Pre každú z nich sme spravili 5000 simulácií budúcich trajektórií za predpokladu normálneho rozdelenia chýb, s využitím bootstrap metódy aj generovaním chýb z t -rozdelenia s odhadnutým počtom stupňov voľnosti. Pre každú zo sto realizácií sme z nasimulovaných dát odhadli strednú hodnotu a rozptyl v čase 110 a 99-percentný predikčný interval. Pre jednotlivé realizácie sme tieto hodnoty spočítali aj pomocou analytických vzorcov. Následne sme zo sto hodnôt pre každú metódu vypočítali ich priemery. Porovnanie výsledkov, ktoré sme dostali, sa nachádza v tabuľke 5.8.

	$\mu_{110 100}$	$\text{var}_{110 100}$	Horná hranica pred. int.	Šírka pred. int.
Normálne chyby	4.4871	2.8166	8.7309	8.5029
Bootstrap	4.5240	2.7029	10.5240	11.5739
Odhadnutý model	4.4850	3.8911	10.5346	11.8882
Analytické vzorce	4.4816	2.8265	8.7513	8.5394
Skutočná hodnota	4.7005	3.0000	10.5414	11.6818

Tabuľka 5.8: Porovnanie výsledkov zo sto simulácií modelu $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim t_3$, pre rôzne metódy odhadov.

Odhady zo simulácií s normálne rozdelenými chybami sú opäť najbližšie k výsledkom z analytických vzorcov. V prípade predikčných intervalov sa to dá očakávať, vzhľadom na to, že analytické vzorce pre ne boli odvodené na základe predpokladu $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Ak však porovnáme výsledky so skutočnými hodnotami, vidíme, že simulácie založené na bootstrap metóde a na odhadnutom modeli poskytnú lepšie odhady hornej hranice 99-percentného predikčného intervalu aj jeho šírky.

5.1.3 Chyby z posunutého exponenciálneho rozdelenia

Stále pracujme s modelom

$$y_t = \log(t) + \epsilon_t, \quad t = 1, 2, \dots, 100,$$

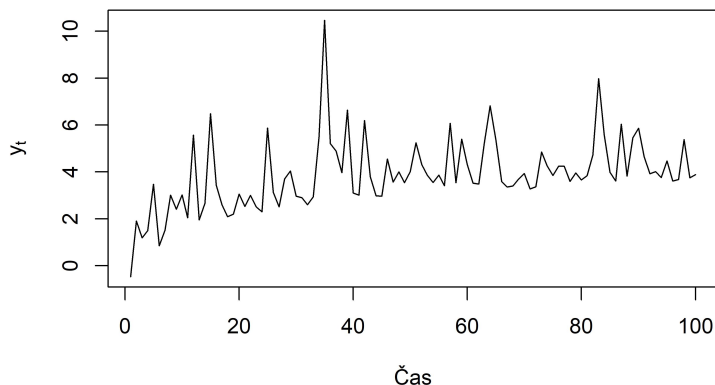
kde tentokrát uvažujeme chyby z o konštantu c posunutého exponenciálneho rozdelenia s parametrom λ s hustotou

$$f(x; \lambda) = \lambda \exp(-\lambda(x + c)) \cdot \mathbb{I}_{[x \geq -c]}.$$

Konkrétne budeme uvažovať rozdelenie s $\lambda = 1$ posunuté o $c = 1$, ktorého hustota má tvar

$$f(x) = \exp(-x - 1) \cdot \mathbb{I}_{[x \geq -1]}.$$

Ďalej budeme takéto rozdelenie označovať $\text{Exp}(1) - 1$. Pre $\epsilon_t \sim \text{Exp}(1) - 1$ platí $E[\epsilon_t] = \frac{1}{\lambda} - 1 = 1 - 1 = 0$, $\text{var}[\epsilon_t] = \frac{1}{\lambda^2} = 1$. Toto rozdelenie sme pre chybové členy vybrali z dôvodu jeho nesymetrie. Jedna z možných realizácií sto pozorovaní tohto časového radu je na obrázku 5.13.



Obr. 5.13: Simulované pozorovania časového radu $y_t = \log(t) + \epsilon_t, \epsilon_t \sim \text{Exp}(1) - 1$.

Funkcia $ets()$, rovnako ako v predchádzajúcich príkladoch, opäť určila model $\text{ETS}(A, A_d, N)$ za najlepší podľa AIC kritéria. Jeho parametre, počiatkové stavy a rozptyl chybových členov ϵ_t odhadla tak, ako uvádza tabuľka 5.9.

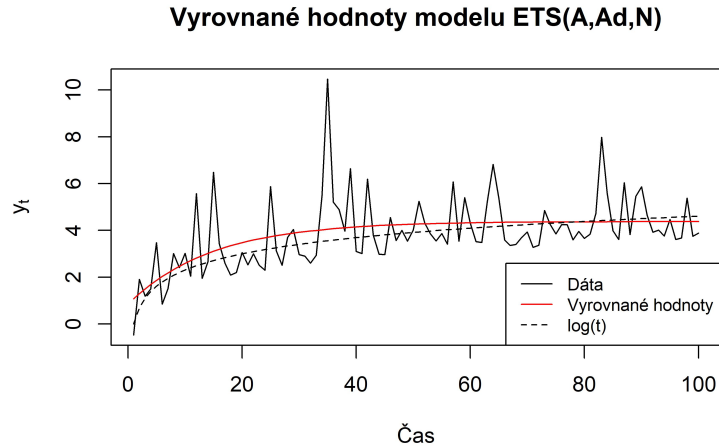
Parametre		Počiatkové stavy		Rozptyl	
$\hat{\alpha}$	0.0001	L_0	0.8473	$\hat{\sigma}^2$	1.6937
$\hat{\beta}$	0.0001	T_0	0.2481		
$\hat{\phi}$	0.9345				

Tabuľka 5.9: Odhady parametrov modelu $\text{ETS}(A, A_d, N)$ pre časový rad $y_t = \log(t) + \epsilon_t \sim \text{Exp}(1) - 1$.

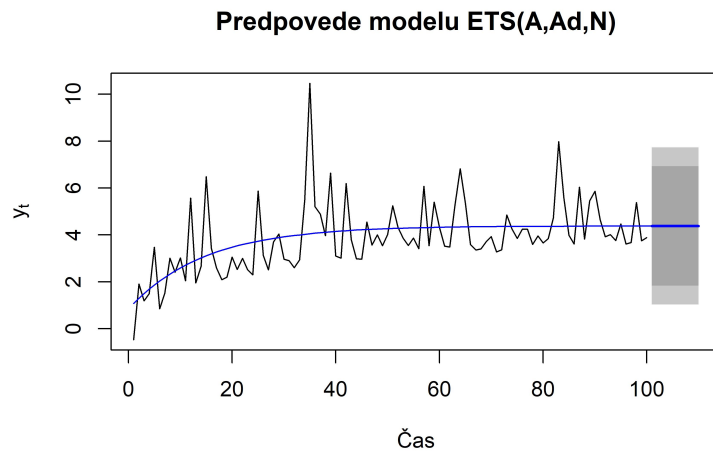
Odhady $\hat{\alpha}, \hat{\beta}$ sú znovu čísla blízke nule, takže vyrovnané hodnoty tohto modelu, zobrazené na obrázku 5.14, sú v čase t takmer identické s $L_{t-1} + \hat{\phi}T_{t-1}$. Pritom smernica T_t v čase t je približne 93 percent predchádzajúcej hodnoty smernice.

Na obrázku 5.15 môžeme vidieť bodové predpovede takto odhadnutého modelu pre 10 nasledujúcich období. Širší bledosivý pás predstavuje príslušné 99-percentné predikčné intervaly, tmavosivou sú zobrazené 95-percentné predikčné intervaly. Ich číselné hodnoty sú uvedené v tabuľke 5.10.

Rovnako ako v predchádzajúcich príkladoch nás aj teraz zaujímali charakteristiky pravdepodobnostného rozdelenia budúcich hodnôt. Preto sme simulovali



Obr. 5.14: Pozorovania časového radu $y_t = \log(t) + \epsilon_t, \epsilon_t \sim \text{Exp}(1) - 1$, a vyrovnané hodnoty odhadnutého modelu $\text{ETS}(A, A_d, N)$.



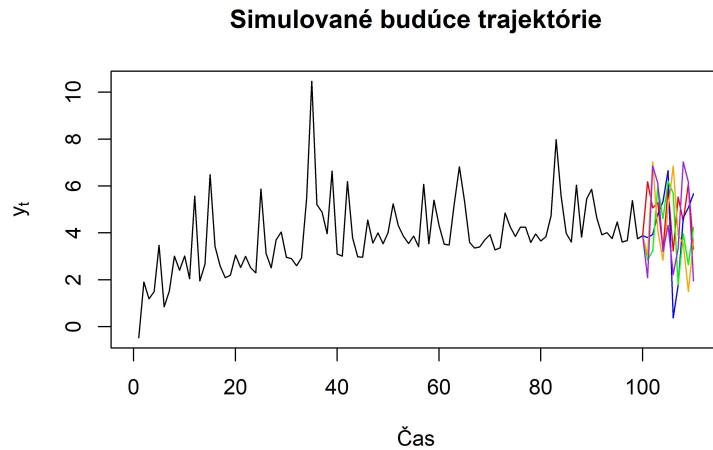
Obr. 5.15: Bodové predpovede modelu $\text{ETS}(A, A_d, N)$ o 10 krokov dopredu a predikčné intervaly pre rad $y_t = \log(t) + \epsilon_t, \epsilon_t \sim \text{Exp}(1) - 1$.

možné budúce trajektórie pre nasledujúcich 10 období a to opäť dvoma spôsobmi. Najskôr sme chyby $\epsilon_t, t = 101, \dots, 110$, do rekurzívnych vzorcov (4.1), (4.2) generovali z normálneho rozdelenia s nulovou strednou hodnotou a rozptylom $\hat{\sigma}^2$, potom sme využili metódu bootstrap. V oboch prípadoch sme simulovali 5000 trajektórií, pričom päť z nich sme pre jednotlivé metódy zobrazili na obrázkoch 5.16 a 5.17.

Z 5000 nasimulovaných hodnôt v čase 110 sme znovu odhadovali podmienenú strednú hodnotu $\mu_{110|100}$ a podmienený rozptyl $\text{var}_{110|100}$ budúcej hodnoty y_{110} a pomocou kvantilov sme určili aj 95 a 99-percentné predikčné intervaly pre hodnotu v tomto čase. Taktiež sme dané charakteristiky a predikčné intervaly vypočítali aj pomocou analytických vzorcov odvodených v štvrtej kapitole, a na záver sme určili presné hodnoty. Presnú strednú hodnotu a rozptyl sme počítali ana-

	Bodová predpoveď	95% pred. int.	99% pred. int.
$\hat{y}_{101}(100)$	4.3786	(1.8278, 6.9293)	(1.0263, 7.7308)
$\hat{y}_{102}(100)$	4.3788	(1.8280, 6.9295)	(1.0265, 7.7310)
$\hat{y}_{103}(100)$	4.3790	(1.8283, 6.9297)	(1.0268, 7.7312)
$\hat{y}_{104}(100)$	4.3792	(1.8284, 6.9299)	(1.0269, 7.7314)
$\hat{y}_{105}(100)$	4.3794	(1.8286, 6.9301)	(1.0271, 7.7316)
$\hat{y}_{106}(100)$	4.3795	(1.8288, 6.9303)	(1.0273, 7.7318)
$\hat{y}_{107}(100)$	4.3797	(1.8290, 6.9304)	(1.0275, 7.7320)
$\hat{y}_{108}(100)$	4.3799	(1.8291, 6.9306)	(1.0276, 7.7321)
$\hat{y}_{109}(100)$	4.3800	(1.8292, 6.9307)	(1.0277, 7.7323)
$\hat{y}_{110}(100)$	4.3801	(1.8294, 6.9309)	(1.0279, 7.7324)

Tabuľka 5.10: Hodnoty bodových predpovedí modelu ETS(A, A_d, N) a príslušné predikčné intervaly pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \text{Exp}(1) - 1$.



Obr. 5.16: Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \text{Exp}(1) - 1$, s chybami generovanými z $\mathcal{N}(0, \hat{\sigma}^2)$.

logicky ako v predchádzajúcich častiach. Predikčný interval sme dostali z toho, že

$$y_{110} = \log(110) + \epsilon_{110},$$

kde $\epsilon_{110} = Z - 1$, pričom $Z \sim \text{Exp}(1)$. Potom ak označíme 0.025 a 0.975-kvantily exponenciálneho rozdelenia s parametrom 1 ako $z_{0.025}$, resp. $z_{0.975}$, platí

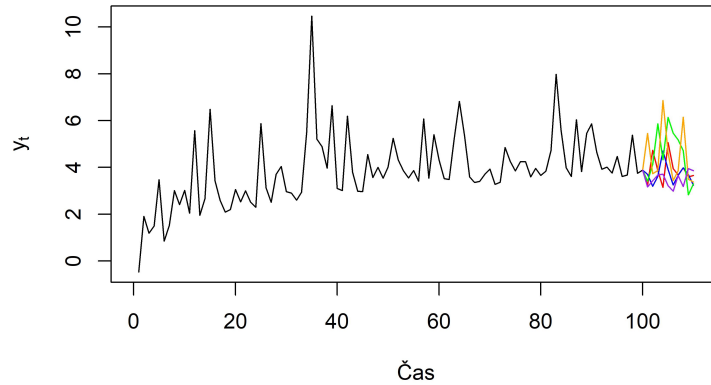
$$P(z_{0.025} \leq Z \leq z_{0.975}) = 0.95$$

a teda pre y_{110} máme

$$P(z_{0.025} - 1 + \log(110) \leq y_{110} \leq z_{0.975} - 1 + \log(110)) = 0.95.$$

99-percentný interval by sme odvodili analogicky a dostali by sme $(z_{0.005} - 1 + \log(110), z_{0.995} - 1 + \log(110))$, kde z_p je p -kvantil exponenciálneho rozdelenia s parametrom 1.

Budúce trajektórie s použitím bootstrap metódy



Obr. 5.17: Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \text{Exp}(1) - 1$, s využitím metódy bootstrap.

Číselné hodnoty výsledkov všetkých popísaných metód sú uvedené v tabuľke 5.11. Ak porovnáваме odhady zo simulácií so skutočnými hodnotami, simulácie s normálnymi chybami nám poskytli nepatrne presnejší odhad strednej hodnoty, bootstrap nepatrne presnejší odhad rozptylu. Ak porovnáваме odhadnuté predikčné intervaly so skutočnými predikčnými intervalmi, môžeme si všimnúť, že bootstrap nám v prípade 95-percentného intervalu poskytne jednoznačne najlepší odhad jeho hraníc. Pre 99-percentný predikčný interval bootstrap najlepší odhadol spodnú hranicu, celý interval je však o dosť širší ako ten skutočný.

	$\mu_{110 100}$	$\text{var}_{110 100}$	95% pred. int.	99% pred. int.
Normálne chyby	4.4066	1.6809	(1.8740, 6.9323)	(1.1153, 7.7534)
Bootstrap	4.3499	1.6127	(3.0059, 7.7635)	(2.8203, 10.7977)
Analytické vzorce	4.3801	1.6937	(1.8294, 6.9309)	(1.0279, 7.7324)
Skutočná hodnota	4.7005	1.0000	(3.7258, 7.3894)	(3.7055, 8.9988)

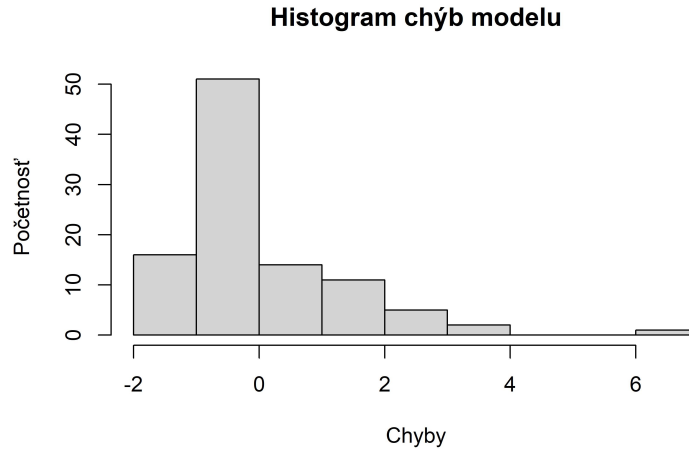
Tabuľka 5.11: Porovnanie výsledkov pre pravdepodobnostné predpovede radu $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \text{Exp}(1) - 1$ o 10 rokov dopredu.

Ak by sme nepracovali so simulovanými ale s reálnymi dátami, informáciou o presnom rozdelení chybových členov by sme nedisponovali. Vždy si však môžeme vykresliť histogram chýb nášho modelu a posúdiť vhodnosť predpokladu ich normálneho rozdelenia. Takýto histogram je na obrázku 5.18.

Pohľad na tento histogram normálne rozdelenie nepripomína, dá sa v ňom však vidieť klesajúca exponenciála. Skúsili sme teda nájsť vhodný parametrický model pre chybové členy a odhadnúť jeho parametre. Zvolili sme si posunuté exponenciálne rozdelenie s parametrom a , ktoré má hustotu

$$f(x; a) = \frac{1}{a} \exp\left(-\frac{x}{a} - 1\right), \quad x \geq -a,$$

a potom maximálne vierohodný odhad parametra a je $\hat{a} = -\min_{1 \leq t \leq 100}(e_t)$.



Obr. 5.18: Histogram chybových členov modelu $ETS(A, A_d, N)$ pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \text{Exp}(1) - 1$.

Pre chybové členy nášho modelu platí $\hat{a} = 1.5586$. Ďalej sme vygenerovali 5000 možných budúcich trajektórií pre 10 nasledujúcich období, a to tak, že sme opäť používali rekurzívne vzorce (4.1) a (4.2), v ktorých sme za $\epsilon_t, t = 101, \dots, 110$, dosádzali hodnoty generované z posunutého exponenciálneho rozdelenia s parametrom $a = 1.5586$. Päť takto simulovaných vývojev je na obrázku 5.19. Odhady podmienenej strednej hodnoty a podmieneného rozptylu v čase 110 sú

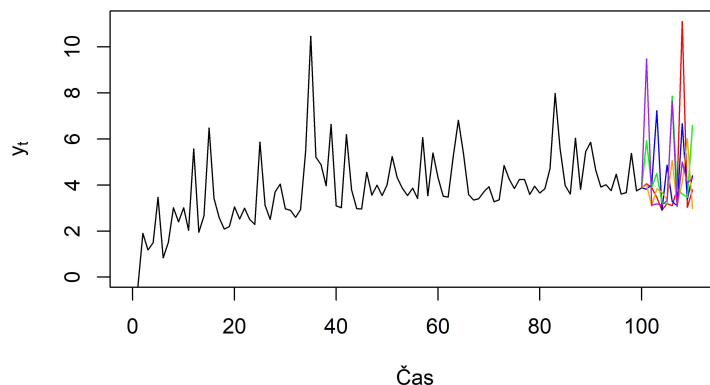
$$\hat{\mu}_{110|100} = 4.3684, \quad \widehat{\text{var}}_{110|100} = 2.0933,$$

predikčné intervaly, ktorých hranice sme dostali ako 0.025 a 0.975 výberové kvantily, resp. 0.005 a 0.995 výberové kvantily, nasimulovaných hodnôt v čase 110, sú (2.8608, 8.3225), resp. (2.8289, 9.4475). Odhad strednej hodnoty a rozptylu je medzi bootstrapom a simuláciami za predpokladu normálne rozdelených chýb, predikčné intervaly pokrývajú tie skutočné. Najlepšie výsledky, v porovnaní so skutočnými hodnotami, nám stále poskytla bootstrap metóda, avšak 99-percentný predikčný interval vyšiel lepší zo simulácií z odhadnutého parametrického modelu.

Pre porovnanie presnosti jednotlivých simulačných metód aplikovaných na dáta, ktoré sa riadia modelom $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \text{Exp}(1) - 1$, sme znovu vygenerovali sto jeho realizácií a pre každú sme simulovali možné budúce vývoje. V simuláciách sme najskôr predpokladali, že chyby majú normálne rozdelenie $\mathcal{N}(0, \hat{\sigma}^2)$, vybrali sme ich náhodne zo známych minulých chýb a tiež sme odhadovali ich rozdelenie. Potom sme odhadovali strednú hodnotu, rozptyl a 99-percentný predikčný interval pre budúcu hodnotu v čase $t = 110$. Pre každú realizáciu sme strednú hodnotu, rozptyl, aj 99-percentný predikčný interval vypočítali aj pomocou analytických vzorcov. Na záver sme 100 získaných hodnôt pre každú charakteristiku pravdepodobnostného rozdelenia spriemerovali, aby sme výsledky mohli porovnať so skutočnými hodnotami. Toto porovnanie môžeme vidieť v tabuľke 5.12.

Podobne ako v predchádzajúcich príkladoch, aj tentokrát sa nám potvrdilo, že simulácie s použitím bootstrap metódy poskytujú najlepšie výsledky v porovnaní so skutočnými hodnotami. Zaujímavé však je, že pomocou analytických vzorcov

Budúce trajektórie z odhadnutého parametrického modelu



Obr. 5.19: Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \text{Exp}(1) - 1$ s chybami z odhadnutého modelu.

	$\mu_{110 100}$	$\text{var}_{110 100}$	Horná hranica pred. int.	Šírka pred. int.
Normálne chyby	4.5421	1.1137	7.2401	5.3887
Bootstrap	4.6214	1.0590	8.8498	5.6823
Odhadnutý model	4.5189	1.9809	10.1994	7.2899
Analytické vzorce	4.5439	1.1150	7.2428	5.3979
Skutočná hodnota	4.7005	1.0000	8.9988	5.2933

Tabuľka 5.12: Porovnanie výsledkov zo sto simulácií modelu $y_t = \log(t) + \epsilon_t$, $\epsilon_t \sim \text{Exp}(1) - 1$, pre rôzne metódy odhadov.

aj s využitím simulácií s normálne rozdelenými chybami dostávame predikčné intervaly, ktoré sú približne rovnako široké ako je ten skutočný.

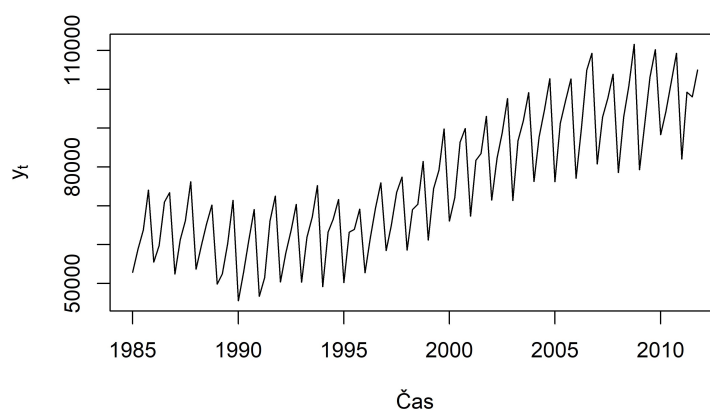
5.1.4 Zhrnutie

Na základe doteraz prezentovaných príkladov sa javia simulácie s využitím bootstrap metódy ako najpresnejší spôsob odhadovania predikčného intervalu pre budúce hodnoty časového radu. Pri nich totiž pracujeme bez predpokladu o pravdepodobnostnom rozdelení chybových členov a teda fungujú vcelku presne pre chyby z rôznych rozdelení. V prípade t -rozdelenia chybových členov sme však dostali z odhadnutého modelu porovnateľne presný odhad 99-percentného predikčného intervalu ako bootstrap metódou.

Čo sa týka odhadov podmienenej strednej hodnoty a podmieneného rozptylu v budúcich časoch, tie vychádzajú približne rovnako presne pre simulácie s normálne rozdelenými chybami aj s využitím bootstrap metódy. S využitím generovania chýb z odhadnutého parametrického modelu sme však v oboch príkladoch dostali menej presný odhad podmieneného rozptylu.

5.2 Reálne dáta

V tejto časti budeme aplikovať doteraz prezentované metódy aj na reálne dáta. Budeme pracovať konkrétne s časovým radom štvrtročných objemov predaja stolového vína v Austrálii v tisíckach litrov v období od roku 1985 do roku 2011², to znamená so 108 pozorovaniami. Tie môžeme vidieť na obrázku 5.20.



Obr. 5.20: Pozorovania štvrtročných objemov predaja stolového vína medzi rokmi 1985 a 2011.

Vidíme, že každý rok je objem predaja najnižší na jeho začiatku a najvyšší koncom roka, čo môže byť spôsobené sviatočným obdobím. Tieto sezónne výkyvy však pôsobia približne rovnaké v jednotlivých rokoch a nezväčšujú sa s rastúcim objemom predaja, ktorý môžeme sledovať približne od roku 1996. Preto predpokladáme, že vhodným modelom pre tieto dáta by mohol byť model s aditívnou sezónnosťou s dĺžkou sezóny $m = 4$.

Na jeho odhad sme opäť použili funkciu *ets()*, ktorú sme nechali vybrať najvhodnejší model automaticky. Najmenšiu hodnotu AIC mal model ETS(A,A,A) s aditívnou chybou, aditívnym trendom a, ako sme očakávali, aditívnou sezónnosťou. Pomocou rovníc ho môžeme zapísať ako

$$y_t = (1, 1, 0, 0, 0, 1) \cdot \begin{pmatrix} L_{t-1} \\ T_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \\ S_{t-4} \end{pmatrix} + \epsilon_t,$$

$$\begin{pmatrix} L_t \\ T_t \\ S_t \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} L_{t-1} \\ T_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \\ S_{t-4} \end{pmatrix} + \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ 0 \\ 0 \end{pmatrix} \cdot \epsilon_t.$$

²zdroj: Australian Bureau of Statistics, www.abs.gov.au, katalógové číslo 8504.0, Table 2.

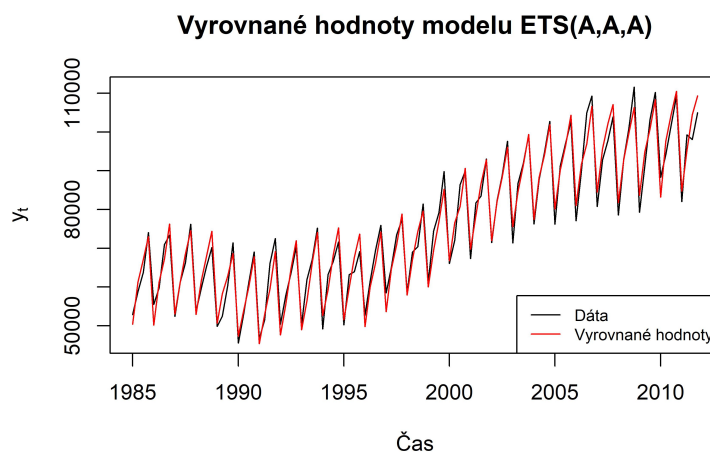
Odhadnuté hodnoty vyrovnávacích konštánt α, β a γ sa nachádzajú v tabuľke 5.13. Taktiež v nej môžeme vidieť odhadnuté počiatkové stavy $L_0, T_0, S_0, S_{-1}, S_{-2}, S_{-3}$, a odhad rozptylu chybových členov σ^2 .

Parametre		Počiatkové stavy		Rozptyl	
$\hat{\alpha}$	0.3022	L_0	62819.5175	$\hat{\sigma}^2$	9788468
$\hat{\beta}$	0.0307	T_0	56.1370		
$\hat{\gamma}$	0.1458	S_0	10990.8777		
		S_{-1}	3980.6449		
		S_{-2}	-2461.5984		
		S_{-3}	-12509.9241		

Tabuľka 5.13: Odhady parametrov modelu ETS(A,A,A) pre štvrtročné objemy predaja stolového vína v Austrálii.

Hodnota S_0 v tomto prípade predstavuje odhad sezónnej zložky pre posledný štvrtrok roku 1984, ktorý naše dáta neobsahujú. S_{-1} je jej odhad pre tretí štvrtrok, S_{-2} pre druhý a S_{-3} je odhad sezónnej zložky pre prvý štvrtrok 1984. Môžeme si všimnúť, že vplyvom takto odhadnutej sezónnej zložky naozaj dostávame vyšší objem predaja stolového vína na konci roka, a naopak nižší na jeho začiatku, presne ako sme konštatovali v úvodnom popise pozorovaní časového radu.

Vyrovnané hodnoty odhadnutého modelu ETS(A,A,A), vypočítané ako $\hat{y}_t = L_t + S_t$, môžeme vidieť na obrázku 5.21.

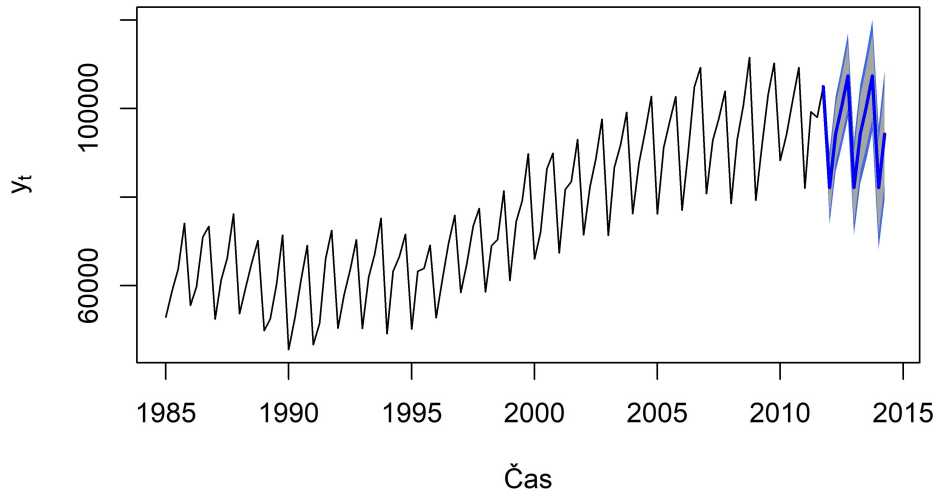


Obr. 5.21: Pozorovania štvrtročných objemov predaja stolového vína v Austrálii a vyrovnané hodnoty odhadnutého modelu ETS(A,A,A).

Bodové predpovede nášho modelu pre 10 nasledujúcich období sú spolu so sivými 95-percentnými a širšími modrými 99-percentnými predikčnými intervalmi na obrázku 5.22. Predpovedané budúce hodnoty sú zobrazené modrou krivkou.

Číselné hodnoty predpovedí a hraníc predikčných intervalov, ktoré sme dostali použitím funkcie `forecast.ets()`, sú uvedené v tabuľke 5.14. Tieto hodnoty

Predpovede modelu ETS(A,A,A)



Obr. 5.22: Bodové predpovede modelu ETS(A,A,A) o 10 krokov dopredu a predikčné intervaly pre časový rad štvrtročných objemov predaja stolového vína v Austrálii.

	Bodová predpoved	95% pred. int.		99% pred.int	
1Q 2012	82 161.0161	(76 028.9693,	88 293.0629)	(74 102.1406,	90 219.8917)
2Q 2012	94 275.1167	(87 812.1988,	100 738.0346)	(85 781.4028,	102 768.8306)
3Q 2012	100 383.1080	(93 546.3608,	107 219.8551)	(91 398.0992,	109 368.1167)
4Q 2012	107 410.5110	(100 158.7275,	114 662.2945)	(97 880.0520,	116 940.9700)
1Q 2013	82 155.2059	(74 102.6523,	90 207.7595)	(71 572.3565,	92 738.0552)
2Q 2013	94 269.3064	(85 745.5376,	102 793.0753)	(83 067.1753,	105 471.4376)
3Q 2013	100 377.2977	(91 346.5742,	109 408.0212)	(88 508.9153,	112 245.6802)
4Q 2013	107 404.7008	(97 833.2553,	116 976.1462)	(94 825.6892,	119 983.7123)
1Q 2014	82 149.3956	(71 675.1157,	92 623.6756)	(68 383.8584,	95 914.9329)
2Q 2014	94 263.4962	(83 204.5384,	105 322.4540)	(79 729.5620,	108 797.4304)

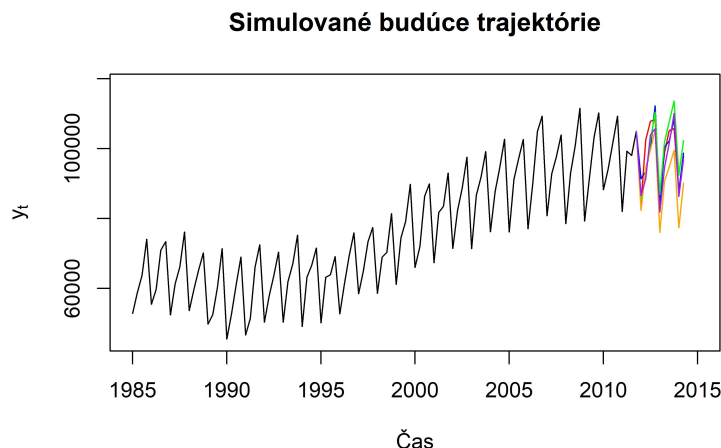
Tabuľka 5.14: Hodnoty bodových predpovedí modelu ETS(A,A,A) a príslušné predikčné intervaly pre štvrtročné objemy predaja stolového vína v Austrálii.

sa zhodujú s výsledkami, ktoré by sme dostali aplikáciou analytických vzorcov odvodených v štvrtej kapitole.

Opäť nás však zaujímalo aj pravdepodobnostné rozdelenie budúcich hodnôt podmienené informáciou do času $T = 108$, konkrétne jeho stredná hodnota a rozptyl, ktoré sme aj v prípade reálnych dát odhadovali zo simulácií budúcich trajektórií. Tie sme počítali s využitím rekurzívnych vzorcov (4.1), (4.2) spôsobom, ktorý je podrobne popísaný v časti 5.1.1 na strane 34.

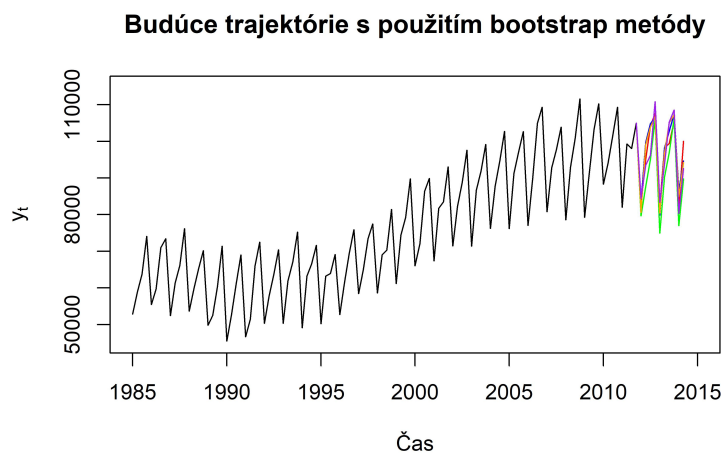
Najskôr sme vypočítali 5000 možných budúcich trajektórií tak, že sme hodnoty $\epsilon_t, t = 109, \dots, 118$, ktoré potrebujeme do výpočtu budúcich hodnôt y_t a \mathbf{x}_t , generovali ako náhodné čísla z normálneho rozdelenia $\mathcal{N}(0, \hat{\sigma}^2)$ s nulovou strednou hodnotou a rozptylom rovným odhadu rozptylu chybových členov z modelu.

Päť z nich je na obrázku 5.23.



Obr. 5.23: Simulované budúce trajektórie pre časový rad štvrtročných objemov predaja stolového vína v Austrálii s chybami z $\mathcal{N}(0, \hat{\sigma}^2)$.

V ďalších 5000 simuláciách sme budúce hodnoty $\epsilon_t, t = 109, \dots, 118$, získali bootstrap metódou, čo znamená, že sme ich vybrali náhodne s opakovaním z množiny známych minulých chýb $e_t = y_t - \hat{\mathbf{w}}' \mathbf{x}_{t-1}, t = 1, \dots, 108$. Ukážku piatich z nich môžeme vidieť na obrázku 5.24.



Obr. 5.24: Simulované budúce trajektórie pre časový rad štvrtročných objemov predaja stolového vína v Austrálii s využitím bootstrap metódy.

V oboch prípadoch sme sa zamerali na predpovede o desať krokov dopredu, t.j. na objem predaja stolového vína v Austrálii v druhom štvrtroku 2014. Strednú hodnotu a rozptyl rozdelenia v čase $t = 118$, ktorý reprezentuje práve tento štvrtrok, podmienené informáciou do času $T = 108$, sme odhadli ako výberový priemer a výberový rozptyl nasimulovaných hodnôt pre čas 118. Predikčné intervaly pre hodnoty v tomto budúcom čase sme odhadovali ako $\left(\frac{\alpha}{2}\right)$ a $\left(1 - \frac{\alpha}{2}\right)$

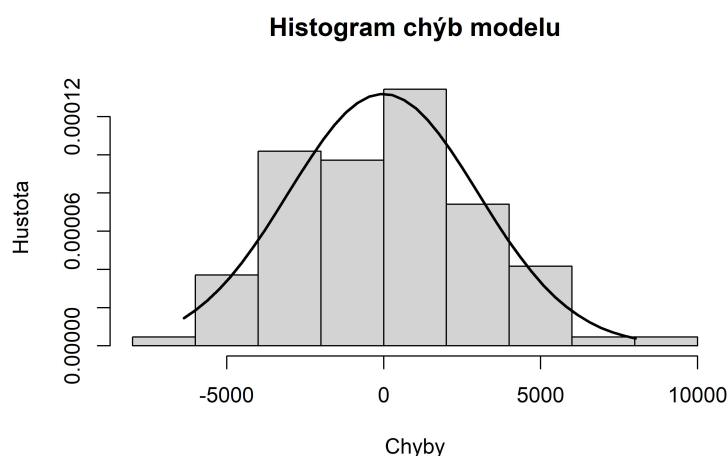
výberové kvantily nasimulovaných hodnôt y_{118} , pričom sme uvažovali postupne $\alpha = 0.05$ a $\alpha = 0.01$.

Podmienenu strednú hodnotu $\mu_{118|108}$ a podmienený rozptyl $\text{var}_{118|108}$, rovnako ako aj 95 a 99-percentné predikčné intervaly, sme ďalej vypočítali aj pomocou analytických vzorcov zo štvrtej kapitoly. Výsledky zo všetkých použitých metód sú zhrnuté v tabuľke 5.15.

	$\mu_{118 108}$	$\text{var}_{118 108}$	95% pred. int.	99% pred. int.
Normálne chyby	94 307.8920	30 908 289.7851	(83 396.7276, 105 132.5072)	(79 986.9182, 108 208.5118)
Bootstrap	94 266.7634	27 194 570.4720	(83 452.5373, 103 903.4873)	(79 791.5145, 109 028.9223)
Analytické vzorce	94 263.4962	31 837 006.1021	(83 204.5384, 105 322.4540)	(79 729.5620, 108 797.4304)

Tabuľka 5.15: Porovnanie výsledkov pre pravdepodobnostné predpovede o 10 krokov dopredu pre štvrtročné objemy predaja stolového vína v Austrálii.

Vidíme, že obe simulačné metódy poskytli podobné výsledky, ktoré sú blízke aj výsledkom z analytických vzorcov. V prípade predikčného intervalu odhadnutého zo simulácií s normálnymi chybami sa to dá očakávať, keďže analytický predikčný interval je odvodený rovnako s predpokladom, že chyby sú normálne rozdelené. Podobné výsledky však poskytli aj simulácie s využitím bootstrap metódy, ktoré sa aj na základe príkladov so simulovanými dátami javili ako najpresnejšie. Platí to dokonca i v prípade 99-percentných intervalov, pri ktorých sa v predchádzajúcich príkladoch, kde sme uvažovali chyby z iného ako normálneho rozdelenia, rozdiel medzi simulačnými metódami prejavil. Môžeme teda predpokladať, že rozdelenie chýb je naozaj blízke normálnemu, čo môžeme vidieť aj na histograme chýb modelu na obrázku 5.25. Pre jednoduchšie porovnanie sme ním preložili krivku hustoty normálneho rozdelenia so strednou hodnotou rovnou priemeru chybových členov a rozptylom rovným ich výberovému rozptylu. V takom prípade môžeme považovať bodové predpovede aj odhady predikčných intervalov oboch simulačných metód za spoľahlivé.



Obr. 5.25: Histogram chybových členov modelu ETS(A,A,A) pre štvrtročné objemy predaja stolového vína v Austrálii.

Keďže pracujeme s reálnymi dátami a teda nepoznáme skutočné rozdelenie chýb, nemôžeme odhady predikčných intervalov porovnať so žiadnymi teoretickými hodnotami. Pokiaľ však ide o bodové predpovede, môžeme sa pozrieť na skutočnú pozorovanú hodnotu objemu predaja stolového vína v Austrálii v druhom štvrtroku 2014 a porovnať ich s ním. Predaj v tomto čase dosiahol 94 381 tisíc litrov. Simulácie za predpokladu normálne rozdelených chýb teda predpovedali jeho hodnotu o čosi lepšie ako bootstrap či analytické vzorce.

Záver

Cieľom našej diplomovej práce bolo prezentovať štatistické modely pre základné metódy exponenciálneho vyrovnávania, popísať ich využitie pri odhadovaní pravdepodobnostného rozdelenia budúcich hodnôt, ktoré nám umožňuje konštruovať pravdepodobnostné predpovede, a následne dané modely a metódy aplikovať na rôzne časové rady.

V teoretickej časti práce sme sa najskôr venovali popisu metód exponenciálneho vyrovnávania ako algoritmov na výpočet vyrovnaných hodnôt a bodových predpovedí analyzovaných časových radov. Zaoberali sme sa konkrétne metódami jednoduchého a dvojitého exponenciálneho vyrovnávania, Holtovou metódou a jej rozšírením pre dáta so sezónnymi vplyvmi, Holtovou-Wintersovou metódou. Teórii týkajúcu sa týchto metód sme spracovávali vychádzajúc primárne z (Cipra, 2008). Predstavili sme však aj modifikácie Holtovej a Holtovej-Wintersovej metódy a to konkrétne ich verzie s tlmeným trendom, ktorých predpovede nerastú či neklesajú v budúcnosti stále rovnakým tempom a preto nedospejú k neprimerane vysokým, resp. nízkym budúcim hodnotám. Tieto modifikácie sme spracovávali najmä podľa (Hyndman a kol., 2008). Popis jednotlivých metód sme doplnili aj ilustráciami na konkrétnych príkladoch.

V druhej polovici teoretickej časti práce, ktorú predstavuje tretia a štvrtá kapitola, sme sa zamerali na štatistické stavové modely, ktoré sú založené na prezentovaných predpovedných metódach. Tento prístup k problematike exponenciálneho vyrovnávania nám umožnil nielen konštruovať bodové predpovede budúcich hodnôt, ale tiež odhadovať ich pravdepodobnostné rozdelenie, vďaka čomu sme mohli odhadovať aj predikčné intervaly s požadovanou spoľahlivosťou. V rámci našej práce sme sa zaoberali šiestimi čisto aditívnymi modelmi z celkovo tridsiatich existujúcich modelov exponenciálneho vyrovnávania.

Tretia kapitola sa venovala formulácii stavových modelov, popisu odhadu ich parametrov metódou maximálnej vierohodnosti a metóde výberu najvhodnejšieho modelu pre analyzovaný časový rad. V štvrtej kapitole sme odvodili analytické vyjadrenie pre strednú hodnotu a rozptyl minulosťou podmieneného pravdepodobnostného rozdelenia budúcich hodnôt, rovnako ako pre hranice predikčných intervalov pre budúce hodnoty. Vzorce pre výpočet predikčných intervalov pre niektoré metódy sa síce nachádzajú aj v (Cipra, 2008), avšak nie sú bližšie vysvetlené a nie je jasné, ako sú konštruované. Ďalej sme prezentovali simulačné prístupy, pomocou ktorých môžeme vygenerovať veľké množstvo možných budúcich trajektorií daného časového radu a z tohto množstva hodnôt, ktoré máme k dispozícii pre vybraný budúci čas, odhadovať podmienenú strednú hodnotu a podmienený rozptyl rozdelenia v tomto čase. Pri spracovaní tretej a štvrtej kapitoly sme vychádzali z (Hyndman a kol., 2008) a (Hyndman a Athanasopoulos, 2018).

Piata kapitola a teda praktická časť našej práce pozostáva z aplikácie popísaných metód a modelov na simulované a ďalej aj na reálne dáta.

V prvých troch príkladoch sme simulovali časové rady, v ktorých sme chyby generovali postupne zo štandardného normálneho rozdelenia $\mathcal{N}(0,1)$, z t -rozdelenia s tromi stupňami voľnosti a z posunutého exponenciálneho rozdelenia. Pre všetky takto získané časové rady sme následne simulovali ich budúce trajektórie. Najskôr sme uvažovali, že budúce chyby pochádzajú z normálneho rozdelenia, potom

sme využívali metódu bootstrap. V príkladoch, kde sme pre chyby uvažovali iné ako normálne rozdelenie, sme tiež odhadli parametrický model pre chybové členy a generovali budúce chyby z odhadnutého rozdelenia. Z nasimulovaných budúcich hodnôt sme odhadli podmienenú strednú hodnotu, podmienený rozptyl a hranice 95 a 99-percentných predikčných intervalov. Všetky tieto hodnoty sme vypočítali tiež pomocou analytických vzorcov. Tento postup, vrátane generovania nových dát, sme zopakovali vždy 100 krát a vypočítali priemerné hodnoty pre dané veličiny, ktoré sme porovnali s teoretickými, odvodenými na základe znalosti skutočného rozdelenia chybových členov.

V príkladoch so simulovanými dátami sa ako najlepší prístup javil odhad zo simulácií, ktoré využívali bootstrap metódu. Tá, keďže nepracuje s predpokladom o pravdepodobnostnom rozdelení chybových členov, fungovala spoľahlivo vo všetkých prípadoch, dokonca aj pri odhade 99-percentného predikčného intervalu. Simulácie s chybami z normálneho rozdelenia produkovali v prípade, že skutočné rozdelenie normálne nie je, pre 99-percentný predikčný interval dosť nepresné predpovede. Výsledky simulácií z odhadnutého parametrického modelu boli v prípade t -rozdelenia rovnako veľmi blízke skutočným, v prípade posunutého exponenciálneho rozdelenia sme až takú presnosť nedosiahli. Na základe toho by sme v praxi odporúčali používať simulácie s využitím bootstrap metódy, ktorá poskytovala najlepšie výsledky vo všetkých prípadoch.

V ďalšej časti piatej kapitoly sme sa venovali aplikácii popísaných metód a modelov aj na reálne dáta. Opäť sme simulovali budúce trajektórie za predpokladu, že chyby majú normálne rozdelenie, aj s využitím bootstrap metódy, a odhadli sme podmienenú strednú hodnotu, podmienený rozptyl a predikčné intervaly pre predpoveď o 10 krokov dopredu. Dané hodnoty sme vypočítali aj s využitím analytických vzorcov. V tomto prípade všetky metódy poskytli podobné výsledky, na základe čoho sme usúdili, že skutočné rozdelenie chybových členov je zrejme blízke normálnemu. V takom prípade môžeme považovať výsledky všetkých použitých metód za spoľahlivé.

Myslíme si, že cieľ diplomovej práce môžeme považovať za splnený. V práci sme priniesli prehľad metód využívaných na výpočet bodových predpovedí aj prístupov, pomocou ktorých vieme odhadovať pravdepodobnostné rozdelenie budúcich hodnôt a tak konštruovať pravdepodobnostné predpovede. Dôkladne sme popísali aj aplikáciu daných metód na príklady, s ktorými sa môže čitateľ stretnúť a výsledky jednotlivých metód porovnali.

Zoznam použitej literatúry

- BERGMEIR, C., HYNDMAN, R. J. a BENÍTEZ, J. M. (2016). Bagging Exponential Smoothing Methods using STL Decomposition and Box–Cox Transformation. *International journal of forecasting*, **32**(2), 303–312.
- BROWN, R. G. (1959). *Statistical Forecasting for Inventory Control*. First Edition. McGraw–Hill, New York. ISBN 978-0070081451.
- CIPRA, T. (2008). *Finanční ekonometrie*. 1. vydání. Ekopress, Praha. ISBN 978-80-86929-43-9.
- CLEVELAND, R. B., CLEVELAND, W. S., MCRAE, J. E. a TERPENNING, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of official statistics*, **6**(1), 3–73.
- GARDNER JR, E. S. a MCKENZIE, E. (1985). Forecasting Trends in Time Series. *Management Science*, **31**(10), 1237–1246.
- HOLT, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, **20**(1), 5–10.
- HYNDMAN, R., KOEHLER, A. B., ORD, J. K. a SNYDER, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. First Edition. Springer Science & Business Media. ISBN 978-3-540-71918-2.
- HYNDMAN, R. J. a ATHANASOPOULOS, G. (2018). *Forecasting: Principles and Practice*. Second Edition. OTexts, Melbourne.
- HYNDMAN, R. J. a KHANDAKAR, Y. (2008). Automatic Time Series Forecasting: the forecast Package for R. *Journal of Statistical Software*, **26**(3), 1–22. URL <http://www.jstatsoft.org/article/view/v027i03>.
- ORD, J. K., KOEHLER, A. B. a SNYDER, R. D. (1997). Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models. *Journal of the American Statistical Association*, **92**(440), 1621–1629.
- PRÁŠKOVÁ, Z. (2016). *Základy náhodných procesů II*. 2., upravené vydání. Univerzita Karlova, Nakladatelství Karolinum, Praha. ISBN 978-80-246-3529-3.
- R CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- WINTERS, P. R. (1960). Forecasting Sales by Exponentially Weighted Moving Averages. *Management science*, **6**(3), 324–342.
- ZVÁRA, K. (2008). *Regrese*. 1. vydání. MatfyzPress, Praha. ISBN 978-80-7378-041-8.

Zoznam obrázkov

2.1	Vyrovnanie metódou jednoduchého exponenciálneho	8
2.2	Vyrovnanie časového radu pre priemer lemu	12
2.3	Porovnanie predpovedí klasickej a tlmenej Holtovej metódy.	13
2.4	Časový rad s aditívnou (vľavo) a multiplikatívnou sezónnosťou.	14
2.5	Vyrovnanie a predpoveď časového radu logaritmických	16
2.6	Porovnanie predpovedí klasickej a tlmenej Holtovej-Wintersovej metódy.	17
5.1	Simulované pozorovania časového	32
5.2	Pozorovania časového radu $y_t = \log(t) + \epsilon_t, \epsilon_t \sim \mathcal{N}(0,1)$,	32
5.3	Bodové predpovede modelu ETS(A, A_d, N) o 10 krokov	33
5.4	Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$,	35
5.5	Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$,	35
5.6	Simulované pozorovania časového radu	38
5.7	Pozorovania časového radu $y_t = \log(t) + \epsilon_t, \epsilon_t \sim t_3$,	38
5.8	Bodové predpovede modelu ETS(A, A_d, N) o 10 krokov	39
5.9	Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$,	40
5.10	Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$,	40
5.11	Histogram chybových členov modelu ETS(A, A_d, N)	42
5.12	Simulované budúce trajektórie pre časový rad	42
5.13	Simulované pozorovania časového radu	44
5.14	Pozorovania časového radu $y_t = \log(t) + \epsilon_t, \epsilon_t \sim \text{Exp}(1) - 1$,	45
5.15	Bodové predpovede modelu ETS(A, A_d, N) o 10 krokov	45
5.16	Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$,	46
5.17	Simulované budúce trajektórie pre časový rad $y_t = \log(t) + \epsilon_t$,	47
5.18	Histogram chybových členov modelu ETS(A, A_d, N)	48
5.19	Simulované budúce trajektórie pre časový rad	49
5.20	Pozorovania štvrtročných objemov predaja	50
5.21	Pozorovania štvrtročných objemov predaja stolového vína	51
5.22	Bodové predpovede modelu ETS(A, A, A) o 10 krokov dopredu	52
5.23	Simulované budúce trajektórie pre časový rad štvrtročných	53
5.24	Simulované budúce trajektórie pre časový rad štvrtročných	53
5.25	Histogram chybových členov modelu ETS(A, A, A)	54

Zoznam tabuliek

3.1	Špecifikácia aditívnych modelov (ETS).	22
5.1	Odhady parametrov modelu $ETS(A, A_d, N)$	33
5.2	Hodnoty bodových predpovedí modelu $ETS(A, A_d, N)$	34
5.3	Porovnanie výsledkov pre pravdepodobnostné predpovede	36
5.4	Porovnanie výsledkov zo sto simulácií modelu $y_t = \log(t) + \epsilon_t$,	37
5.5	Odhady parametrov modelu $ETS(A, A_d, N)$	37
5.6	Hodnoty bodových predpovedí modelu $ETS(A, A_d, N)$	39
5.7	Porovnanie výsledkov pre pravdepodobnostné predpovede	41
5.8	Porovnanie výsledkov zo sto simulácií modelu $y_t = \log(t) + \epsilon_t$,	43
5.9	Odhady parametrov modelu $ETS(A, A_d, N)$	44
5.10	Hodnoty bodových predpovedí modelu $ETS(A, A_d, N)$	46
5.11	Porovnanie výsledkov pre pravdepodobnostné predpovede	47
5.12	Porovnanie výsledkov zo sto simulácií modelu $y_t = \log(t) + \epsilon_t$,	49
5.13	Odhady parametrov modelu $ETS(A, A, A)$	51
5.14	Hodnoty bodových predpovedí modelu $ETS(A, A, A)$ a príslušné	52
5.15	Porovnanie výsledkov pre pravdepodobnostné predpovede o 10	54