UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Vybrané metody analýzy přerušovaných časových řad

Katedra matematické analýzy a aplikací matematiky
Vedoucí práce: **doc. RNDr. Eva Fišerová, Ph.D.**
Vypracoval(a): **Bc. Klára Drastichová**
Studijní program: Aplikovaná matematika (N0541A170026)
Studijní obor: Aplikovaná matematika
Forma studia: prezenční
Rok odevzdání: 2024

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Klára Drastichová

Název práce: Vybrané metody analýzy přerušovaných časových řad

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Eva Fišerová, Ph.D.

Rok obhajoby práce: 2024

Abstrakt: Přerušované časové řady představují speciální typ časových řad,v jejichž průběhu se v konkrétním časovém okamžiku vyskytne nějaká intervence. Cílem diplomové práce je seznámení s problematikou přerušovaných časových řad a vyhodnocením efektu intervence. V teoretické části práce jsou popsány základní pojmy a zejména metoda segmentové regrese se známým i neznámým časem intervence. Dále je práce soustředěna na možnosti využití Boxovy-Jenkinsovy metodologie při analýze přerušovaných časových řad. V praktické části jsou získané teoretické poznatky aplikovány při analýze reálné časové řady.

Klíčová slova: přerušovaná časová řada, segmentová regrese, Boxova-Jenkinsova metodologie, intervence

Počet stran: 85

Počet příloh: 2

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Klára Drastichová

Title: Selected methods for the analysis of interrupted time series

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Eva Fišerová, Ph.D.

The year of presentation: 2024

Abstract: Interrupted time series are a special type of time series in which an intervention occurs at a specific point in time. The aim of this thesis is to introduce the subject of interrupted time series and to evaluate the effect of an intervention. The theoretical part of the thesis describes the basic concepts of interrupted time series with particular focus on the segmented regression method with known and unknown time of intervention. Furthermore, the thesis focuses on the possibilities of using the Box-Jenkins methodology in the analysis of interrupted time series. In the practical part, the obtained theoretical knowledge is applied in the analysis of a real time series.

Key words: interrupted time series, segmented regression, Box-Jenkins methodology, intervention

Number of pages: 85

Number of appendices: 2

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením paní doc. RNDr. Evy Fišerové, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Seznam obrázků

Poděkování

Ráda bych na tomto místě poděkovala paní doc. RNDr. Evě Fišerové, Ph.D. za její cenné rady, ochotu a čas, který mi během psaní diplomové práce věnovala. Děkuji také své rodině a přátelům za obrovskou podporu během celého studia.

Úvod

Tématem této diplomové práce jsou vybrané metody pro analýzu přerušovaných časových řad. Jedná se o speciální typ časových řad, v jejichž průběhu došlo k nějaké intervenci a průběh časové řady se tak mohl či nemusel změnit. Cílem této diplomové práce je představit několik přístupů k této problematice a ukázat analýzu přerušované časové řady na reálném příkladě.

Struktura práce je rozdělena do dvou hlavních částí, kterými jsou část teoretická a praktická. Teoretická část se věnuje třem tématům, a to základnímu popisu přerušovaných časových řad, detailnímu představení analýzy pomocí segmentové regrese a nakonec ARIMA modelům. Zaměřuje se také na situace, kdy se v časové řadě nachází více intervencí anebo kdy je potřeba tyto intervence identifikovat.

Praktická část práce se zabývá analýzou reálné přerušované časové řady, která se týká úrazovosti dětí v automobilech v Japonsku. Jako intervence je zde zkoumána změna legislativy o dětských bezpečnostních systémech v automobilech a cílem je analyzovat efekt této intervence. Při analýze jsou aplikovány teoretické přístupy a metody diskutované v teoretické části.

V obou hlavních částech této práce přistupujeme k analýze přerušovaných časových řad dvěma způsoby. Zaprvé, pokud známe přesný čas intervence, a zadruhé, pokud tuto informaci nemáme k dispozici. Přístup k analýze je v obou případech odlišný, a proto je znalost času intervence potřeba při analýze správně zohlednit.

Kapitola 1

Přerušované časové řady

Přerušované časové řady se řadí mezi speciální typ časových řad, které se využívají v odvětvích, jako například v medicíně, biologii či psychologii. Jsou považovány za nejsilnější kvaziexperimentální[1] přístup pro ohodnocení efektu intervence v průběhu času. V této kapitole se zaměříme na základní postup při analýze přerušovaných časových řad a na nejčastější přístupy k této problematice, jako je segmentová regrese a ARIMA modely. Ve stručnosti též popíšeme dekompozici časových řad na jednotlivé složky.

1.1. Úvod

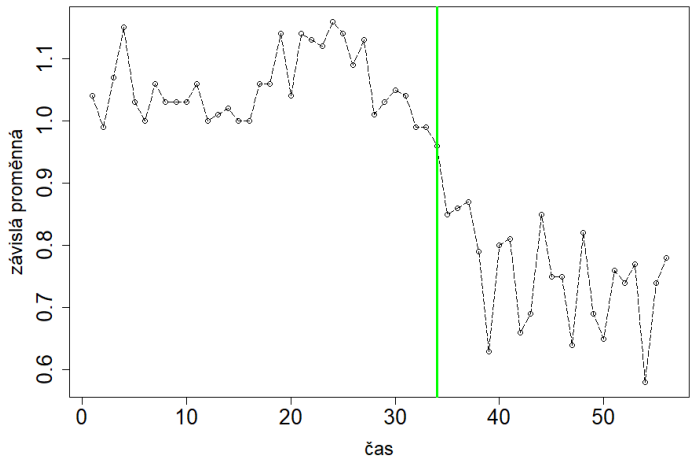
Model přerušovaných časových řad využíváme v situaci, kdy máme k dispozici data sledovaná v čase, v jejichž průběhu se v konkrétním časovém bodě vyskytne nějaká intervence nebo zásah. Cílem zkoumání takovýchto časových řad je analyzovat, zda má tento zásah vliv na následující průběh řady. Jinými slovy chceme zjistit, zda a jak se změní chování závisle proměnné po nastání tohoto konkrétního zásahu. Jako příklad můžeme uvést sledování počtu pacientů s rakovinou plic za rok před a po vydání zákazu kouření ve veřejných prostorech.

[1] Kvaziexperiment je typ výzkumu, který hledá kauzální vztahy mezi proměnnými. Od klasického experimentu se liší tím, že nevyžaduje randomizované rozdělení do skupin.

Hypotetickou přerušovanou časovou řadu bychom mohli zapsat jako

$$Y_1 \qquad Y_2 \qquad Y_3 \quad \dots \quad Y_i \qquad Y_{i+1} \qquad Y_{i+2} \quad \dots \quad Y_{T_n},$$

kde $Y_t,\ t = 1, \dots, T_n$, představuje pozorování časové řady v čase t, včetně hodnoty Y_i v konkrétním čase intervence.[15] Příklad takové řady lze vidět na obrázku 1.1, kde známe časový bod intervence (bod Y_i vyznačen zelenou svislou čarou). Zde bychom již od pohledu mohli odhadnout, že intervence měla na průběh časové řady vliv, jelikož po jejím nastání začaly hodnoty závislé proměnné viditelně klesat. Na to, jak tento efekt ohodnotit, se zaměříme v následujících kapitolách.



Obrázek 1.1: Přerušovaná časová řada s vyznačeným časem intervence (zeleně)

1.2. Postup při analýze přerušovaných časových řad

K analýze přerušovaných časových řad se musí přistupovat velmi individuálně, jelikož charakter každé časové řady může být jiný. Dá se ovšem držet několika základních kroků, které nyní popíšeme. [1]

1) Identifikace intervence: Přerušované časové řady vyžadují jasné rozlišení

segmentů před intervencí a po ní. Je tedy vhodné správně definovat období, kdy intervence nastala. Může se jednat jak o jednorázovou událost (např. vydání zákonu v konkrétní den) či o průběžnou změnu (např. zavření škol v době covidu - zde je nutné znát přesný začátek a konec tohoto období). V některých situacích může být tato identifikace poněkud náročná. Přerušované časové řady proto fungují nejlépe v situacích, kdy se po intervenci (či po nějakém období od intervence) očekává relativně rychlá změna.

2) Volba modelu: Klíčovým krokem při analýze časových řad je správný výběr modelu. Existuje několik možností, jak k analýze přerušovaných časových řad přistupovat. Při volbě modelu je nutné ověřit předpoklady pro použití jednotlivých metod. Mezi základní a nejpoužívanější přístupy patří:

- segmentová regrese
- ARIMA modely
- Praisova-Winstenova metoda (zobecněná metoda nejmenších čtverců)
- restringovaná metoda maximální věrohodnosti (REML)
- robustní přístupy k analýze přerušovaných časových řad

3) Odhad a interpretace parametrů: Pro správnou interpretaci výsledků je u každého modelu třeba odhadnout parametry, a ty následně interpretovat. Způsob odhadu se bude lišit na základě toho, jakou z metod použijeme.

4) Ohodnocení efektu intervence: Stěžejním krokem analýzy přerušovaných časových řad je ohodnotit, zda intervence měla na následující průběh časové řady vliv. V následujících kapitolách popíšeme vyjádření efektu intervence pomocí segmentové regrese a ARIMA modelů.

5) Predikce: Jako poslední krok nás často zajímají predikce hodnot závisle proměnné v budoucím čase. To lze predikovat také u přerušovaných časových řad na základě použitého modelu.

1.3. Dekompozice časových řad

Na úvod je také potřeba uvést, že časová řada se může skládat z následujících složek:

- trendová složka Tr_t
- sezónní složka S_t
- cyklická složka C_t
- náhodná složka ϵ_t

V této souvislosti mluvíme o dekompozici časové řady, jelikož časovou řadu můžeme do těchto čtyř složek rozložit. Existují dva základní způsoby, a to aditivní dekompozice, při které jednotlivé složky sčítáme:

$$Y_t = Tr_t + S_t + C_t + \epsilon_t, \quad t = 1, \ldots, T_n,$$

a multiplikativní dekompozice, při které složky násobíme:

$$Y_t = Tr_t \times S_t \times C_t \times \epsilon_t, \quad t = 1, \ldots, T_n.$$

Trendová složka vyjadřuje dlouhodobé chování časové řady, tedy popisuje trend v celé délce jejího průběhu. Sezónní složka naopak popisuje periodické změny v časové řadě, které se v jejím průběhu opakují (po jednotlivých sezónách, jako např. týden, měsíc či čtvrtletí). Cyklická složka vyjadřuje dlouhodobé kolísání kolem trendu (po dobu delší než jeden rok) a náhodná složka zachycuje všechny ostatní vlivy působící na časovou řadu, které nejsme schopni nijak vyjádřit. [10]

U náhodné složky požadujeme tři základní předpoklady, a to aby měla nulovou střední hodnotu:

$$E(\epsilon_t) = 0, \ \forall t = 1, \ldots, T_n,$$

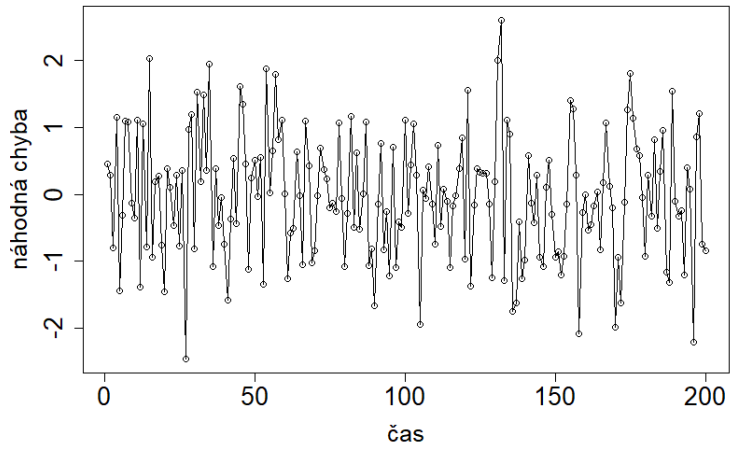
konstantní rozptyl:

$$\mathrm{var}(\epsilon_t) = \sigma^2, \ \forall t = 1, \ldots, T_n,$$

a aby byly její hodnoty vzájemně nekorelované:

$$\mathrm{cov}(\epsilon_i, \epsilon_j) = 0, \ \forall i \neq j.$$

Pokud jsou tyto předpoklady splněny, nazýváme takovýto proces bílý šum. Ukázku bílého šumu můžeme vidět na obrázku 1.2.



Obrázek 1.2: Příklad bílého šumu

Kapitola 2

Segmentová regrese

V této kapitole se zaměříme na jeden z nejpoužívanějších přístupů k přerušovaným časovým řadám, kterým je analýza pomocí segmentové regrese. Kapitolu rozdělíme na dvě části na základě toho, zda známe časový bod zlomu či nikoliv. V obou těchto částech popíšeme matematický model a jeho interpretaci. Zaměříme se také na segmentovou regresi v softwaru R.

2.1. Segmentová regrese se znalostí času intervence

Segmentová regrese je model, který popisuje závislost vysvětlované proměnné na čase pomocí po částech lineární funkce. Časový bod, ve kterém nastává intervence, nazýváme bod zlomu (neboli change point). V této kapitole budeme předpokládat, že tento časový bod intervence je známý. Cílem je pak proložit data na základě metody nejmenších čtverců, a to jak v části před bodem zlomu, tak po něm (tyto dvě části budeme nazývat segmenty). Předpokládáme tak lineární závislost mezi vysvětlovanou proměnnou a časem v každém segmentu.

2.1.1. Model a odhad regresních koeficientů

V modelu segmentové regrese se znalostí intervence se bude vyskytovat vždy jedna vysvětlovaná/závislá proměnná a následující tři vysvětlující/nezávislé proměnné:

- t ... spojitá proměnná vyjadřující dobu (např. dny, měsíce, roky,...) uplynulou od počátku sledovaného období, nabývající hodnot 1,2,...,T_n
- *intervence* ... binární umělá proměnná, která přiřazuje číslo 0 pozorováním před intervencí (tedy v časech $1, \ldots, T_i - 1$) a číslo 1 pozorováním v časech $T_i, \ldots, T_n$
- *cas od intervence* ... spojitá proměnná, která indikuje čas od nastání intervence (pozorováním před bodem zlomu přiřazujeme 0, od bodu zlomu pak 1,2,...,$T_n - T_i + 1$)

Pomocí takto zavedených proměnných lze psát model segmentové regrese jako

$$Y_t = \mu + \alpha t + \beta \times intervence_t + \gamma \times cas\ od\ intervence_t + \epsilon_t\ , \tag{2.1}$$

kde Y_t vyjadřuje hodnotu závislé proměnné v čase t, koeficienty $\mu, \alpha, \beta, \gamma$ představují neznámé regresní koeficienty a ϵ_t chybový člen, který zahrnuje variabilitu nevysvětlenou modelem. I zde platí základní předpoklad lineární regrese, a tedy že chyby musí tvořit bílý šum. [1, 6]

Model (2.1) můžeme vyjádřit také pomocí maticového zápisu [25]:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

kde

- $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_{T_n})'$ je vektor pozorovaných hodnot závisle proměnné v časových bodech 1,2,...,T_n

- $\boldsymbol{\beta} = (\mu, \alpha, \beta, \gamma)'$ symbolizuje vektor regresních koeficientů
- $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_{T_n})'$ představuje vektor náhodných chyb všech pozorování, pro který platí $E(\boldsymbol{\epsilon})=\mathbf{0}$ a $\text{var}(\boldsymbol{\epsilon})=\sigma^2\boldsymbol{I}$
- matice $\boldsymbol{X}$ vypadá následovně:

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & T_i & 1 & 1 \\ 1 & T_i+1 & 1 & 2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & T_n & 1 & T_n - T_i + 1 \end{pmatrix}.$$

Příklad dat se všemi výše zmíněnými proměnnými by mohl vypadat následovně (tučným písmem je vyznačen časový bod, ve kterém nastala intervence, tedy bod zlomu):

t	Y_t	*intervence*	*cas od intervence*
...	...	...	...
85	28,09	0	0
86	30,61	0	0
87	29,52	0	0
88	27,67	0	0
89	24,58	1	1
90	24,01	1	2
91	21,76	1	3
92	27,85	1	4
93	28,32	1	5
...	...	...	...

Odhad koeficientů $\boldsymbol{\beta}$ v regresním modelu (2.2) provedeme na základě metody nejmenších čtverců [9] pomocí vzorce

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

kde $\boldsymbol{X}$ a $\boldsymbol{Y}$ představují matici a vektor popsané výše. Lze odvodit střední hodnotu tohoto odhadu

$$E(\widehat{\boldsymbol{\beta}}) = E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

a také jeho varianční matici

$$\begin{aligned}\mathrm{var}(\widehat{\boldsymbol{\beta}}) &= \mathrm{var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\mathrm{var}(\boldsymbol{Y})]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\sigma^2\boldsymbol{I}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

Při těchto úpravách jsme využili toho, že pro střední hodnotu vektoru $\boldsymbol{Y}$ platí následující

$$E(\boldsymbol{Y}) = E(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{X}\boldsymbol{\beta} + E(\boldsymbol{\epsilon}) = \boldsymbol{X}\boldsymbol{\beta}$$

a pro varianční matici vektoru $\boldsymbol{Y}$ platí vztah

$$\mathrm{var}(\boldsymbol{Y}) = \mathrm{var}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}.$$

Po odhadnutí vektoru $\boldsymbol{\beta}$ můžeme vypočítat také očekávanou hodnotu v konkrétním čase T_0 a poté její interval spolehlivosti. Pro očekávanou hodnotu $\widehat{Y}_0$ v čase T_0 za předpokladu normality náhodných chyb platí

$$\widehat{Y}_0 = \boldsymbol{x_0}'\widehat{\boldsymbol{\beta}} \sim N_1\big(\boldsymbol{x_0}'\boldsymbol{\beta}, \sigma^2\boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}\big),$$

kde vektor $\boldsymbol{x_0}$ závisí na tom, v jakém segmentu se čas T_0 nachází:

$$\boldsymbol{x_0} = \begin{cases} (1, T_0, 0, 0)' & \text{pokud } T_0 < T_i \\ (1, T_0, 1, T_0 - T_i + 1)' & \text{pokud } T_0 \geq T_i. \end{cases}$$

Na základě výše odvozeného bodového odhadu a jeho rozptylu lze vypočítat meze $100(1-\alpha)\%$ intervalu spolehlivosti pro $E(Y_0)$ jako

$$\boldsymbol{x_0}'\widehat{\boldsymbol{\beta}} \pm u(1 - \frac{\alpha}{2})\sigma\sqrt{\boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}},$$

kde předpokládáme, že je parametr σ^2 známý a $u(1-\frac{\alpha}{2})$ představuje hodnotu kvantilu normálního rozdělení. Pokud by byl parametr σ^2 neznámý, pracujeme místo normálního rozdělení se Studentovým t-rozdělením a meze intervalu spolehlivosti by byly následující:

$$\boldsymbol{x_0}'\widehat{\boldsymbol{\beta}} \pm t_{n-4}(1-\frac{\alpha}{2})\widehat{\sigma}\sqrt{\boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}}.$$

Odhad parametru σ^2 provádíme na základě reziduálního vektoru $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ pomocí vztahu

$$\widehat{\sigma}^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-4},$$

kde n představuje počet měření. Pokud bychom za čas T_0 dosazovali různé hodnoty, vznikne při spojitě se měnícím čase tzv. pás spolehlivosti kolem regresní přímky. Nelze tvrdit, že tento pás spolehlivosti pokrývá celou regresní přímku, jelikož uvedená spolehlivost $1-\alpha$ platí pro každý vybraný čas zvlášť. [5]

Nakonec zaveďme také předpovědní interval, který představuje interval spolehlivosti pro budoucí hodnotu Y_0 v čase T_0, kde $T_0 > T_n$. Predikovaná budoucí hodnota $\widehat{Y}_0$ vychází z odhadnutého modelu a získáme ji opět jako $\widehat{Y}_0 = \boldsymbol{x_0}'\widehat{\boldsymbol{\beta}}$. Meze $100(1-\alpha)\%$ předpovědního intervalu se známým parametrem σ^2 vypočítáme pomocí vztahu

$$\boldsymbol{x_0}'\widehat{\boldsymbol{\beta}} \pm u(1-\frac{\alpha}{2})\sigma\sqrt{1+\boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}}.$$

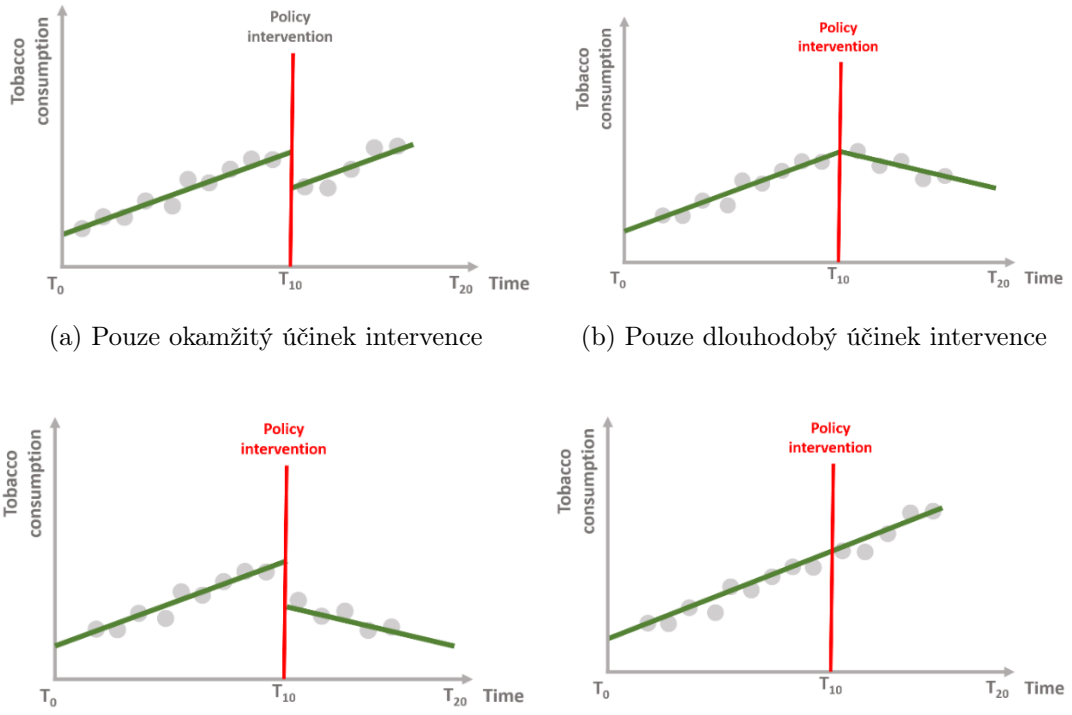
V případě neznámého σ bychom opět kvantil normálního rozdělení nahradili kvantilem Studentova t-rozdělení a neznámé σ jeho odhadem.

2.1.2. Interpretace regresních koeficientů a efektu intervence

Nejprve je třeba si uvědomit, že intervence může mít na časovou řadu dva různé typy efektů. Prvním je okamžitý účinek (neboli skoková změna v čase) a druhým dlouhodobý účinek (změna ve směrnici obou segmentů). Pro přerušo-

vanou časovou řadu proto můžou nastat následující čtyři základní situace, které jsou vyobrazené na obrázcích 2.1(a) až 2.1(d):

- pouze okamžitý účinek
- pouze dlouhodobý účinek
- okamžitý i dlouhodobý účinek
- žádný účinek



(a) Pouze okamžitý účinek intervence

(b) Pouze dlouhodobý účinek intervence

(c) Okamžitý i dlouhodobý účinek intervence

(d) Žádný účinek intervence

Obrázek 2.1: Možné efekty intervence, zdroj: [6]

Nyní již lze interpretovat regresní koeficienty výše zmíněného modelu (2.1). Pro lepší interpretaci však tento model nyní vyjádříme pomocí indikátorových proměnných. Ty rozlišují, zda se nacházíme v segmentu před bodem zlomu T_i

(tedy $I(t \geq T_i) = 0$) nebo v segmentu za ním (tedy $I(t \geq T_i) = 1$):

$$Y_t = \mu + \alpha t + \beta \times I(t \geq T_i) + \gamma(t - T_i) \times I(t \geq T_i) + \epsilon_t,\ t = 1, \ldots, T_n. \quad (2.3)$$

Tento model můžeme rozepsat pro tři situace, a to pro přímku před intervencí, po intervenci a pro bod v čase intervence. V prvním segmentu, tedy v části před intervencí, bude platit předpis

$$Y_t = \mu + \alpha t + \epsilon_t,\ t = 1, \ldots, T_i - 1.$$

Koeficienty μ a α v prvním segmentu interpretujeme stejně jako koeficienty základního lineárního modelu. Absolutní člen μ odhaduje hodnotu závisle proměnné v časovém bodě 0 a koeficient α říká, o kolik se v průměru změní hodnota závisle proměnné při jednotkové změně času. Pro přímku druhého segmentu platí předpis

$$Y_t = \mu + \alpha t + \beta + \gamma(t - T_i) + \epsilon_t,\ t = T_i, \ldots, T_n, \quad (2.4)$$

kde koeficient β vyjadřuje okamžitý účinek, tedy rozdíl funkčních hodnot dvou přímek v čase intervence. Pokud bude koeficient β nulový, okamžitý účinek se nebude vyskytovat a graf bude spojitý (půjde o situaci na obrázcích 2.1(b) nebo 2.1(d)). Jestliže bude nenulový, bude vyjadřovat velikost skoku v čase T_i. Graf funkce může mít v každém bodě samozřejmě pouze jednu funkční hodnotu, kterou bude v čase T_i hodnota dosazená do přímky druhého segmentu:

$$Y_{T_i} = \mu + \beta + \alpha T_i + \epsilon_{T_i}.$$

Tuto hodnotu bychom mohli nazvat absolutním členem druhé přímky, jelikož pro druhý segment je počáteční (neboli nulovou) hodnotou právě čas T_i.

V předpisu přímky pro druhý segment (2.4) se vyskytuje také koeficient γ, který vyjadřuje dlouhodobý účinek, neboli o kolik se liší směrnice přímky v druhé

části oproti směrnici přímky první části. Pokud bychom chtěli vyjádřit sklon přímky v druhém segmentu, vypočítáme jej jako součet koeficientů (α+γ), k čemuž dojdeme pomocí následujícího odvození. Předpis druhé přímky (2.4) upravíme

$$Y_t = \mu + \alpha(t + T_i - T_i) + \beta + \gamma(t - T_i) + \epsilon_t,\ t = T_i, \ldots, T_n,$$

a následně přeuspořádáme první závorku

$$Y_t = \mu + \alpha(T_i + (t - T_i)) + \beta + \gamma(t - T_i) + \epsilon_t,\ t = T_i, \ldots, T_n.$$

Po roznásobení a přeuspořádání pak dostáváme

$$Y_t = \mu + \alpha T_i + \beta + (\alpha + \gamma) \times (t - T_i) + \epsilon_t,\ t = T_i, \ldots, T_n,$$

z čehož můžeme vyčíst, že součet $(\alpha + \gamma)$ vyjadřuje směrnici druhé přímky. Opět vidíme, že její absolutní člen v čase T_i je $(\mu + \alpha T_i + \beta)$. [6]

Efekt intervence se tedy dá vyjádřit pomocí regresních koeficientů. Další způsob posouzení efektu intervence spočívá v porovnání predikované funkční hodnoty segmentové regrese s hodnotou, kterou bychom predikovali v situaci, kdyby žádná intervence nenastala.[15] Je tedy potřeba predikovat budoucí hodnoty pro přímku prvního segmentu, a ty následně srovnat s odhadnutou hodnotou segmentové regrese v druhém segmentu. Porovnání lze provést absolutně nebo relativně, a to v jakémkoliv časovém bodě po intervenci. Záleží na tom, zda chceme zjistit rozdíl těsně po intervenci či ve vzdálenějším časovém indexu. [23]. Absolutní efekt intervence vypočítáme jako rozdíl

$$\widehat{Y}_{(s\ intervenci)} - \widehat{Y}_{(bez\ intervence)}$$

a relativní změnu bychom vyjádřili následovně

$$\frac{\widehat{Y}_{(s\ intervenci)} - \widehat{Y}_{(bez\ intervence)}}{\widehat{Y}_{(bez\ intervence)}}.$$

2.1.3. Test existence bodu zlomu

V situaci, kdy máme odhadnuté parametry pro přímky v obou segmentech, je vhodné otestovat, zda má vůbec takto složitý model smysl. Někdy totiž intervence nemusí mít na průběh časové řady významný vliv, a je potom lepší použít pouze jednoduchý model lineární regrese (jednu přímku pro celou časovou řadu). Pro toto otestování slouží Chowův test, který je založený na principu F–statistiky. Testujeme, zda by nebylo lepší přejít k jednoduššímu modelu s méně koeficienty, tzv. podmodelu, či zda je vhodné data proložit dvěma přímkami na základě segmentové regrese. Hypotézy formulujeme následovně:

H_0: model před a po intervenci je stejný (jedna regresní přímka pro celou časovou řadu)

H_A: model po intervenci je jiný než před intervencí (dvě regresní přímky).

Při výpočtu testové statistiky využijeme reziduální součty čtverců, které získáme sečtením druhých mocnin reziduí. Potřebujeme reziduální součet pro celkovou přímku podmodelu (označíme RSS_c) a také reziduální součty pro jednotlivé přímky obou segmentů zvlášť (označíme RSS_1 a RSS_2). Testovou statistiku pak vypočítáme jako

$$F = \frac{(RSS_c - (RSS_1 + RSS_2))/k}{(RSS_1 + RSS_2)/(n - 2k)},$$

kde k je počet parametrů v modelu před intervencí ($k = 2$) a n celkový počet bodů časové řady. Testová statistika se za platnosti nulové hypotézy řídí F-rozdělením o $(k, n - 2k)$ stupních volnosti. S $(1 - \frac{\alpha}{2})$ kvantilem tohoto rozdělení srovnáme výslednou hodnotu F-statistiky a na hladině významnosti α rozhodneme o zamítnutí či nezamítnutí nulové hypotézy.

2.1.4. Předpoklady a úskalí segmentové regrese

Jak již bylo zmíněno, pro jakoukoliv analýzu přerušovaných časových řad je potřeba dostatečně velké množství dat. To platí také při použití segmentové regrese, kde je vhodné mít dostatek pozorování ve všech segmentech časové řady, tedy před i po intervenci (neexistují však žádné konkrétní limity pro jejich počet). Kvalitu regrese totiž neovlivňuje pouze množství dat, ale také další faktory. Mezi ně patří například variabilita v datech, rozdělení pravděpodobnosti pro data před a po intervenci či výskyt sezónnosti.

Další vhodnou podmínkou pro analýzu časových řad pomocí segmentové regrese je přesná znalost nastání intervence. To může být považováno za úskalí tohoto přístupu, jelikož v některých situacích si o přesném času intervence nemusíme být jisti. S takovou situací si v softwaru R umí poradit např. knihovna `segmented`, které se budeme podrobněji věnovat v kapitole 2.2.

Jako poslední omezení lze uvést fakt, že všechny segmenty časové řady prokládáme pouze lineárním trendem (některé složitější situace by mohly vyžadovat jiný než lineární model). Analýzu pomocí segmentové regrese tedy můžeme popsat jako jednoduchý a účinný přístup k přerušovaným časovým řadám, který však má několik omezení. [1, 23]

2.1.5. Přerušované časové řady s více intervencemi

Často se můžeme setkat také se situací, kdy se během jedné časové řady vyskytne více než jedna intervence. Může se jednat například o situaci, kdy nás zajímá efekt opatření zavedených v různých časových bodech. Dále se může jednat o případ, kdy je intervence zavedena a později zrušena, přičemž v obou těchto časových bodech nás zajímá účinek intervence na průběh časové řady.

Model pro segmentovou regresi se dvěma intervencemi T_{i_1} a T_{i_2}, kde $T_{i_1} < T_{i_2}$, zapíšeme pomocí binárních proměnných $intervence1$ a $intervence2$ jako

$$Y_t = \mu + \alpha t + \beta \times intervence1_t + \gamma \times cas\ od\ intervence1_t + \psi \times intervence2_t + \phi \times cas\ od\ intervence2_t + \epsilon_t,\ t = 1, \ldots, T_n,$$

kde

$$intervence1 = \begin{cases} 1 & \text{pro } t = T_{i_1}, \ldots, T_{i_2} - 1 \\ 0 & \text{jinak,} \end{cases}$$

$$intervence2 = \begin{cases} 1 & \text{pro } t = T_{i_2}, \ldots, T_n \\ 0 & \text{jinak.} \end{cases}$$

V modelu máme nyní dva nové koeficienty ψ a ϕ, které odpovídají druhému bodu intervence. Interpretace koeficientů je analogická jako u modelu s jedním bodem zlomu (viz kapitola 2.1.2). Takto popsaný model bychom mohli zobecnit na jakýkoliv větší počet bodů intervence, což už by ale vyžadovalo poměrně dlouhou časovou řadu s dostatečným množstvím dat. [14, 16]

2.2. Segmentová regrese bez znalosti času intervence

Tato kapitola bude věnována podobné problematice jako doposud, a to analýze přerušovaných časových řad s intervencí pomocí segmentové regrese. Nyní ale budeme předpokládat situaci, ve které sice víme, že během určitého časového období intervence nastala, nevíme ovšem kdy přesně. Doposud jsme znalost přesného časového bodu intervence při analýze předpokládali, nyní se zaměříme na jeho detekci. Tato situace odpovídá balíčku `segmented` v softwaru R, jehož algoritmus v této kapitole popíšeme. Celá tato kapitola vychází ze zdrojů [4, 16, 17].

2.2.1. Model a detekce intervence

Pro modelování segmentové regrese bez znalosti času intervence zavedeme následující reparametrizaci vztahu mezi závisle a nezávisle proměnnou (v našem

případě čas t). Budeme modelovat střední hodnotu závisle proměnné $E(Y)$ v závislosti na čase t pomocí následujícího nelineárního vztahu:

$$\mu + \alpha t + \gamma(t - T_i) \times I(t \geq T_i),\ t = 1, \ldots, T_n. \tag{2.5}$$

Symbol T_i zde označuje hledaný bod zlomu a $I(t \geq T_i)$ opět symbolizuje indikátorovou proměnnou, která rozlišuje, ve kterém segmentu časové řady se nacházíme. Jedná se o nelineární model, jelikož po roznásobení dostáváme v modelu součin dvou neznámých koeficientů γ a T_i.

Zápis pro indikátorovou proměnnou nyní zjednodušme pomocí symbolu $+$, tedy místo $(t - T_i) \times I(t \geq T_i)$ nyní budeme psát $(t - T_i)_+$, pro které platí

$$(t - T_i)_+ = \begin{cases} (t - T_i) & \text{pokud } t \geq T_i \\ 0 & \text{jinak.} \end{cases}$$

Model (2.5) tedy můžeme přepsat jako

$$\mu + \alpha t + \gamma(t - T_i)_+,\ t = 1, \ldots, T_n. \tag{2.6}$$

Dále se v modelu vyskytuje neznámý koeficient α, který symbolizuje velikost sklonu přímky v prvním segmentu a koeficient γ, který značí změnu sklonu přímky druhého segmentu oproti směrnici přímky prvního segmentu. Směrnici druhé přímky získáme opět sečtením $(\alpha + \gamma)$, jako tomu bylo při znalosti intervence. V tomto modelu se však nevyskytuje koeficient β, jelikož knihovna `segmented` neumožňuje hodnocení okamžitého efektu, tudíž půjde vždy o spojitou lomenou přímku (přímky na sebe budou v bodě intervence navazovat).

Část $(t - T_i)_+$ v modelu (2.6) lze dále aproximovat pomocí Taylorova rozvoje prvního řádu kolem počátečního času intervence $T_i^{(0)}$ jako

$$(t - T_i)_+ = (t - T_i^{(0)})_+ + (T_i - T_i^{(0)})(-I(t \geq T_i^{(0)})),$$

kde $(-I(t \geq T_i^{(0)})$ vyjadřuje první derivaci výrazu $(t - T_i)_+$ v bodě $T_i^{(0)}$, jelikož

$$\left.\frac{\partial\big((t-T_i)_+\big)}{\partial T_i}\right|_{T_i=T_i^{(0)}} = -1 \times I(t \geq T_i^{(0)}).$$

Tuto úpravu pomocí Taylorova rozvoje můžeme zpětně dosadit do původně nelineárního modelu (2.6), čímž dostaneme model, který bude lineární v parametrech, jelikož se zde již nevyskytuje součin neznámých koeficientů ($T_i^{(0)}$ zde považujeme za známou počáteční hodnotu):

$$\mu + \alpha t + \gamma(t - T_i^{(0)})_+ - \delta \times I(t \geq T_i^{(0)}),\ t = 1, \ldots, T_n.$$

Nyní máme vše připravené pro detekci bodu zlomu, která bude probíhat pomocí následujícího iteračního algoritmu [16]:

1. V každém kroku $s+1$, $s = 0, 1, \ldots$, dosadíme bod $T_i^{(s)}$ do modelu

$$\mu + \alpha t + \gamma(t - T_i^{(s)})_+ + \delta(-I(t \geq T_i^{(s)}) \tag{2.7}$$

   s tím, že pro první krok potřebujeme zadat počáteční čas $T_i^{(0)}$. Ten zvolíme buď na základě odhadu (v některých případech můžeme tušit, kde intervence přibližně nastala) nebo zvolíme počáteční bod zlomu v prostředním bodě časové řady.

2. Neznámé koeficienty tohoto modelu odhadneme pomocí metody nejmenších čtverců, čímž získáme odhady $\widehat{\mu}, \widehat{\alpha}, \widehat{\gamma}$ a $\widehat{\delta}$.

3. Nyní se pokusíme zlepšit odhad bodu zlomu na základě výše odhadnutých koeficientů. Pokud si rozepíšeme předpisy přímek pro oba segmenty v s-tém kroku podle vzorce (2.7), dostáváme

$$\begin{gathered}\text{před intervencí: } \mu + \alpha t, \quad t = 1, \ldots, T_i^{(s)} - 1, \\ \text{po intervenci: } \mu + \alpha t + \gamma(t - T_i^{(s)}) - \delta, \quad t = T_i^{(s)}, \ldots, T_n.\end{gathered}$$

Přímky dáme nyní do rovnosti a za čas t dosadíme čas intervence $T_i^{(s+1)}$, který budeme chtít vyjádřit. Dostáváme

$$\mu + \alpha T_i^{(s+1)} = \mu + \alpha T_i^{(s+1)} + \gamma(T_i^{(s+1)} - T_i^{(s)}) - \delta$$

$$\gamma T_i^{(s+1)} = \gamma T_i^{(s)} + \delta$$

Nakonec vyjádříme $T_i^{(s+1)}$ a za koeficienty δ a γ dosadíme jejich odhady:

$$T_i^{(s+1)} = T_i^{(s)} + \frac{\widehat{\delta}}{\widehat{\gamma}}.$$

4. Opakujeme kroky 1. – 3. až do nastání konvergence. Za zastavovací kritérium zde považujeme to, když je intervence v kroku (s+1) téměř shodná s intervencí v s-tém kroku, tedy $\widehat{\delta} \approx 0$. Pro nalezený bod zlomu nakonec můžeme psát $\widehat{T}_i \equiv T_i^{(s)}$.

Pro odhadnuté $\widehat{T}_i$ lze dále vypočítat standardní chybu $SE(\widehat{T}_i)$ pomocí lineární aproximace podílu odhadů koeficientů $\widehat{\delta}$ a $\widehat{\gamma}$ (tzv. Delta metoda) [16]. Mějme vektor odhadů parametrů $(\widehat{\delta}, \widehat{\gamma})'$ a jejich varianční matici. Odhad varianční matice jejich podílu provedeme následovně

$$\widehat{\text{var}}\left(\frac{\widehat{\delta}}{\widehat{\gamma}}\right) = \left(\frac{\partial(\widehat{\delta}/\widehat{\gamma})}{\partial\widehat{\delta}}, \frac{\partial(\widehat{\delta}/\widehat{\gamma})}{\partial\widehat{\gamma}}\right) \times \text{var}\left(\frac{\widehat{\delta}}{\widehat{\gamma}}\right) \times \left(\frac{\partial(\widehat{\delta}/\widehat{\gamma})}{\partial\widehat{\delta}}, \frac{\partial(\widehat{\delta}/\widehat{\gamma})}{\partial\widehat{\gamma}}\right)'.$$

Vypočítáme parciální derivace a rozepíšeme varianční strukturu

$$\left(\frac{1}{\widehat{\gamma}}, \frac{\widehat{\delta}}{\widehat{\gamma}^2}\right) \times \begin{pmatrix} \text{var}(\widehat{\delta}) & \text{cov}(\widehat{\delta}, \widehat{\gamma}) \\ \text{cov}(\widehat{\gamma}, \widehat{\delta}) & \text{var}(\widehat{\gamma}) \end{pmatrix} \times \left(\frac{1}{\widehat{\gamma}}, \frac{\widehat{\delta}}{\widehat{\gamma}^2}\right)'.$$

Po roznásobení dostáváme

$$\frac{1}{\widehat{\gamma}^2}\text{var}(\widehat{\delta}) + \frac{\widehat{\delta}}{\widehat{\gamma}^3}\text{cov}(\widehat{\gamma}, \widehat{\delta}) + \frac{\widehat{\delta}}{\widehat{\gamma}^3}\text{cov}(\widehat{\delta}, \widehat{\gamma}) + \frac{\widehat{\delta}^2}{\widehat{\gamma}^4}\text{var}(\widehat{\gamma}),$$

což nakonec upravíme a pro získání $SE(\widehat{T}_i)$ odmocníme a dostaneme

$$\widehat{SE}(\widehat{T}_i) = \sqrt{\frac{\text{var}(\widehat{\delta}) + 2(\widehat{\delta}/\widehat{\gamma})\ \text{cov}(\widehat{\delta},\widehat{\gamma}) + (\widehat{\delta}/\widehat{\gamma})^2\ \text{var}(\widehat{\gamma})}{\widehat{\gamma}^2}}.$$

Na základě tohoto výpočtu lze za předpokladu normality definovat 95% interval spolehlivosti pro bod zlomu jako

$$\big(\widehat{T}_i - 1.96 \times \widehat{SE}(\widehat{T}_i)\ ,\ \widehat{T}_i + 1.96 \times \widehat{SE}(\widehat{T}_i)\big).$$

Alternativní přístup k detekci intervence

Druhý způsob, jak nalézt v časové řadě bod zlomu, je s využitím F-testu, který jsme zmínili již v kapitole o testu existence bodu zlomu, viz 2.1.3. Nyní jej použijeme k nalezení bodu intervence, a to tak, že v každém bodě (kromě několika krajních bodů) vypočítáme hodnotu této F-statistiky a najdeme ten časový bod, ve kterém bude testová statistika největší. Tato hodnota totiž bude nejvíce odpovídat alternativní hypotéze, a tedy se tam bude nejpravděpodobněji vyskytovat největší změna v trendu. Podrobněji tento přístup popíšeme v praktické části při analýze reálné časové řady. [24]

2.2.2. Model s více intervencemi

Opět se zaměříme také na situaci, kdy se v časové řadě vyskytuje více než jedna intervence s tím, že nevíme, v jakých časových bodech. Vyjdeme z modelu (2.6), který zobecníme na případ K bodů zlomu $T_{i_1} \dots T_{i_K}$:

$$\mu + \alpha t + \sum_{k=1}^{K} \gamma_k (t - T_{i_k})_+. \tag{2.8}$$

Parametr α v modelu vyjadřuje sklon přímky v segmentu $t < T_{i_1}$ a koeficienty $\gamma_k,\ k = 1, \dots, K$, symbolizují změnu sklonu mezi segmenty před a za k-tým

bodem zlomu T_{i_k}. Pro každý segment $T_{i_{k^*}} < t < T_{i_{k^*+1}}$ lze dopočítat také velikost sklonu pomocí součtu $\alpha + \sum_k^{k^*} \gamma_k$. [4]

Při detekci bodů zlomu můžeme v softwaru R opět využít knihovnu `segmented`, která umí pracovat i s více než jednou intervencí. Algoritmus je založen na stejném principu jako při jedné intervenci (viz 2.2.1), pouze s tím rozdílem, že v každém kroku odhadujeme více parametrů, a to $\mu, \alpha, \gamma_k, \delta_k$ pro $k = 1, \ldots, K$.

2.2.3. Test dlouhodobého účinku intervence

Pokud hledáme bod zlomu, je také dobré otestovat, zda se v časové řadě vůbec nějaká intervence vyskytuje. S ohledem na podkapitolu 2.2.1 se zaměříme na zmíněný model (2.6), konkrétně na koeficient γ, který vyjadřuje změnu sklonu přímky druhého segmentu oproti přímce prvního segmentu. Pokud by se v časové řadě žádný bod zlomu nevyskytoval, byl by koeficient γ roven nule. Nulová a alternativní hypotéza pro otestování existence T_i tedy bude následující:

$$H_0 : \gamma = 0$$
$$H_A : \gamma \neq 0.$$

Vzhledem k nedostatku znalostí o rozdělení pravděpodobností za předpokladu nulové hypotézy, nelze k otestování hypotézy použít klasické statistické testy. Jednou z možností, která je využívána také v knihovně `segmented`, je Daviesův test (v R `davies.test`).[4] Jedná se o asymptotický test, který pro K pevně zvolených bodů zlomu $T_{i_1} < T_{i_2} < \cdots < T_{i_K}$ napočítá K hodnot Waldovy testové statistiky pro γ_k jako

$$S(T_{i_k}) = \frac{\widehat{\gamma_k}}{\mathrm{SE}(\widehat{\gamma_k})}.$$

Tyto hodnoty $\{S(T_{i_k})\}_{k=1,\ldots,K}$ mají pro pevné T_{i_k} standardizované normální rozdělení. S využitím těchto testových statistik vypočítá p-hodnotu na základě vzorce

$$\text{p-hodnota} \approx \Phi(-M) + V\exp(-M^2/2)(8\pi)^{-1/2},$$

kde $\Phi(.)$ symbolizuje standardizované normální rozdělení, $M = max\{S(T_{i_k})\}_{k=1,\dots,K}$ značí maximum hodnot testové statistiky a V označuje celkový rozptyl hodnot $\{S(T_{i_k})\}_{k=1,\dots,K}$. [17]

Je nutné poznamenat, že Daviesův test je vhodný nástroj pro otestování existence bodu zlomů, ale není dobré jej používat pro určení počtu bodů zlomu. Na tuto problematiku se zaměříme v následující podkapitole.

2.2.4. Identifikace počtu bodů zlomu

K identifikaci toho, kolik bodů zlomu se v časové řadě vyskytuje, můžeme využít dva přístupy - výběr modelu pomocí informačních kritérií nebo sekvenční testování hypotéz.

Informační kritéria: Informační kritéria slouží k porovnání různých modelů. Lze použít například Akaikeho informační kritérium (AIC) či Bayesovské informační kritérium (BIC), pro něž platí, čím menší hodnota, tím lepší model. [5] Za naším účelem vytvoříme modely s různým počtem bodů zlomu a na základě těchto kritérií poté vybereme ten nejlepší. Tento výběr modelu nám pak určí, kolik intervencí se v časové řadě vyskytuje. Kritérium AIC odpovídá následujícímu vztahu

$$AIC = -2\ \ln(L) + 2s, \tag{2.9}$$

kde L označuje maximální hodnotu věrohodnostní funkce modelu a s značí počet odhadovaných parametrů v modelu. Druhou možností je kritérium BIC, které silněji penalizuje počet parametrů v modelu a vypočítáme jej následovně

$$BIC = -2\ \ln(L) + s\ ln(n).$$

Sekvenční testování hypotéz: Sekvenční testování hypotéz spočívá na podobném principu. Postupně vyzkoušíme různé počty bodů zlomu a testujeme hypotézy pomocí výše zmíněného Daviesova testu.

Začneme hypotézou s jedním bodem zlomu, která vypadá následovně:

$$H_0 : \gamma_1 = 0 \ (K = 0)$$
$$H_A : \gamma_1 \neq 0 \ (K > 1).$$

Pokud hypotézu na základě Daviesova testu zamítneme (p-hodnota bude menší než hladina významnosti 0.05), zvýšíme K o jedna a pokračujeme v testování:

$$H_0 : \gamma_2 = 0 \ (K = 1)$$
$$H_A : \gamma_2 \neq 0 \ (K > 2).$$

Takto pokračujeme v postupném zvyšování K a testování, dokud nedojdeme k nezamítnutí nulové hypotézy. Na základě toho rozhodneme o vhodném počtu bodů zlomu v časové řadě. [4]

Kapitola 3

ARIMA modely

V této kapitole se zaměříme na teorii zvanou Boxova-Jenkinsova metodologie, která zahrnuje modely MA (modely klouzavých součtů), modely AR (autoregresní modely), smíšené modely ARMA a na závěr integrované modely ARIMA. Jedná se o hojně využívané stochastické modely, pomocí kterých lze modelovat jak trendovou složku, tak i sezónní složku. Všechny výše zmíněné modely postupně popíšeme. Nejprve se ale zaměříme na pojmy, bez kterých se tato kapitola neobejde, a to jsou stacionarita a autokorelace.

3.1. Stacionarita

Pro analýzu časových řad pomocí Boxovy-Jenkinsovy metodologie potřebujeme předpokládat, že je časová řada stacionární. To jednoduše řečeno znamená, že je časová řada neměnná v čase a nevyskytuje se v jejím průběhu žádný výrazný trend ani sezónnost.

3.1.1. Typy stacionarity

Rozlišujeme dva typy stacionarity, a to slabou a silnou stacionaritu. Aby byl proces slabě stacionární, musí splňovat následující vlastnosti:

- má konstantní střední hodnotu, tj. $E(Y_t) = \mu$ pro každý časový okamžik t
- má konstantní rozptyl, tj. $\text{var}(Y_t) = \sigma^2$ pro každý časový bod t
- korelace mezi dvěma body časové řady závisí pouze na tom, jak daleko jsou od sebe tyto body vzdálené, nikoliv na tom, kde se v časové řadě nacházejí, tj. $\text{cor}(Y_t, Y_{t+s})$=$\text{cor}(Y_r, Y_{r+s})$ pro jakékoliv časové body t a r a jakoukoliv vzdálenost s.

Silně stacionární proces si klade přísnější podmínky, jelikož navíc předpokládá, že:

- sdružené rozdělení pravděpodobnosti každé množiny veličin $\{Y_{t_1}, \ldots, Y_{t_k}\}$ je stejné jako sdružené rozdělení množiny náhodných veličin $\{Y_{t_1+s}, \ldots, Y_{t_k+s}\}$ posunuté o nějaký časový okamžik s. Jinak řečeno musí platit rovnost následujících pravděpodobností:

$$P(Y_{t_1} \leq c_1, \ldots, Y_{t_k} \leq c_k) = P(Y_{t_1+s} \leq c_1, \ldots, Y_{t_k+s} \leq c_k)$$

  pro jakékoliv časové body $t_1, \ldots, t_k$, jakýkoliv posun s a pro jakékoliv reálná čísla $c_1, \ldots, c_k$.

Při analýze přerušovaných časových řad bude stačit předpoklad slabě stacionárního procesu. Stacionaritou budeme dále v textu rozumět slabou stacionaritu. [10, 21]

3.1.2. Diferencování časové řady

V praxi často narazíme na časové řady, které stacionární nebudou (v naší situaci nejspíše přerušované časové řady s velkým bodem zlomu). I s takovými řadami se dá v Boxově-Jenkinsově metodologii pracovat, jelikož většinu časových řad dokážeme na stacionární procesy převést. To se nejčastěji provádí pomocí

diferencování procesu, což znamená, že se místo původních hodnot časové řady zaměříme na rozdíly po sobě jdoucích pozorování:

$$Y'_t = Y_t - Y_{t-1},\ t = 2, \ldots, T_n.$$

Časová řada diferencí prvního řádu tedy bude mít $(n-1)$ složek, konkrétně $Y'_2, Y'_3, \ldots, Y'_{T_n}$. [10]

V některých případech, kdy první diferencování ke stacionaritě nepomůže, lze řadu převést na řadu druhých diferencí:

$$\begin{aligned} Y''_t &= Y'_t - Y'_{t-1} \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2},\ t = 3, \ldots, T_n, \end{aligned} \tag{3.1}$$

která bude obsahovat $(n-2)$ složek, a to $Y''_3, Y''_4, \ldots, Y''_{T_n}$.

3.2. Autokorelace

Pozorování časové řady můžou být často korelovány s pozorováními v předchozích časových bodech. Tuto závislost nelze měřit pomocí klasické korelace, jelikož ta měří lineární vztah mezi dvěma proměnnými. Budeme ji měřit pomocí autokorelace, která vyjadřuje lineární vztah mezi zpožděnými hodnotami stejné časové řady. Jinými slovy se jedná o závislost mezi náhodnými složkami jedné proměnné v čase. V této kapitole dále popíšeme pojem autokorelační funkce a parciální autokorelační funkce.

3.2.1. Autokorelační funkce

Při tvorbě autokorelační funkce (neboli ACF z anglického AutoCorrelation Function) musíme předpokládat, že je časová řada stacionární. Autokorelační

funkci pak definujeme pomocí korelačního koeficientu mezi dvěma hodnotami téže časové řady:

$$\rho_k = \mathrm{cor}(Y_t, Y_{t-k}) = \frac{\mathrm{cov}(Y_t, Y_{t-k})}{\mathrm{var}(Y_t)}, \quad k = \ldots, -1, 0, 1, \ldots$$

Kovarianci a rozptyl ve vzorci odhadujeme následovně:

$$\widehat{\mathrm{cov}}(Y_t, Y_{t-k}) = \frac{1}{n} \sum_{t=k+1}^{n} (Y_t - \overline{Y})(Y_{t-k} - \overline{Y}),$$

$$\widehat{\mathrm{var}}(Y_t) = \frac{1}{n} \sum_{t=1}^{n} (Y_t - \overline{Y})^2,$$

kde $\overline{Y}$ odpovídá výběrovému průměru:

$$\overline{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t.$$

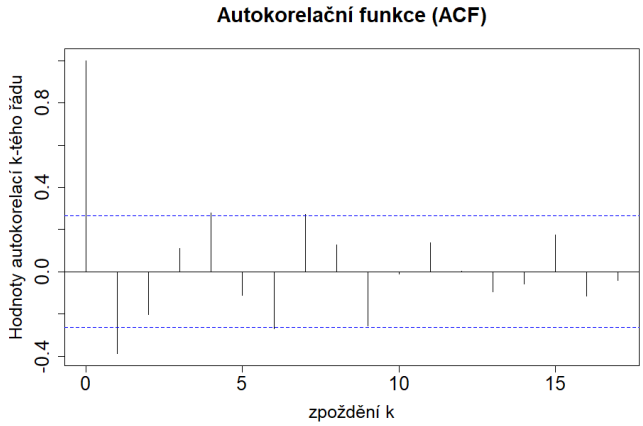
Index k ve vzorcích vyjadřuje, o kolik zpožděných časových bodů se máme posunout a vypočítat pak korelaci (nebo kovarianci) a n značí délku časové řady. Například ρ_1=cor(Y_t, Y_{t-1}) vyjadřuje korelaci mezi dvěma po sobě jdoucími hodnotami (bez ohledu na to, kde se v časové řadě nacházejí), kterou nazýváme jako autokorelace 1. řádu. ρ_2 už vyjadřuje autokorelaci 2. řádu, kdy nás zajímá, zda spolu souvisí také hodnoty zpožděné o 2 časové období, tedy cor(Y_t, Y_{t-2}). Takto bychom mohli vypočítat autokorelaci jakéhokoliv řádu, což můžeme graficky znázornit pomocí tzv. korelogramu, viz obrázek 3.1. Na osu x nanášíme zpoždění k (anglicky lag) a pro každé k odhadneme autokorelaci ρ_k, která nabývá hodnot mezi -1 a 1. Pro $k = 0$ bude vždy platit, že $\rho_0 = 1$, jelikož se jedná o korelaci dvou stejných časových bodů, tj. $\rho_0 = \mathrm{cor}(Y_t, Y_t) = 1$. Protože se jedná o sudou funkci, stačí korelogram vykreslovat pouze v pravé části grafu, tedy pro $k \geq 0$. [3, 15]

Na obrázku 3.1 lze vidět také modrou hranici pro statisticky významné auto-

korelace, která se vypočítá jako $\pm 2/\sqrt{n}$, kde n vyjadřuje délku časové řady. Každý odhadnutý autokorelační koeficient $\widehat{\rho}_k$ má asymptotické normální rozdělení:

$$\widehat{\rho}_k \sim N\left(\rho_k, \frac{4}{n}\right).$$

Pravděpodobnost toho, že ρ_k náleží do pásu ohraničeného hranicemi $\pm 2/\sqrt{n}$, je pro každé $k = 1, \ldots, n$, přibližně 95 %. [10]



Obrázek 3.1: Ukázka korelogramu pro autokorelační funkci

3.2.2. Parciální autokorelační funkce

Parciální autokorelační funkce (PACF z anglického Partial AutoCorrelation Function) opět měří korelaci mezi Y_t a Y_{t-k}, ale nyní bez vlivu hodnot mezi nimi. Závislost dvou hodnot tedy „očistíme“ od těch, které by nám tento vztah mohly ovlivňovat. První parciální autokorelace je stejná jako hodnota klasické autokorelace 1. řádu ρ_1, jelikož se mezi dvěma po sobě jdoucími hodnotami žádná jiná hodnota, jejíž vliv bychom mohli potlačit, nevyskytuje. Parciální autokorelace 2. řádu mezi hodnotami Y_t a Y_{t-2} již chceme spočítat bez vlivu Y_{t-1} atd. K tomuto výpočtu nyní nelze použít jednoduchý korelační koeficient, ale je nutné zadefinovat parciální korelační koeficient. [9]

Mějme dvě náhodné veličiny Y a Z a náhodný vektor $\mathbf{X} = (X_1, \ldots, X_k)'$. Budeme chtít vypočítat parciální korelační koeficient mezi veličinami Y a Z, který chceme očistit od vlivu náhodných veličin $X_1, \ldots, X_k$ (neboli zafixovat jejich hodnoty). Nejdříve je potřeba odhadnout lineární závislost Y na $X_1, \ldots, X_k$ a také Z na $X_1, \ldots, X_k$ pomocí metody nejmenších čtverců, čímž získáme $\widehat{Y}$ a $\widehat{Z}$. Parciální korelační koeficient pak vypočítáme jako korelační koeficient mezi $Y - \widehat{Y}$ a $Z - \widehat{Z}$, tedy mezi těmi částmi, které nejsou vysvětlené pomocí veličin $X_1, \ldots, X_k$:

$$\rho_{Y,Z.X} = \mathrm{pcor}(Y, Z) = \mathrm{cor}(Y - \widehat{Y}, Z - \widehat{Z}).$$

Takto zadefinovaný parciální korelační koeficient pak můžeme využít při tvorbě autokorelační parciální funkce, kdy pro každé k vypočítáme tentokrát hodnotu parciálního korelačního koeficientu mezi hodnotami Y_t a Y_{t-k}:

$$\rho_{kk} = \mathrm{pcor}(Y_t, Y_{t-k}) = \mathrm{cor}(Y_t - \widehat{Y}_t, Y_{t-k} - \widehat{Y}_{t-k}),$$

kde

$$\widehat{Y}_t = \widehat{\beta}_1 Y_{t-1} + \cdots + \widehat{\beta}_{k-1} Y_{t-k+1},$$

$$\widehat{Y}_{t-k} = \widehat{\beta}_1 Y_{t-k+1} + \cdots + \widehat{\beta}_{k-1} Y_{t-1},$$

čímž odstraníme efekt hodnot mezi Y_t a Y_{t-k}, tedy $Y_{t-1}, \ldots, Y_{t-k+1}$. To, že jsou koeficienty $\beta_1, \ldots, \beta_{k-1}$ stejné, vyplývá ze stacionarity a následujícího výpočtu. Zvolíme např. $k = 3$ a vypočítáme parciální korelační koeficient mezi Y_t a Y_{t-3}:

$$\rho_{33} = \mathrm{pcor}(Y_t, Y_{t-3}) = \mathrm{cor}(Y_t - \widehat{Y}_t, Y_{t-3} - \widehat{Y}_{t-3}),$$

kde

$$\widehat{Y}_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2},$$

$$\widehat{Y}_{t-3} = \beta_1 Y_{t-2} + \beta_2 Y_{t-1}$$

je nejlepší lineární aproximace Y_t a Y_{t-3} pomocí Y_{t-1} a Y_{t-2}. Neznámé koeficienty β_1 a β_2 odhadujeme na základě minimalizace střední čtvercové chyby. Nejprve vypočítejme střední čtvercovou chybu pro Y_t

$$E(Y_t - \widehat{Y}_t)^2 = E(Y_t - \beta_1 Y_{t-1} - \beta_2 Y_{t-2})^2 = \mathrm{var}(Y_t) + \beta_1^2 \mathrm{var}(Y_{t-1}) + \beta_2^2 \mathrm{var}(Y_{t-2}) \\ - 2\beta_1 \mathrm{cov}(Y_t, Y_{t-1}) + 2\beta_1\beta_2 \mathrm{cov}(Y_{t-1}, Y_{t-2}) - 2\beta_2 \mathrm{cov}(Y_t, Y_{t-2})$$

a následně vypočítáme parciální derivace tohoto výsledku podle β_1 jako

$$\frac{\partial E(Y_t - \widehat{Y}_t)^2}{\partial \beta_1} = 2\beta_1 \mathrm{var}(Y_{t-1}) - 2\mathrm{cov}(Y_t, Y_{t-1}) + 2\beta_2 \mathrm{cov}(Y_{t-1}, Y_{t-2}).$$

Výraz položíme roven 0 a vyjádříme koeficient β_1:

$$\beta_1 = \frac{\mathrm{cov}(Y_t, Y_{t-1}) - \beta_2 \mathrm{cov}(Y_{t-1}, Y_{t-2})}{\mathrm{var}(Y_{t-1})} = \rho_1 - \beta_2 \rho_1,$$

kde ρ_1 odpovídá výše zavedené korelaci $\mathrm{cor}(Y_t, Y_{t-1})$. Pro parciální derivaci podle β_2 platí

$$\frac{\partial E(Y_t - \widehat{Y}_t)^2}{\partial \beta_2} = 2\beta_2 \mathrm{var}(Y_{t-2}) + 2\beta_1 \mathrm{cov}(Y_{t-1}, Y_{t-2}) - 2\mathrm{cov}(Y_t, Y_{t-2}).$$

Odtud

$$\beta_2 = \frac{\mathrm{cov}(Y_t, Y_{t-2}) - \beta_1 \mathrm{cov}(Y_{t-1}, Y_{t-2})}{\mathrm{var}(Y_{t-2})} = \rho_2 - \beta_1 \rho_1.$$

Stejné výpočty provedeme pro Y_{t-3}, kde střední čtvercovou chybu vyjádříme jako

$$E(Y_{t-3} - \widehat{Y}_{t-3})^2 = E(Y_{t-3} - \beta_1 Y_{t-2} - \beta_2 Y_{t-1})^2 = \mathrm{var}(Y_{t-3}) + \beta_1^2 \mathrm{var}(Y_{t-2}) + \\ \beta_2^2 \mathrm{var}(Y_{t-1}) - 2\beta_1 \mathrm{cov}(Y_{t-3}, Y_{t-2}) + 2\beta_1\beta_2 \mathrm{cov}(Y_{t-2}, Y_{t-1}) - 2\beta_2 \mathrm{cov}(Y_{t-3}, Y_{t-1}).$$

Opět vypočítáme parciální derivace, které položíme rovny nule a vyjádříme hle-

dané koeficienty. Pro první parciální derivaci dostáváme

$$\frac{\partial E(Y_{t-3} - \widehat{Y}_{t-3})^2}{\partial \beta_1} = 2\beta_1 \mathrm{var}(Y_{t-2}) - 2\mathrm{cov}(Y_{t-3}, Y_{t-2}) + 2\beta_2 \mathrm{cov}(Y_{t-2}, Y_{t-1}).$$

Odtud

$$\beta_1 = \frac{\mathrm{cov}(Y_{t-2}, Y_{t-3}) - \beta_2 \mathrm{cov}(Y_{t-1}, Y_{t-2})}{\mathrm{var}(Y_{t-2})} = \rho_1 - \beta_2 \rho_1.$$

Pro druhou parciální derivaci platí

$$\frac{\partial E(Y_{t-3} - \widehat{Y}_{t-3})^2}{\partial \beta_2} = 2\beta_2 \mathrm{var}(Y_{t-1}) + 2\beta_1 \mathrm{cov}(Y_{t-2}, Y_{t-1}) - 2\mathrm{cov}(Y_{t-3}, Y_{t-1}).$$

Odtud

$$\beta_2 = \frac{\mathrm{cov}(Y_{t-1}, Y_{t-3}) - \beta_1 \mathrm{cov}(Y_{t-1}, Y_{t-2})}{\mathrm{var}(Y_{t-1})} = \rho_2 - \beta_1 \rho_1.$$

Můžeme vidět, že pro Y_t a Y_{t-3} jsme na základě minimalizace střední čtvercové chyby dostali stejné odhady koeficientů β_1, β_2, a to ve tvaru

$$\begin{aligned} \widehat{\beta}_1 &= \rho_1 - \widehat{\beta}_2 \rho_1 = (1 - \widehat{\beta}_2)\rho_1 \\ \widehat{\beta}_2 &= \rho_2 - \widehat{\beta}_1 \rho_1. \end{aligned}$$

Po vyřešení této soustavy dosazením $\widehat{\beta}_1$ do $\widehat{\beta}_2$ a po nahrazení teoretických autokorelací jejich odhady, můžeme psát

$$\begin{aligned} \widehat{\beta}_1 &= \widehat{\rho}_1 \Big(1 - \frac{\widehat{\rho}_2 - \widehat{\rho}_1^2}{1 - \rho_1^2}\Big) \\ \widehat{\beta}_2 &= \frac{\widehat{\rho}_2 - \widehat{\rho}_1^2}{1 - \widehat{\rho}_1^2}. \end{aligned}$$

Toto tvrzení lze zobecnit pro libovolné zpoždění k. Odhadnuté hodnoty parciálního autokorelačního koeficientu $\widehat{\rho}_{kk}$ lze nakonec vynést do korelogramu. [10, 21]

3.3. Testy stacionarity a autokorelace

Pro ověření podmínek stacionarity využíváme především dva testy, a to KPSS test a rozšířený Dickeyův-Fullerův test, které nyní představíme.

KPSS test

První možností, jak otestovat, zda je řada stacionární či zda bude potřeba diferencování, je KPSS test (pojmenováno podle autorů Kwiatkowski, Phillips, Schmidt, Shin). Tento test pracuje s následující nulovou a alternativní hypotézou:

$$H_0 : \text{časová řada je stacionární}$$
$$H_A : \text{časová řada není stacionární.}$$

Uvažujme nyní regresní model, který rozložíme na náhodný absolutní člen r_t, deterministický trend βt a náhodnou chybu ϵ_t:

$$Y_t = r_t + \beta t + \epsilon_t,$$

kde r_t je náhodná procházka, pro kterou platí:

$$r_t = r_{t-1} + u_t.$$

Pro chybu u_t předpokládáme, že jsou její hodnoty nezávislé, stejně rozdělené s nulovou střední hodnotou a rozptylem σ_u^2. S využitím tohoto rozptylu můžeme nulovou a alternativní hypotézu přepsat jako

$$H_0 : \sigma_u^2 = 0$$
$$H_A : \sigma_u^2 > 0.$$

Z toho vyplývá, že za platnosti nulové hypotézy bude absolutní člen roven číslu r_0 a bude platit:

$$Y_t = r_0 + \beta t + \epsilon_t.$$

Testová statistika, kterou při KPSS testu využíváme k ohodnocení hypotézy, je ve tvaru:

$$\text{KPSS} = \frac{1}{n^2} \sum_{t=1}^{n} S_t^2 / \widehat{\sigma}_\epsilon^2,$$

kde $S_t^2 = \sum_{i=1}^{t} \widehat{e}_i^2,\ t = 1, \ldots, n$, definuje částečný součet reziduí a $\widehat{\sigma}_\epsilon^2$ odhad rozptylu chyb ϵ_t, který získáme pomocí vztahu

$$\widehat{\sigma}_\epsilon^2 = \lim_{n \to +\infty} \frac{1}{n} E\big(S_t^2\big).$$

Výslednou hodnotu testové statistiky poté porovnáme s kritickou hodnotou KPSS testu a rozhodneme o zamítnutí či nezamítnutí nulové hypotézy. [7, 8, 13]

Rozšířený Dickeyův-Fullerův test

Druhou možností, jak otestovat stacionaritu časové řady, je rozšířený Dickeyův-Fullerův test. Tento test již nebudeme detailně popisovat (lze odkázat např. na literaturu [19]), ovšem je potřeba zmínit, že nulová a alternativní hypotéza je nyní přesně naopak než u KPSS testu. Testujeme zde tedy následující hypotézu:

$$H_0 : \text{časová řada není stacionární}$$
$$H_A : \text{časová řada je stacionární.}$$

Časová řada proto bude stacionární, pokud nulovou hypotézu tohoto testu zamítneme. K otestování lze využít v softwaru R funkci `adf.test` (z anglického Augmented Dickey-Fuller test).

K otestování přítomnosti autokorelace v časových řadách můžeme využít například následující Ljung-Boxův test (pojmenovaný opět po autorech).

Ljung - Boxův test

Pro otestování autokorelace zavedeme následující nulovou a alternativní hypotézu Ljung-Boxova testu:

$$H_0 : \rho_1 = \cdots = \rho_m = 0 \text{ (v datech se nevyskytuje autokorelace)}$$
$$H_A : \exists i : \rho_i \neq 0 \text{ (v datech se vyskytuje autokorelace).}$$

O zamítnutí či nezamítnutí nulové hypotézy rozhodujeme na základě následující testové Q-statistiky:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\widehat{\rho}_k^2}{n-k},$$

kde n značí délku časové řady a ρ_k autokorelaci k-tého řádu. Písmeno m vyjadřuje, do jakého řádu chceme autokorelaci testovat, tedy dosazujeme $\rho_1, \ldots, \rho_m$. Za platnosti nulové hypotézy se testová statistika asymptoticky řídí χ^2 rozdělením s m stupni volnosti. Nulovou hypotézu proto zamítneme na hladině významnosti α tehdy, jestliže testová statistika Q překročí hranici $(1-\alpha)$ kvantilu rozdělení χ^2_m. Po zamítnutí nulové hypotézy můžeme říct, že se v datech nějaká autokorelace vyskytuje. [2, 15]

3.4. Procesy Boxovy-Jenkinsovy metodologie

V této kapitole se již přesuneme k popisu jednotlivých procesů AR, MA, ARMA a ARIMA, které jsou součástí Boxovy-Jenkinsovy metodologie. Zaměříme se také na odvození jednotlivých charakteristik těchto procesů, jako je střední hodnota, rozptyl a autokorelace. V této kapitole je čerpáno ze zdrojů [3, 10, 15, 21].

3.4.1. Procesy MA(q)

Proces klouzavých součtů řádu q, neboli zkráceně MA(q) z anglického Moving Average, představuje posloupnost náhodných veličin danou předpisem

$$Y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}.$$

Reálné hodnoty $\theta_1, \ldots, \theta_q$ symbolizují parametry procesu a posloupnost $\{\epsilon_t\}$ zde

představuje proces bílého šumu, pro který platí

$$\epsilon_t \sim N(0, \sigma^2).$$

Nejjednodušší proces klouzavých součtů je MA(1), neboli proces prvního řádu, kde na výslednou hodnotu Y_t má vliv pouze ϵ_t a ϵ_{t-1}:

$$Y_t = \epsilon_t + \theta_1 \epsilon_{t-1}.$$

Procesy MA(q) řadíme mezi slabě stacionární procesy, jelikož mají nulovou střední hodnotu, konstantní rozptyl v čase a korelace hodnot závisí pouze na jejich vzdálenosti. Nyní tyto charakteristiky odvodíme, přičemž budeme vycházet zejména z vlastností bílého šumu, tedy že platí: $\epsilon_t \sim N(0, \sigma^2)$ a $\text{cov}(\epsilon_i, \epsilon_j) = 0, \forall i, j, i \neq j$.

Střední hodnota

$$\begin{aligned} E(Y_t) &= E\big(\epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}\big) \\ &= E(\epsilon_t) + \theta_1 E(\epsilon_{t-1}) + \theta_2 E(\epsilon_{t-2}) + \cdots + \theta_q E(\epsilon_{t-q}) \\ &= 0 + \theta_1 \times 0 + \theta_2 \times 0 + \cdots + \theta_q \times 0 \\ &= 0 \end{aligned}$$

Rozptyl

$$\begin{aligned} \text{var}(Y_t) &= \text{var}\big(\epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}\big) \\ &= \text{var}(\epsilon_t) + \text{var}(\theta_1\epsilon_{t-1}) + \cdots + \text{var}(\theta_q\epsilon_{t-q}) + 2\text{cov}(\epsilon_t, \theta_1\epsilon_{t-1}) + \cdots + \\ &\quad 2\text{cov}(\epsilon_t, \theta_q\epsilon_{t-q}) \\ &= \sigma^2 + \theta_1^2\sigma^2 + \cdots + \theta_q^2\sigma^2 + 0 + \cdots + 0 \\ &= \sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2) \end{aligned}$$

Autokorelace

$$\begin{aligned}\operatorname{cor}(Y_t, Y_{t-k}) &= \frac{\operatorname{cov}(Y_t, Y_{t-k})}{\sqrt{\operatorname{var}(Y_t)}\sqrt{\operatorname{var}(Y_{t-k})}} \\ &= \frac{\operatorname{cov}(\epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}, \epsilon_{t-k} + \theta_{1+k}\epsilon_{t-k-1} + \cdots + \theta_{q+k}\epsilon_{t-k-q})}{\sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2)} \\ &= \frac{\sigma^2(\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q)}{\sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2)} \\ &= \frac{\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \cdots + \theta_q^2}\end{aligned}$$

3.4.2. Procesy AR(p)

Autoregresní procesy řádu p jsou další typ modelu, který řadíme do Boxovy-Jenkinsovy metodologie. Jedná se o náhodnou posloupnost, která je nyní dána předpisem

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t,$$

kde ϵ_t je bílý šum a $\phi_1, \ldots, \phi_p$ představují reálné parametry procesu. Nyní hodnotu Y_t předpovídáme pomocí lineární kombinace zpožděných hodnot stejné proměnné (do jaké minulosti se díváme, záleží na indexu p). Autoregresní proces prvního řádu bude ve tvaru

$$Y_t = \phi Y_{t-1} + \epsilon_t.$$

Tento typ procesů nelze automaticky považovat za stacionární, jako tomu bylo u procesů MA(q). Aby byly stacionární, musí splňovat podmínku, že všechna řešení rovnice

$$1 - \sum_{j=1}^{p} \phi_j x^j = 0$$

leží vně jednotkového kruhu. Pro procesy AR(1) lze tuto podmínku na staciona-

ritu přepsat jako

$$|\phi| < 1,$$

která vyplývá z následujícího odvození. Rozepíšeme vztah pro proces AR(1):

$$\begin{aligned} Y_t &= \phi Y_{t-1} + \epsilon_t \\ &= \phi(\phi Y_{t-2} + \epsilon_{t-1}) + \epsilon_t = \phi^2 Y_{t-2} + \phi\epsilon_{t-1} + \epsilon_t \\ &= \phi(\phi(\phi Y_{t-3} + \epsilon_{t-2}) + \epsilon_{t-1}) + \epsilon_t = \phi^3 Y_{t-3} + \phi^2\epsilon_{t-2} + \phi\epsilon_{t-1} + \epsilon_t, \end{aligned}$$

což lze souhrnně zapsat jako

$$Y_t = \phi^k Y_{t-k} + \sum_{i=0}^{k-1} \phi^i \epsilon_{t-i}. \tag{3.2}$$

Z toho plyne, že pokud bychom za parametr ϕ ve vztahu (3.2) dosadili číslo takové, že $|\phi| < 1$, byl by člen $\phi^k Y_{t-k}$ zanedbatelný. Druhá část rovnice (3.2) však představuje proces klouzavých součtů, o kterých jsme již dokázali, že jsou vždy stacionární. Proto můžeme říct, že za podmínky $|\phi| < 1$ je stacionární také proces AR(1).

Odvoďme nyní střední hodnotu, rozptyl a korelaci pro stacionární procesy AR(1), kde využijeme právě této podmínky a také toho, že stacionární procesy mají konstantní střední hodnotu i rozptyl (tedy např. platí $E(Y_t) = E(Y_{t-1})$).

Střední hodnota

$$\begin{aligned} E(Y_t) &= E(\phi Y_{t-1} + \epsilon_t) \\ E(Y_t) &= \phi E(Y_{t-1}) + E(\epsilon_t) \\ E(Y_t) &= \phi E(Y_t) + 0 \\ E(Y_t)(1-\phi) &= 0 \\ E(Y_t) &= 0 \end{aligned}$$

Rozptyl

$$\begin{aligned}
\text{var}(Y_t) &= \text{var}(\phi Y_{t-1} + \epsilon_t) \\
&= \phi^2 \text{var}(Y_{t-1}) + \text{var}(\epsilon_t) + \phi 2\text{cov}(Y_{t-1}, \epsilon_t) \\
&= \phi^2 \text{var}(Y_t) + \sigma^2 + 0 \\
\text{var}(Y_t)(1 - \phi^2) &= \sigma^2 \\
\text{var}(Y_t) &= \frac{\sigma^2}{1 - \phi^2}
\end{aligned}$$

Před odvozením autokorelace se nejdříve zaměříme na výpočet kovariance mezi Y_t a Y_{t-k}, kde v prvním kroku dosadíme za Y_t odvozený vztah (3.2):

Kovariance

$$\begin{aligned}
\text{cov}(Y_t, Y_{t-k}) &= \text{cov}(\phi^k Y_{t-k} + \sum_{i=0}^{k-1} \phi^i \epsilon_{t-i}, Y_{t-k}) \\
&= \phi^k \text{cov}(Y_{t-k}, Y_{t-k}) + \sum_{i=0}^{k-1} \phi^i \text{cov}(\epsilon_{t-i}, Y_{t-k}) \\
&= \phi^k \text{var}(Y_{t-k}) + 0 \\
&= \phi^k \sigma^2
\end{aligned}$$

Autokorelace

$$\begin{aligned}
\text{cor}(Y_t, Y_{t-k}) &= \frac{\text{cov}(Y_t, Y_{t-k})}{\sqrt{\text{var}(Y_t)}\sqrt{\text{var}(Y_{t-k})}} \\
&= \frac{\phi^k \sigma^2}{\frac{\sigma^2}{1-\phi^2}} \\
&= \phi^k (1 - \phi^2)
\end{aligned}$$

3.4.3. Procesy ARMA(p,q)

Smíšeným procesem řádu p a q, který označujeme ARMA(p,q), rozumíme posloupnost náhodných veličin $\{Y_t\}$, která je dána následujícím vztahem:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}.$$

Jedná se o kombinovaný model autoregrese, které přísluší stupeň p, a klouzavých součtů se stupněm q. Odhady koeficientů $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ jsou v programu R získávány pomocí metody maximální věrohodnosti.

3.4.4. Procesy ARIMA(p,d,q)

Nejobecnějším modelem Boxovy-Jenkinsovy metodologie je nestacionární integrovaný smíšený model ARIMA(p,d,q), kde písmeno „I“ značí integraci stupně d. Tento model kombinuje diferencování procesu s modelem klouzavých součtů stupně q a modelem autoregresního procesu stupně p. Nyní budeme vyjadřovat posloupnost diferencí $\{Y_t^d\}$, která je dána předpisem

$$Y_t^d = \phi_1 Y_{t-1}^d + \phi_2 Y_{t-2}^d + \cdots + \phi_p Y_{t-p}^d + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}.$$

Výše popsané procesy AR(p) a MA(q) jsou pouze speciálními případy ARIMA procesů. Model AR(p) bychom mohli zapsat jako ARIMA(p,0,0) a model MA(q) jako ARIMA(0,0,q). Proces ARIMA(0,0,0) pak odpovídá bílému šumu.

3.4.5. Konstrukce predikcí pro ARIMA modely

V modelech Boxovy-Jenkinsovy metodologie často potřebujeme znát také budoucí vývoj časové řady. K tomu je třeba zavést konstrukci budoucí hodnoty v čase T_n o k kroků dopředu, kterou budeme značit jako $\widehat{Y}_{T_n+k}(T_n)$. Konstrukci predikcí si ukážeme na procesu ARMA(p,q), který jsme zavedli v kapitole 3.4.3.

Předpověď $\widehat{Y}_{T_n+k}(T_n)$ konstruujeme jako lineární funkci hodnot Y_{T_n},$Y_{T_n-1}, \ldots$, respektive ϵ_{T_n},$\epsilon_{T_n-1}, \ldots$ (z předpokladu stacionarity a invertibility procesu):

$$\widehat{Y}_{T_n+k}(T_n) = \psi_k \epsilon_{T_n} + \psi_{k+1}\epsilon_{T_n-1} + \ldots,$$

přičemž koeficienty $\psi_k, \psi_{k+1}, \ldots$, určíme tak, aby minimalizovaly střední čtvercovou chybu předpovědi. Dále můžeme psát, že platí

$$\widehat{Y}_{T_n+k}(T_n) = [Y_{T_n+k}],$$

kde hranaté závorky značí podmíněnou střední hodnotu Y_{T_n+k} při pevných hodnotách Y_{T_n},$Y_{T_n-1}, \ldots$ Na základě posloupnosti procesu ARMA(p,q) pro hodnotu Y_{T_n+k} platí

$$Y_{T_n+k} = \phi_1 Y_{T_n+k-1} + \cdots + \phi_p Y_{T_n+k-p} + \epsilon_{T_n+k} + \theta_1 \epsilon_{T_n+k-1} + \cdots + \theta_q \epsilon_{T_n+k-q},$$

přičemž pro predikci tedy můžeme psát

$$\widehat{Y}_{T_n+k}(T_n) = \phi_1[Y_{T_n+k-1}] + \cdots + \phi_p[Y_{T_n+k-p}] + [\epsilon_{T_n+k}] + \theta_1[\epsilon_{T_n+k-1}] + \cdots + \theta_q[\epsilon_{T_n+k-q}]. \tag{3.3}$$

Nakonec je potřeba rozepsat, že pro podmíněné střední hodnoty platí

$$[Y_{T_n+h}] = \begin{cases} \widehat{Y}_{T_n+h}(T_n) & \text{pro } h > 0 \\ Y_{T_n+h} & \text{pro } h \leq 0, \end{cases} \tag{3.4}$$

$$[\epsilon_{T_n+h}] = \begin{cases} 0 & \text{pro } h > 0 \\ \epsilon_{T_n+h} & \text{pro } h \leq 0. \end{cases} \tag{3.5}$$

Při výpočtu predikovaných hodnot postupujeme rekurentně, tedy postupně napočítáme hodnoty $\widehat{Y}_{T_n+1}(T_n), \widehat{Y}_{T_n+2}(T_n+1), \ldots$, a následně do vzorce (3.3) dosadíme

vztahy (3.4) a (3.5).

Na závěr zkonstruujme předpovědní interval pro budoucí pozorování v čase $T_n + k$:

$$\left(\widehat{Y}_{T_n+k}(T_n) - 2\sqrt{\mathrm{var}(e_{T_n+k}(T_n))}, \widehat{Y}_{T_n+k}(T_n) + 2\sqrt{\mathrm{var}(e_{T_n+k}(T_n))}\right),$$

kde $e_{T_n+k}(T_n)$ označuje chybu předpovědi

$$e_{T_n+k}(T_n) = Y_{T_n+k} - \widehat{Y}_{T_n+k}(T_n) = \epsilon_{T_n+k} + \psi_1\epsilon_{T_n+k-1} + \cdots + \psi_{k-1}\epsilon_{T_n+1}$$

a pro její rozptyl platí

$$\mathrm{var}(e_{T_n+k}(T_n)) = (1 + \psi_1^2 + \cdots + \psi_{k-1}^2)\sigma_\epsilon^2.$$

Více o předpovědních intervalech viz [3] nebo [10].

Kapitola 4

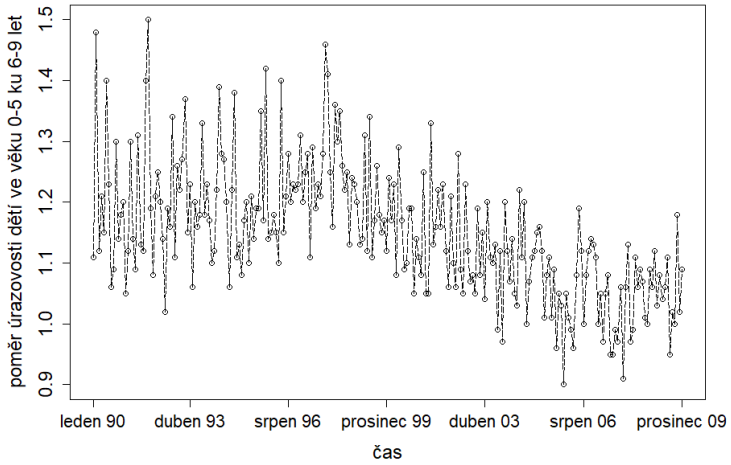
Praktická část

Poslední část této práce bude věnována aplikaci již popsané teorie přerušovaných časových řad. Teorii ukážeme na reálné časové řadě, jejíž analýzu budeme postupně detailně rozvíjet a popisovat. Přistoupíme k analýze této řady jak s předpokladem znalosti času intervence, tak nejprve bez této znalosti. Celý proces analýzy bude proveden pomocí softwaru R [20].

4.1. Data

Budeme analyzovat data o úrazech dětí v motorových vozidlech v Japonsku. Data pochází z policejních údajů, která byla získávána každý měsíc v období od ledna 1990 do prosince 2009. Naším zájmem jsou měsíční poměry úrazovosti dětí ve věku 0–5 let ve srovnání s dětmi ve věku 6–9 let. Jedná se o časovou řadu o délce 240 měsíců, v průběhu které došlo ke změně legislativy, týkající se zavedení povinných dětských zádržných systémů v automobilech u dětí ve věku 0–5 let (více o povinných typech dětských autosedaček v Japonsku např. zde [26]). Cílem této analýzy je proto zjištění, jak tato legislativní změna, kterou považujeme za intervenci, ovlivnila úrazy malých dětí ve věku 0–5 let a zda došlo k prevenci dopravních úrazů. Zkoumaná data vykreslená do grafu vidíme na obrázku 4.1

(dále budeme časovou osu x značit pouze body od 1 do 240). [18, 22]



Obrázek 4.1: Časová řada poměrů úrazovosti dětí ve věku 0–5 let ve srovnání s dětmi ve věku 6–9 let v Japonsku

4.2. Analýza časové řady bez znalosti intervence

V této části se budeme věnovat přerušované časové řadě, při níž budeme nejdříve předpokládat, že čas intervence není znám. Zaměříme se proto na detekci bodu zlomu pomocí tří různých metod a v závěru této části porovnáme výsledné odhady se skutečným časem intervence.

4.2.1. Lineární regrese

Začneme tím, jak bychom data modelovali v případě, kdy bychom si mysleli, že v průběhu časové řady nenastala žádná intervence. Následně tento model porovnáme s variantou zohledňující možnou intervenci, a posoudíme, zda je zahrnutí intervence vůbec nutné. Zaměříme se pouze na model regresní přímky, jelikož na přímkové regresi jsou založeny také dále používané metody.

Po vytvoření modelu regresní přímky s názvem m, získáme v softwaru R výstup, který vidíme na obrázku 4.2. Poměry úrazovosti dětí ve věku 0–5 let ku dětem ve věku 6–9 let zde považujeme za závisle proměnnou Y_t a v kódech je budeme značit jako *outcome*. Nezávisle proměnnou je zde čas (neboli měsíce označené od 1 do 240), který budeme v modelech nazývat jako *time*.

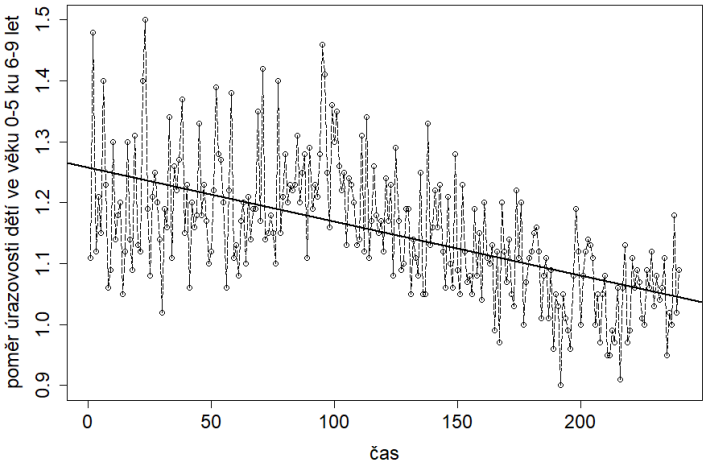
```
> m=lm(outcome~time,data=d)
> summary(m)

Call:
lm(formula = outcome ~ time, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.210760 -0.061023 -0.004779  0.052965  0.286766

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.257e+00  1.172e-02   107.3   <2e-16 ***
time        -8.850e-04  8.431e-05   -10.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09049 on 238 degrees of freedom
Multiple R-squared:  0.3164,    Adjusted R-squared:  0.3136
F-statistic: 110.2 on 1 and 238 DF,  p-value: < 2.2e-16
```

Obrázek 4.2: Lineární model m pro poměry úrazovostí v závislosti na čase

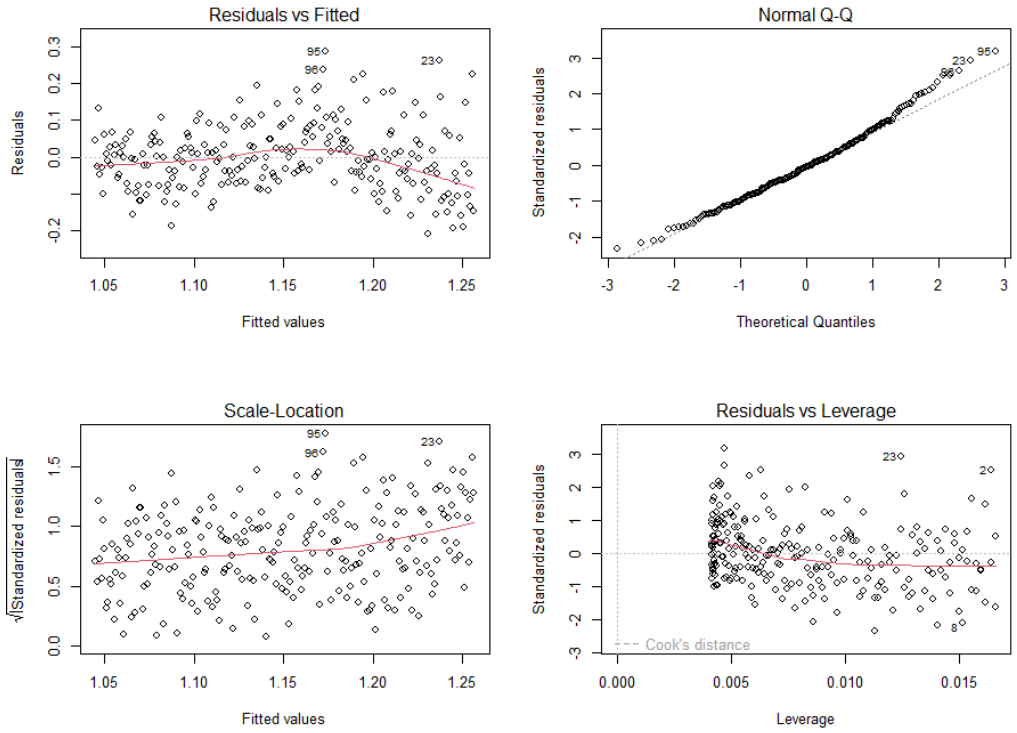
Na základě odhadnutých koeficientů dostáváme přímku s předpisem:

$$\widehat{Y}_t = 1.26 - 0.0009 \times time.$$

Odhadnutou přímku vykreslíme do grafu na obrázku 4.3, kde lze od prvního pohledu vidět, že tato přímka nepředstavuje optimální proložení dat, což platí zejména pro první část časové řady a její konec. To si můžeme potvrdit také pomocí analýzy reziduí tohoto modelu, kterou představují grafy na obrázku 4.4. Vykreslením vyrovnaných hodnot oproti reziduím pozorujeme, že se rezidua nepohybují rovnoměrně kolem nuly (viz červená křivka na prvním grafu). Zároveň lze identifikovat odlehlá pozorování, indikující výrazně větší rezidua. Proložení těchto dat přímkou by tedy mělo jít udělat lépe. V dalších kapitolách se proto zaměříme na to, zda se v této řadě nevyskytuje bod zlomu.



Obrázek 4.3: Časová řada s odhadnutou regresní přímkou na základě lineárního modelu m



Obrázek 4.4: Analýza reziduí lineárního modelu m

4.2.2. Test existence bodu zlomu

Prvním krokem bude ověření existence významného bodu zlomu v této časové řadě. Ačkoliv víme, že k intervenci došlo, není zaručeno, že tato událost měla významný efekt na průběh řady (z grafu však na první pohled vidíme, že nějaká změna pravděpodobně nastala). Otestujme proto existenci bodu zlomu pomocí Daviesova testu, který najdeme v softwaru R jako `davies.test` v knihovně `segmented`. Tento test ověřuje nulovost koeficientu γ, který vyjadřuje změnu sklonu přímky druhého segmentu ve srovnání s prvním. Pokud test hypotézu zamítne, změna sklonu je signifikantní a bod zlomu se v průběhu časové řady vyskytuje. Výstup Daviesova testu pro naše data je vidět na obrázku 4.5.

```
        Davies' test for a change in the slope

data:  formula = outcome ~ time ,   method = lm
model = gaussian , link = identity
segmented variable = time
'best' at = 81, n.points = 10, p-value = 8.81e-06
alternative hypothesis: two.sided
```

Obrázek 4.5: Výstup Daviesova testu pro otestování existence bodu zlomu

Dostáváme, že p–hodnota testu je výrazně menší než hladina významnosti 0.05, tudíž hypotézu o nulovosti parametru γ zamítáme. Bod zlomu tedy v časové řadě existuje, a proto se v další kapitole budeme věnovat jeho detekci. Z výstupu 4.5 můžeme také vyčíst, že Daviesův test určil hledaný bod zlomu v časovém bodě 81. Je však důležité zdůraznit, že tento test není plně určen k identifikaci časů intervencí, proto tento odhad nebudeme brát v potaz.

4.2.3. Detekce času intervence

I. Metoda segmentové regrese

Pokud nevíme, kdy k intervenci v časové řadě došlo, je klíčovým krokem tento čas identifikovat. To provedeme nejdříve pomocí funkce `segmented` ze stejnojmenné knihovny, jejíž model a algoritmus je detailně popsán v kapitole 2.2.1. Základem

pro příkaz `segmented` je lineární model, který jsme již vytvořili na začátku této kapitoly jako `m=lm(outcome ~ time)`. Tento model nyní využijeme při tvorbě segmentové regrese příkazem `seg.m=segmented(m, seg.Z = ~ time)`. Z výstupu pomocí příkazu `summary(seg.m)`, který vidíme na obrázku 4.6, lze vyčíst hned několik informací.

```
        ***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = m, seg.Z = ~time, npsi = 1)

Estimated Break-Point(s):
           Est. St.Err
psi1.time   95 11.965

Meaningful coefficients of the linear terms:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.1913114  0.0176924  67.335   <2e-16 ***
time         0.0003517  0.0003200   1.099    0.273
U1.time     -0.0018823  0.0003623  -5.196       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08554 on 236 degrees of freedom
Multiple R-Squared: 0.3943,  Adjusted R-squared: 0.3866

Boot restarting based on 6 samples. Last fit:
Convergence attained in 1 iterations (rel. change 1.7373e-07)
```

Obrázek 4.6: Model segmentové regrese na základě lineárního modelu m

Základní jsou pro nás odhady koeficientů modelu segmentové regrese. Připomeňme, že nyní pracujeme s modelem (2.5), který vypadal následovně:

$$\mu + \alpha t + \gamma(t - T_i) \times I(t \geq T_i).$$

Dle výstupu na obrázku 4.6 dostáváme odhad absolutního členu jako $\widehat{\mu} = 1.19$. Koeficient α, který udává sklon přímky prvního segmentu, jsme v našem modelu odhadli přibližně na 0.0004 (ve výstupu jako *time*). Koeficient γ (ve výstupu jako $U1.time$) byl odhadnut na -0.0019, což vyjadřuje změnu sklonu přímky druhého segmentu oproti prvnímu segmentu časové řady. Pro získání sklonu také druhého segmentu musíme tyto koeficienty sečíst, a získáme tak číslo -0.0015. Ve výstupu dále vidíme p-hodnoty pro odhadnuté koeficienty modelu. Absolutní

člen vychází na hladině významnosti 0.05 jako signifikantní, ovšem proměnná *time* má p-hodnotu větší než 0.05, tudíž signifikantní není. Sklon prvního segmentu je tedy nevýznamný, a proto později vytvoříme nový model bez této proměnné. Dále vidíme, že pro $U1.time$ nebyla p-hodnota uvedena, a to z důvodu toho, že pro odhad parametru γ zde používáme Daviesův test namísto standardních asymptotických testů (viz kapitola 2.2.3). Jak jsme již otestovali výše, pomocí Daviesova testu vyšel koeficient γ jako významný. Sklony obou přímek lze také snadno získat pomocí příkazu `slope`, který navíc vypíše standardní chybu a 95% intervaly spolehlivosti pro tyto sklony, viz obrázek 4.7. Také odsud lze usoudit, že sklon prvního segmentu není významný, jelikož jeho 95% interval spolehlivosti zahrnuje nulu.

```
> slope(seg.m)
$time
              Est.     St.Err. t value   CI(95%).l   CI(95%).u
slope1  0.00035165 0.00032004  1.0988 -0.00027885  0.00098216
slope2 -0.00153070 0.00016972 -9.0190 -0.00186500 -0.00119630
```

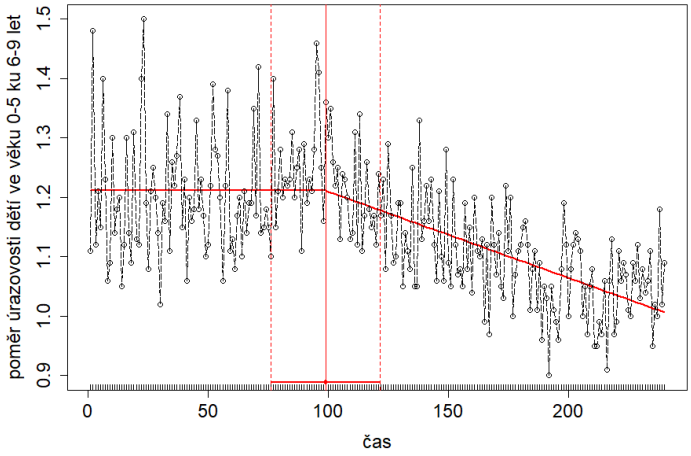
Obrázek 4.7: Směrnice přímek obou segmentů přerušované časové řady

Proto nyní vytvoříme nový model segmentové regrese bez proměnné *time*, se kterým budeme dále pracovat (viz výstup 4.8). Do příkazu `segmented` nyní vložíme lineární model $m1$, který bude vytvořen pouze pomocí absolutního členu, tedy `m1=lm(outcome ~ 1,data=d)`. Z výstupu segmentové regrese pro model $m1$ můžeme na obrázku 4.8 vidět, že opravdu pracujeme pouze s absolutním členem a koeficientem γ, který byl nyní po zaokrouhlení opět odhadnut jako -0.0015 (bez zaokrouhlení je výsledek nepatrně odlišný). Odhad sklonů obou segmentů je nyní velmi jednoduchý, jelikož koeficient α je roven nule, tedy přímka prvního segmentu bude konstantní, a sklon druhého segmentu v tomto případě odpovídá koeficientu γ. Z výstupu na obrázku 4.8 je pro nás stěžejní odhad bodu zlomu, který vidíme hned na jeho začátku. Na základě popsaného algoritmu v kapitole 2.2.1 tato funkce našla hledaný bod zlomu v čase $\widehat{T}_i = 98.9$ se standardní chybou 11.54. Po zaokrouhlení odpovídá časový bod 99 březnu roku 1998.

```
        ***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = m1, seg.Z = ~time)

Estimated Break-Point(s):
             Est. St.Err
psi1.time 98.949 11.535

Meaningful coefficients of the linear terms:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2112245  0.0086597 139.870   <2e-16 ***
U1.time     -0.0014507  0.0001755  -8.266       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08573 on 237 degrees of freedom
Multiple R-Squared: 0.3891,  Adjusted R-squared: 0.384

Boot restarting based on 6 samples. Last fit:
Convergence attained in 2 iterations (rel. change 1.7407e-12)
```

Obrázek 4.8: Upravený model segmentové regrese na základě modelu $m1$

Nyní do obrázku 4.9 vizualizujme data spolu s odhadnutou křivkou segmentové regrese, která vznikla na základě výše odhadnutých koeficientů výsledného modelu. Dostáváme lineárně lomenou funkci, která zahrnuje odhadnutý bod zlomu v čase 99. Pomocí příkazu `confint` lze dopočítat také 95% interval spolehlivosti pro $\widehat{T}_i = 99$, který vychází (76,122) a je v grafu také vyobrazen. Pro srovnání



Obrázek 4.9: Proložení časové řady na základě segmentové regrese pro model $m1$

s dalšími modely rovněž vypočítejme např. hodnotu informačního kritéria BIC, která pro tento model vychází -479.17.

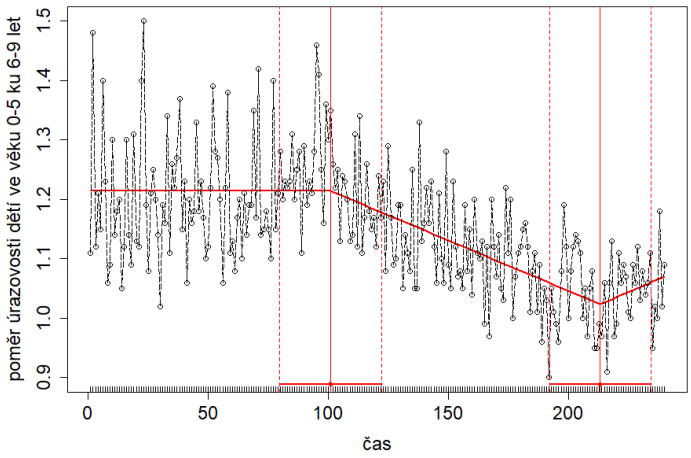
Dosud jsme se v této části věnovali detekci pouze jednoho bodu zlomu. Zbývá vyloučit možnost, že se v časové řadě nachází intervencí více. V této kapitole totiž předpokládáme, že nemáme informaci o jejich počtu ani umístění. Opět využijeme příkaz `segmented`, tentokrát však s doplněním volby `npsi=2`, což indikuje, že hledáme dva body zlomu. Výstup lze vidět na obrázku 4.10, kde místo jednoho odhadnutého bodu $\widehat{T}_i$ nyní vidíme dva, a to $\widehat{T}_{i_1} = 101$ a $\widehat{T}_{i_2} = 213$, a také odhady dvou koeficientů γ_1 a γ_2.

```
        ***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = m1, seg.Z = ~time, npsi = 2)

Estimated Break-Point(s):
               Est. St.Err
psi1.time 101.000 10.744
psi2.time 213.283 10.699

Meaningful coefficients of the linear terms:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2141886  0.0084534 143.633   <2e-16 ***
U1.time     -0.0016977  0.0002483  -6.837       NA
U2.time      0.0034376  0.0021137   1.626       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08496 on 235 degrees of freedom
Multiple R-Squared: 0.4051,  Adjusted R-squared: 0.395

Boot restarting based on 10 samples. Last fit:
Convergence attained in 2 iterations (rel. change -1.4246e-09)
```

Obrázek 4.10: Model segmentové regrese se dvěma body zlomu

Pokud si vypíšeme také BIC $= -474.58$, zjistíme, že je o něco vyšší než u modelu s jedním bodem zlomu (to bylo rovno -479.17). Adjustovaný koeficient determinace se ovšem o něco zvýšil. Druhou možností, jak určit počet bodů zlomu, je pomocí sekvenčního testování hypotéz, které jsme popsali v kapitole 2.2.4. Tímto postupem bychom pomocí Daviesova testu těsně nezamítli hypotézu, že $\gamma_2 = 0$ (p-hodnota $= 0.08$). Tato skutečnost naznačuje, že by jedna intervence mohla být pro tuto časovou řadu dostačující.

Přesto zkusme vykreslit odhadnutou křivku pomocí segmentové regrese pro dvě odhadnuté intervence spolu s jejich 95% intervaly spolehlivosti (viz obrázek 4.11). Lze vidět, že první odhadnutý bod zlomu vychází velmi podobně jako v segmentové regresi s jednou intervencí. Druhý bod zlomu vykazuje poměrně široký interval spolehlivosti, konkrétně (192,234), což může indikovat značnou nejistotu ve stanovení tohoto odhadu. Navíc se druhý odhadnutý bod nachází v okrajové koncové části časové řady, tudíž pro jeho správné posouzení bychom potřebovali data z dalších let. Z grafu se ovšem opravdu zdá, že ke konci časové řady dochází k ukončení poklesu a poměr úrazovosti dětí ve věku 0–5 ku věku dětí 6–9 začíná opět stoupat. Tento nárůst může být ovlivněn také jinými faktory či jinou změnou legislativy. Druhý bod zlomu by tedy vyžadoval důkladné zvážení.



Obrázek 4.11: Proložení časové řady pomocí segmentové regrese se 2 body zlomu

II. F-statistika (knihovna strucchange)

Druhou možností pro detekci intervence je využití F-statistiky, které jsme zmínili v kapitolách 2.2.1 a 2.1.3. Postupujeme tak, že si v každém časovém bodě na intervalu $[i_1, i_2]$ vypočítáme hodnotu F-statistiky pomocí funkce `Fstats` z knihovny `strucchange`. Interval $[i_1, i_2]$ volíme proto, abychom zabránili lokalizaci bodu

zlomu v krajních hodnotách časové řady a zaměřili se pouze na její hlavní část. Standardně funkce vynechá prvních a posledních 15 % časových bodů. Z vypočítaných F-statistik následně vybereme maximální hodnotu, jejíž časový bod bude symbolizovat odhadnutý bod zlomu (alternativně lze místo defaultního maxima/suprema zvolit například výpočet průměrné hodnoty). [24] Pomocí funkcí `Fstats` a `breakpoints` získáme výsledek na obrázku 4.12.

```
> fstat = Fstats(d$outcome ~ d$time)
> breakpoints(fstat)

        Optimal 2-segment partition:

Call:
breakpoints.Fstats(obj = fstat)

Breakpoints at observation number:
104
```

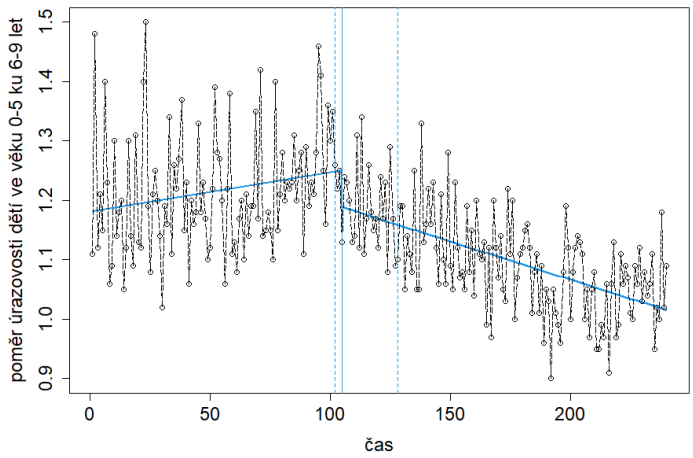
Obrázek 4.12: Odhadnutý bod zlomu na základě F-statistiky

Můžeme vidět, že na základě nejvyšší hodnoty F-statistiky funkce odhadla intervenci v časovém bodě $\widehat{T}_i = 104$ (srpen 1998). Provedeme také test významnosti, který na základě příkazu `sctest` vypíše testovou statistiku (nejvyšší hodnotu) a p-hodnotu, viz obrázek 4.13. Dále pomocí příkazu `confint` vypočítáme 95% interval spolehlivosti intervence a získáme interval (102,128). Dostáváme tedy lehce odlišný výsledek než pomocí funkce `segmented` v předchozí kapitole.

```
> sctest(fstat,type="supF")

        supF test

data:  fstat
sup.F = 36.327, p-value = 4.551e-07
```

Obrázek 4.13: Výstup sctestu (structural change test)

Výslednou křivku s odhadnutou intervencí v bodě 104 a jejím intervalovým odhadem vykreslíme do obrázku 4.14. Z tohoto grafu si můžeme všimnout, že knihovna `strucchange` umí nalézt také okamžitý účinek, tudíž se jedná o nespojitou funkci a v nalezeném bodě zlomu vidíme mírný „skok“. Předpisy pro přímky



Obrázek 4.14: Proložení časové řady pomocí knihovny strucchange

obou segmentů získáme pomocí funkce `coef`:

$$\widehat{Y}_1 = 1.181 + 0.0007 \times time$$
$$\widehat{Y}_2 = 1.322 - 0.0013 \times time.$$

Doplníme také nejistoty odhadů pro koeficienty přímek prvního a druhého segmentu. Do tabulky 4.1 vypíšeme standardní chybu (SE) a meze 95% intervalů spolehlivosti pro tyto koeficienty.

Koeficient	Odhad	SE	95% CI dolní	95% CI horní
μ_1	1.181	3.805e-04	1.143	1.29
α_1	0.0007	1.04e-07	0.0001	0.0013
μ_2	1.322	7.409e-04	1.269	1.375
α_2	-0.0013	2.367e-08	-0.0013	-0.0007

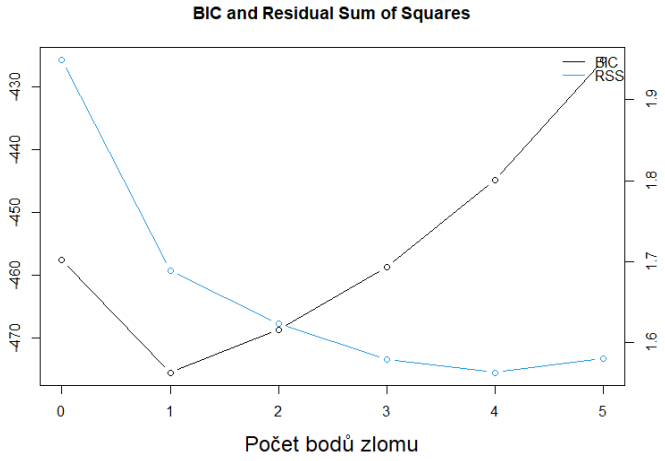
Tabulka 4.1: Odhadnuté koeficienty přímek obou segmentů s 95% intervaly spolehlivosti

Podobně jako pomocí knihovny `segmented`, i zde se podívejme na odhad více bodů zlomu. K tom využijeme opět funkci `breakpoints`, která vypíše, jaké body by odhadla pro 1 až 5 bodů zlomu (viz výstup 4.15, uložený pod názvem `bp`).

```
> bp = breakpoints(outcome ~ time,data=d)
> summary(bp)

        Optimal (m+1)-segment partition:

Call:
breakpoints.formula(formula = outcome ~ time, data = d)

Breakpoints at observation number:

m = 1      104
m = 2   58 102
m = 3   58 102         186
m = 4   58 101 137     186
m = 5   58 94  131 167 204
```

Obrázek 4.15: Body zlomu odhadnuté pomocí funkce breakpoints

Pro vyhodnocení, jaký počet intervencí je pro naši časovou řadu optimální, využijeme příkaz `plot(bp)`, který vykreslí jak BIC, tak reziduální součty čtverců (viz 4.16). Na základě minimálního BIC$=-475.58$ zvolíme pouze jeden bod zlomu.

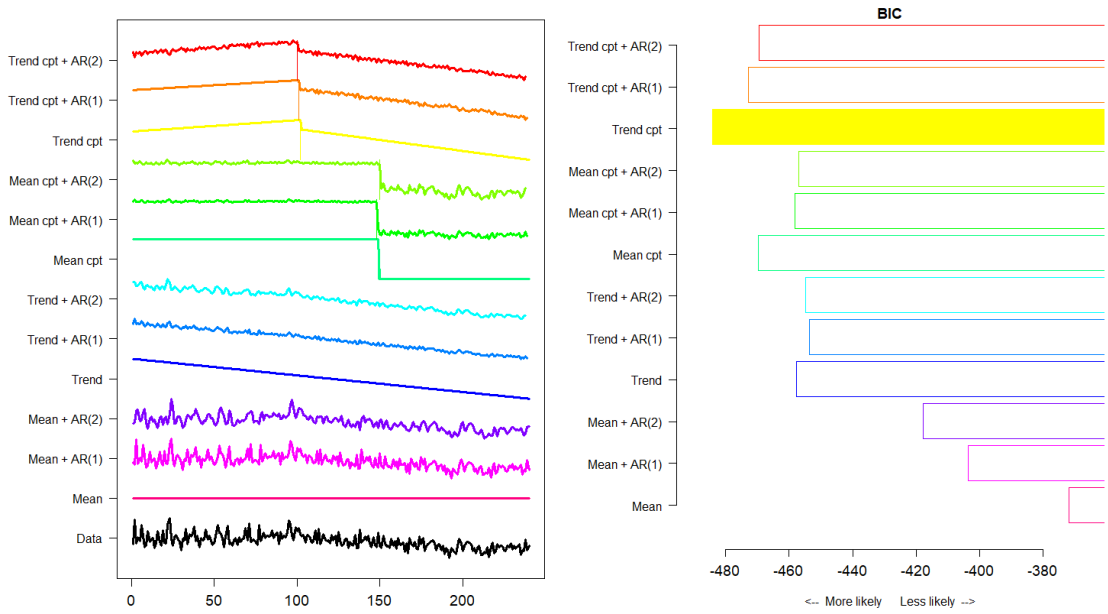


Obrázek 4.16: Hodnoty BIC a reziduálního součtu čtverců (RSS) pro výběr optimálního počtu bodů zlomu

III. Knihovna EnvCpt

Poslední způsob, který si pro detekci bodu zlomu zmíníme, je velmi jednoduché použití knihovny `EnvCpt`. Tato knihovna nabízí 12 modelů, jež lze vytvořit

pomocí funkce `envcpt`. [12] Grafické zobrazení těchto modelů a jejich BIC je demonstrováno na obrázku 4.17.



Obrázek 4.17: Výstup funkce envcpt (vlevo) a hodnoty BIC modelů (vpravo)

Z levého grafu můžeme vidět, že tato funkce vykreslí modely jak s přítomností bodu zlomu (prvních šest křivek), tak bez něj. Jelikož hledáme neznámý bod zlomu, zaměříme se právě na těchto prvních šest případů, které kombinují po částech lineární trend (ve výstupu jako „trend cpt“) a poté po částech konstantní trend (ve výstupu jako „mean cpt“) s autoregresními chybami AR(1) a AR(2) i bez nich. Body zlomu tato funkce odhaduje na základě metody maximální věrohodnosti.

V tabulce 4.2 jsou uvedeny body zlomu nalezené těmito modely spolu s jejich hodnotami BIC. Na základě této tabulky nebo pravého grafu na obrázku 4.17 získáváme nejmenší BIC pro třetí po částech lineární model, který detekoval bod zlomu v časovém bodě 102 (červen roku 1998).

Model	Bod zlomu	BIC
Trend cpt + AR(2)	100	-469.608
Trend cpt + AR(1)	101	-472.797
Trend cpt	102	-484.236
Mean cpt + AR(2)	150	-457.029
Mean cpt + AR(1)	148	-458.291
Mean cpt	149	-469.733

Tabulka 4.2: Odhadnuté body zlomu pomocí funkce EnvCpt a jejich BIC

Shrnutí a vyhodnocení detekce bodu zlomu

V této kapitole jsme nejprve vytvořili jednoduchý model regresní přímky, následně jsme hledali bod zlomu pomocí knihoven `segmented` a `strucchange` a nakonec jsme stručně představili využití knihovny `EnvCpt`. Výsledky přehledně shrneme do tabulky 4.3, kde vidíme nalezené body zlomu a jim odpovídající měsíc a rok, 95% intervaly spolehlivosti bodů zlomu a nakonec BIC jednotlivých modelů.

Model	Bod zlomu	Měsíc a rok	95% CI dolní	95% CI horní	BIC
lin.regrese	x	x	x	x	-457.66
segmented	99	březen 1998	76	122	-479.17
strucchange	104	srpen 1998	102	128	-475.58
EnvCpt	102	červen 1998	x	x	-484.24

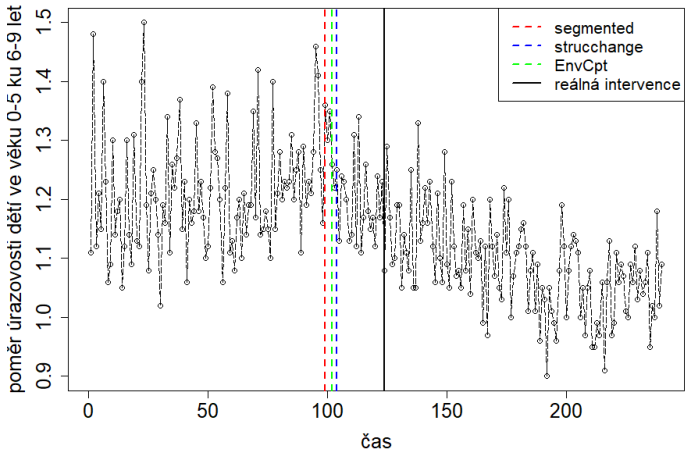
Tabulka 4.3: Shrnutí nalezených bodů zlomu

Z tabulky je zřejmé, že zahrnutím bodu zlomu do analýzy došlo ve všech případech ke snížení hodnoty BIC, což indikuje vytvoření lepšího modelu než pomocí jednoduché lineární regrese. Nejnižší hodnota informačního kritéria byla dosažena pomocí knihovny `EnvCpt` s bodem zlomu lokalizovaným v čase 102. Druhá nejmenší hodnota BIC byla zaznamenána při použití knihovny `segmented`, která identifikovala bod zlomu v čase 99. O něco málo vyšší hodnota BIC odpovídá regresi pomocí knihovny `strucchange`, která na základě F-statistiky odhadla intervenci v časovém bodě 104. Všechny tři metody tedy lokalizovaly bod zlomu v rozmezí od března do srpna roku 1998 (přibližně kolem časového bodu 100).

Nyní již můžeme odhalit, že hledaný bod zlomu je ve skutečnosti opravdu

jeden, a to až v čase 124 (což odpovídá dubnu roku 2000). V této době Japonský národní zákonodárný sbor schválil novelu zákona o silničním provozu, která nařídila povinné používání dětských zádržných systémů u dětí ve věku 0–5 let. Tato legislativní změna byla oznámena 11. května 1999 a vstoupila v platnost 1. dubna 2000, což považujeme za intervenci v naší časové řadě. [18]

Nejblíže jsme se skutečné intervenci přiblížili pomocí knihovny `strucchange`, která dokonce bod 124 zahrnula do svého 95% intervalu spolehlivosti. I přesto jsou však stále všechny odhadnuté body zlomu od toho skutečného relativně vzdálené, jak můžeme vidět na obrázku 4.18. Intervenci jsme ve všech případech odhadli o několik měsíců dříve, než změna zákona vešla v platnost. Navíc, z grafu je také patrné, že největší změna sklonu časové řady nastává přibližně kolem bodu 100, zatímco kolem bodu 124 není viditelný žádný velký účinek. Tento výsledek tedy naznačuje, že jsme detekovali pokles úrazovosti malých dětí ještě před zavedením zákona, což by mohlo odpovídat zvýšené opatrnosti lidí již před chystanou legislativní změnou.



Obrázek 4.18: Shrnutí odhadnutých časů intervence a skutečné intervence

Pokud bychom tyto metody porovnali, všechny nabízejí velmi jednoduché a uživatelsky přívětivé použití. Knihovna `strucchange` se vyznačuje tím, že není

potřeba předem specifikovat počet bodů zlomu a automaticky vykresluje BIC a RSS. Pokud nevyžadujeme proložení pomocí spojité funkce, je výhodou také detekce okamžitého účinku intervence, která u funkce `segmented` chybí. Naopak knihovna `segmented` poskytuje snadnou implementaci modelu segmentové regrese a sklony přímek pro každý segment. Knihovna `EnvCpt` může sloužit jako vhodný úvodní nástroj pro zkoumání časových řad a jednoduchým příkazem vyhodnotí také bod zlomu. Její výhodou je zahrnutí autoregresních chyb do modelů, ale naopak nevýhodou je neuvádění intervalů spolehlivosti.

4.3. Analýza časové řady se znalostí intervence

Na konci minulé kapitoly jsme již zmínili, že změna legislativy týkající se dětských bezpečnostních pásů proběhla v časovém bodě 124 (v dubnu 2000). Viděli jsme, že detekce časového bodu nemusí být úplně snadná, proto je pro analýzu přerušovaných časových řad výhodou znát časový bod intervence dopředu. Na druhou stranu jsme při vyhodnocování detekce bodu zlomu došli k myšlence, že efekt změny legislativy se začal projevovat ještě dřív než skutečně začala platit. Proto je ideální kombinací znalost intervence, ale zároveň analýza jako v kapitole výše. Na situaci při znalosti intervence se zaměříme v této kapitole, kde proložíme stejná data nejprve pomocí segmentové regrese a poté pomocí ARIMA modelů. Zaměříme se také na predikce a na vyjádření efektu intervence.

4.3.1. Analýza pomocí segmentové regrese

V této části kapitoly vytvoříme základní model segmentové regrese při znalosti intervence, který jsme popsali v kapitole 2.1.1. Konkrétně budeme chtít odhadnout koeficienty modelu (2.1), který vypadal následovně:

$$Y_t = \mu + \alpha t + \beta \times intervence_t + \gamma \times cas\ od\ intervence_t + \epsilon_t\ .$$

Abychom mohli tento model vytvořit, musíme mít v datech binární proměnnou intervence (v R značíme jako *interven*) a proměnnou popisující čas od intervence (v R jako *sinceinterven*), které jsou součástí modelu a jsou popsány na začátku kapitoly 2.1.1. Hodnoty závisle proměnné Y_t budeme v kódech opět značit jako *outcome* a čas t jako *time*. Poté lze na základě lineární regrese sestavit model segmentové regrese:

$$mm{=}\mathrm{lm}(outcome \sim time + interven + sinceinterven, \mathrm{data}{=}\mathrm{d}),$$

jehož výsledek vidíme na obrázku 4.19. Opět dostáváme jedinou nevýznamnou proměnnou *time*.

```
Call:
lm(formula = outcome ~ time + interven + sinceinterven, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.184294 -0.064660 -0.004851  0.050056  0.297399

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.1970385  0.0155360  77.049  < 2e-16 ***
time           0.0002418  0.0002174   1.112  0.26721
interven1     -0.0725192  0.0222245  -3.263  0.00127 **
sinceinterven -0.0014068  0.0003243  -4.339 2.13e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08563 on 234 degrees of freedom
Multiple R-squared:  0.3938,    Adjusted R-squared:  0.3861
F-statistic: 50.68 on 3 and 234 DF,  p-value: < 2.2e-16
```

Obrázek 4.19: Model segmentové regrese mm se známou intervencí v čase 124

Vytvoříme proto nový model $mm1$ bez proměnné *time* pomocí příkazu

$$mm1{=}\mathrm{lm}(outcome \sim interven + sinceinterven, \mathrm{data}{=}\mathrm{d}),$$

jehož výstup lze vidět na obrázku 4.20, kde jsou již všechny odhady parametrů statisticky významné. Dle odhadnutých koeficientů nového modelu $mm1$ máme výsledný model:

$$\widehat{Y}_t = 1.212 - 0.059 \times interven_t - 0.001 \times sinceinterven_t + \epsilon_t \ .$$

```
Call:
lm(formula = outcome ~ interven + sinceinterven, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.192033 -0.063087 -0.004921  0.048320  0.287967

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.2120325  0.0077031 157.344  < 2e-16 ***
interven1     -0.0590131  0.0176660  -3.340 0.000971 ***
sinceinterven -0.0011333  0.0002339  -4.846 2.28e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08543 on 237 degrees of freedom
Multiple R-squared:  0.3933,    Adjusted R-squared:  0.3882
F-statistic: 76.83 on 2 and 237 DF,  p-value: < 2.2e-16
```

Obrázek 4.20: Model segmentové regrese $mm1$ bez proměnné *time* se skutečnou intervencí v čase 124

Koeficient $\widehat{\mu} = 1.212$ symbolizuje vyrovnanou hodnotu na počátku časové řady, koeficient $\widehat{\alpha} = 0$ udává sklon přímky prvního segmentu (bude konstantní), odhad $\widehat{\beta} = -0.059$ označuje velikost okamžitého účinku a $\widehat{\gamma} = -0.001$ vyjadřuje dlouhodobý účinek (o kolik se liší směrnice přímky druhého segmentu oproti směrnici přímky prvního segmentu). Sklon druhé přímky odpovídá součtu $\widehat{\alpha}+\widehat{\gamma} = -0.001$.

Druhou možností, jak vytvořit model segmentové regrese bez toho, aniž bychom museli přidávat pomocné proměnné, je následující příkaz v softwaru R:

$$\text{lm}(outcome \sim time * \text{I}(time \geq 124), \text{data=d}).$$

Jedná se o model s interakcí se zavedením indikátorové proměnné I($time \geq 124$), která symbolizuje druhý segment časové řady. Pomocí tohoto modelu získáme stejné přímky, stejný koeficient determinace i BIC, pouze máme jiné proměnné.

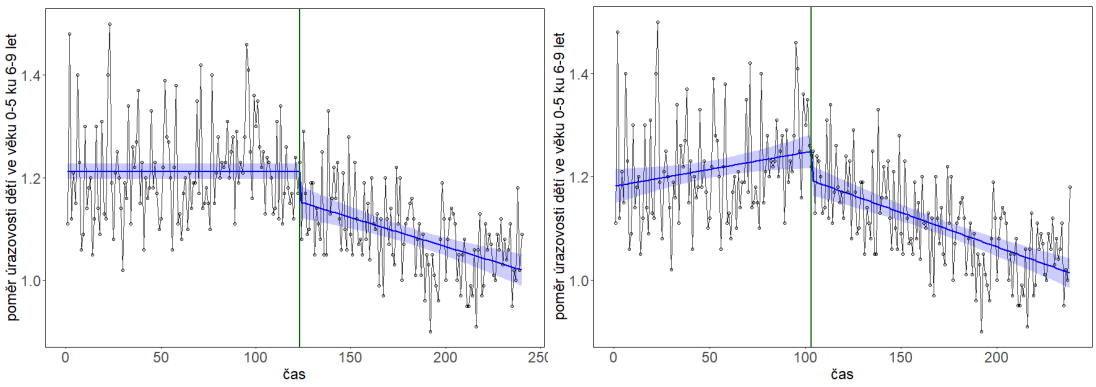
Dále je vhodné otestovat, zda je přechod od základního lineárního modelu k složitějšímu opodstatněný. Lineární model byl vytvořen v předchozí kapitole:

$$m\text{=lm}(outcome \sim time, \text{data=d})$$

a nyní máme složitější model $mm1$ se zahrnutím intervence. Test provedeme pomocí příkazu `anova(mm1,m,test="Chisq")`, který zamítne nulovou hypotézu

o platnosti podmodelu (p-hodnota=4.225e–08). Zamítáme tedy možnost přechodu k základnímu lineárnímu modelu m, jelikož by se jednalo o přílišné zjednodušení. Model segmentové regrese je v této situaci tedy na místě.

Nyní do grafu na levém obrázku 4.21(a) vykreslíme odhadnutou přímku segmentové regrese s pásem spolehlivosti na základě modelu $mm1$ (obrázek 4.20). Zároveň vyznačíme známý čas intervence v čase 124 zelenou svislou přímkou. Z grafu je patrné, že se jedná o nespojitou přímku, jelikož v čase intervence dochází k mírnému skoku (viz koeficient $\widehat{\beta} = -0.059$) a přímka druhého segmentu proto na přímku prvního segmentu nenavazuje. Tento model porovnejme s modelem, kde za bod zlomu považujeme odhadnutou intervenci v čase 102 z minulé kapitoly. Model s touto intervencí vidíme na obrázku 4.22 a výslednou křivku na pravém grafu 4.21(b).



(a) Reálná intervence v čase 124 (b) Odhadnutá intervence v čase 102

Obrázek 4.21: Model segmentové regrese se známým časem intervence

Dostáváme poměrně odlišné výsledky, jelikož v modelu s časem intervence 102 máme tentokrát proměnnou *time* významnou, a proto již není přímka prvního segmentu konstantní. Na pravém grafu jsme tedy zachytili výraznější bod zlomu a větší změnu směrnice druhého segmentu oproti prvnímu segmentu (vlevo $\widehat{\gamma} = -0.001$, CI=$(-0.0016, -0.0006)$ a vpravo $\widehat{\gamma} = -0.002$, CI=$(-0.0027, -0.0014)$), tedy větší dlouhodobý účinek. Na druhou stranu, okamžitý účinek intervence

je v obou případech obdobný. Bodové odhady okamžitého účinku a jejich 95% intervaly spolehlivosti jsou následující: vlevo $\widehat{\beta} = -0.059$, CI=$(-0.094, -0.024)$ a vpravo $\widehat{\beta} = -0.049$, CI=$(-0.094, -0.006)$.

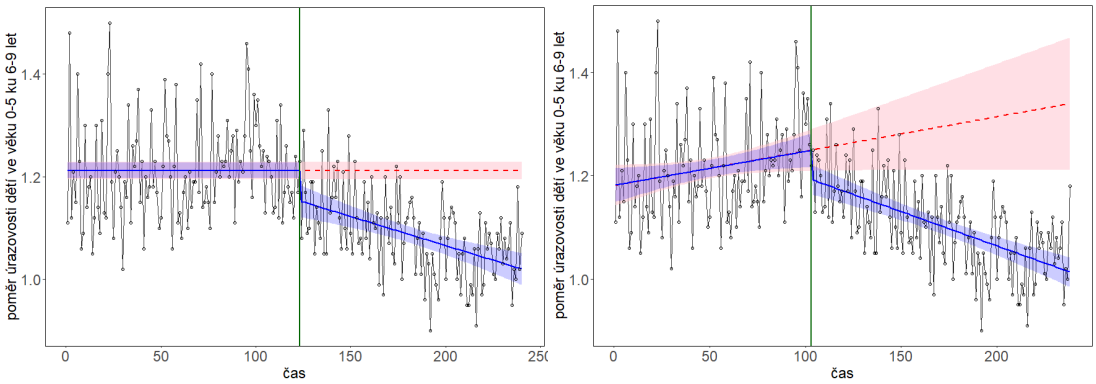
```
Call:
lm(formula = outcome ~ time + interven + sinceinterven, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18095 -0.06270 -0.00695  0.05273  0.30371

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.1809584  0.0170409  69.302  < 2e-16 ***
time           0.0006665  0.0002901   2.298   0.0225 *
interven1     -0.0499463  0.0222517  -2.245   0.0257 *
sinceinterven -0.0020239  0.0003433  -5.895  1.3e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08499 on 234 degrees of freedom
Multiple R-squared:  0.4028,    Adjusted R-squared:  0.3951
F-statistic:  52.6 on 3 and 234 DF,  p-value: < 2.2e-16
```

Obrázek 4.22: Model segmentové regrese s odhadnutou intervencí v čase 102

Odhad efektu intervence

Nyní se zaměříme na číselné ohodnocení efektu intervence, tedy na to, jak posoudit, jaký měla intervence vliv na průběh časové řady. Jednoduchým přístupem je proložit data tak, jako by intervence neměla na časovou řadu vliv. To můžeme vidět na obrázcích 4.23(a) a 4.23(b), kde jsme červenou přímkou znázornili budoucí vývoj časové řady s nevýznamnou intervencí v bodě 124 (vlevo) i v bodě 102 (vpravo). Již na první pohled je zřejmé, že efekt intervence v těchto dvou případech je poměrně odlišný. Zaměřme se nyní spíše na efekt reálné intervence, tedy na čas 124 a obrázek 4.23(a).

Zvolíme například časový bod 180 (odpovídá prosinci roku 2004, tedy 56 měsíců po intervenci) a v něm nejprve vypočítáme vyrovnanou hodnotu segmentové regrese s intervencí (funkční hodnotu modré křivky v bodě 180 na obrázku



(a) Reálná intervence v čase 124 (b) Odhadnutá intervence v čase 102

Obrázek 4.23: Model segmentové regrese pro vyjádření efektu intervence

4.23(a)) pomocí modelu (2.3):

$$\widehat{Y}_{(s\ intervenci)} = \hat{\mu} + \hat{\alpha}t + \hat{\beta} \times I(t \geq T_i) + \hat{\gamma}(t - T_i) \times I(t \geq T_i).$$

Dosadíme za t=180 a jelikož je zde intervence T_i v bodě 124, budou obě indikátorové proměnné $I(t \geq T_i)$ rovny 1:

$$\widehat{Y}_{(s\ intervenci)} = \hat{\mu} + \hat{\alpha} \times 180 + \hat{\beta} \times 1 + \hat{\gamma} \times (180 - 124) \times 1.$$

Po dosazení také odhadnutých koeficientů z modelu segmentové regrese $mm1$ získáme výsledek:

$$\widehat{Y}_{(s\ intervenci)} = 1.212 + 0 \times 180 - 0.059 \times 1 - 0.001 \times 56 = 1.097.$$

Nyní vypočítáme predikovanou funkční hodnotu v bodě 180 pro přímku, v jejímž případě předpokládáme, že žádná intervence nenastala (v grafu červená přímka). Dosadíme tedy pouze do předpisu konstantní přímky:

$$\widehat{Y}_{(bez\ intervence)} = \hat{\mu} = 1.212.$$

Absolutní efekt intervence nakonec vypočítáme jako rozdíl těchto dvou funkčních hodnot:

$$\widehat{Y}_{(s\ intervenci)} - \widehat{Y}_{(bez\ intervence)} = -0.115.$$

Relativní změnu bychom vyjádřili následovně

$$\frac{\widehat{Y}_{(s\ intervenci)} - \widehat{Y}_{(bez\ intervence)}}{\widehat{Y}_{(bez\ intervence)}} = -0.106.$$

Můžeme tedy konstatovat, že poměr úrazovosti dětí ve věku $0-5$ let ku dětem ve věku $6-9$ let se po změně legislativy v dubnu roku 2000 prokazatelně snížil. Dlouhodobý účinek po 56 měsících od intervence jsme vyjádřili nejprve absolutní změnou, kde poměr úrazovosti klesl o 0.115 v porovnání s tím, kdyby k žádné změně legislativy nedošlo. Vyjádřením pomocí relativní změny můžeme říct, že došlo k poklesu až o 10.6 %. Jak okamžitý tak dlouhodobý efekt této intervence je ve zkoumané časové řadě s intervencí v čase 124 prokazatelný.

Nakonec opět stojí za zmínku, že ještě větší efekt je prokazatelný pro bod zlomu v čase 102, tedy již před reálnou změnou legislativy, kdy lidé pravděpodobně začali dodržovat zákon ještě než vstoupil v platnost. Absolutní efekt pro čas 180 vychází v tomto případě -0.21 a relativní změna vyjadřuje pokles až o 16.1 %. Pokud bychom chtěli vypočítat vliv intervence v čase 102 také po 56 měsících, dostáváme absolutní efekt -0.16 neboli relativní pokles o 12.7 %.

4.3.2. Analýza pomocí ARIMA modelů

V poslední části se zaměříme na analýzu přerušované časové řady prostřednictvím ARIMA modelů, které jsme popsali v kapitole 3.4. V této situaci máme opět k dispozici přesný čas intervence, konkrétně čas 124. Cílem je nyní proložit data, vyjádřit efekt intervence a pokusit se predikovat další průběh časové řady.

Analýza stacionarity a autokorelace

Před aplikací ARIMA modelů je nejprve nutné ověřit, zda je časová řada stacionární, a zda se v datech nevyskytuje autokorelace. To provádíme pomocí testů, které jsou popsány v kapitole 3.3.

Nejprve se zaměříme na stacionaritu, kterou můžeme otestovat pomocí KPSS testu nebo rozšířeného Dickeyova-Fullerova testu. Jak již bylo zmíněno v kapitole 3.3, je důležité si uvědomit, že tyto testy mají odlišnou nulovou hypotézu. Pokud zamítneme nulovou hypotézu KPSS testu, časová řada není stacionární (v R pomocí funkce `kpss.test` z knihovny `tseries`). Tento výsledek pozorujeme také v našich datech, jelikož p-hodnota KPSS testu je nižší než hladina významnosti 0.05 (viz obrázek 4.24), tudíž nulovou hypotézu o stacionaritě zamítáme.

```
        KPSS Test for Level Stationarity

data:  d$outcome
KPSS Level = 3.32, Truncation lag parameter = 4, p-value = 0.01
```

Obrázek 4.24: KPSS test stacionarity

Pokud se však podíváme na výsledek rozšířeného Dickeyova-Fullerova testu (funkce `adf.test`), dostáváme opačné tvrzení, a to že časová řada je stacionární. Zamítáme zde totiž nulovou hypotézu o tom, že časová řada není stacionární (viz obrázek 4.25). Dochází tedy k rozdílným výsledkům mezi KPSS testem a Dickeyovým-Fullerovým testem a rozhodnutí o diferencování časové řady bude předmětem dalšího zvážení. Např. článek [11] zmiňuje, že Dickeyův-Fullerův test zcela nezahrnuje strukturální změny či posuny v datech, tudíž bychom se přiklonili k výsledku KPSS testu.

```
        Augmented Dickey-Fuller Test

data:  d$outcome
Dickey-Fuller = -3.6086, Lag order = 6, p-value = 0.03306
alternative hypothesis: stationary
```

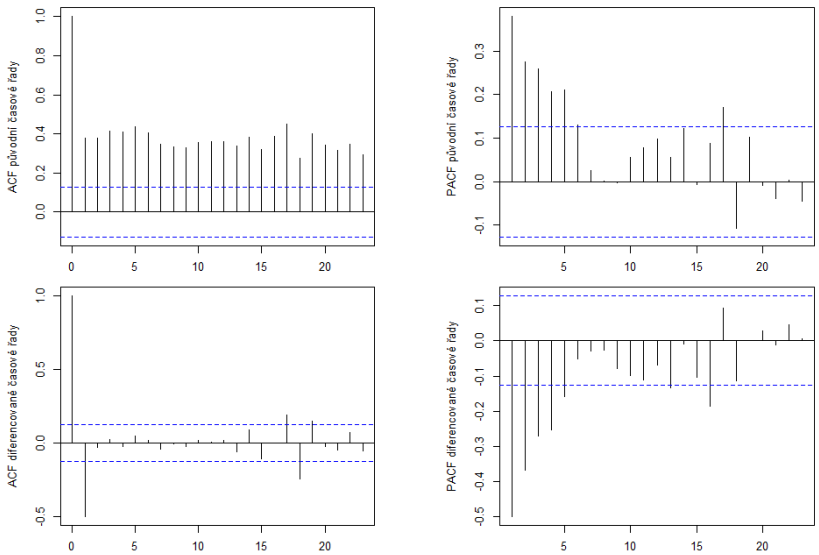
Obrázek 4.25: Rozšířený Dickeyův-Fullerův test stacionarity

Následně se zaměříme na otestování autokorelace, tedy zda mezi sebou hodnoty časové řady vykazují nějakou závislost. Po aplikaci Ljung-Boxova testu (v R jako `Box.test`) zjistíme, že se v datech autokorelace vyskytuje, neboť na základě p-hodnoty zamítáme nulovou hypotézu (viz obrázek 4.26).

```
        Box-Ljung test

data:  d$outcome
X-squared = 34.886, df = 1, p-value = 3.495e-09
```

Obrázek 4.26: Ljung-Boxův test autokorelace

Vzhledem k přítomnosti autokorelace v datech je vhodné vykreslit autokorelační a parciální autokorelační funkce, které jsou popsány v kapitole 3.2. Tím získáme lepší představu o charakteru přítomné autokorelace. Do grafů na obrázku 4.27 vykreslíme autokorelační funkce (ACF) a parciální autokorelační funkce (PACF) pro dvě situace - pro původní časovou řady (horní dva grafy) a pro diferencovanou časovou řadu (dolní dva grafy). Vzhledem k nejednoznačnému výsledku předchozích testů, týkajících se stacionarity časové řady, je vhodné provést analýzu obou možností.

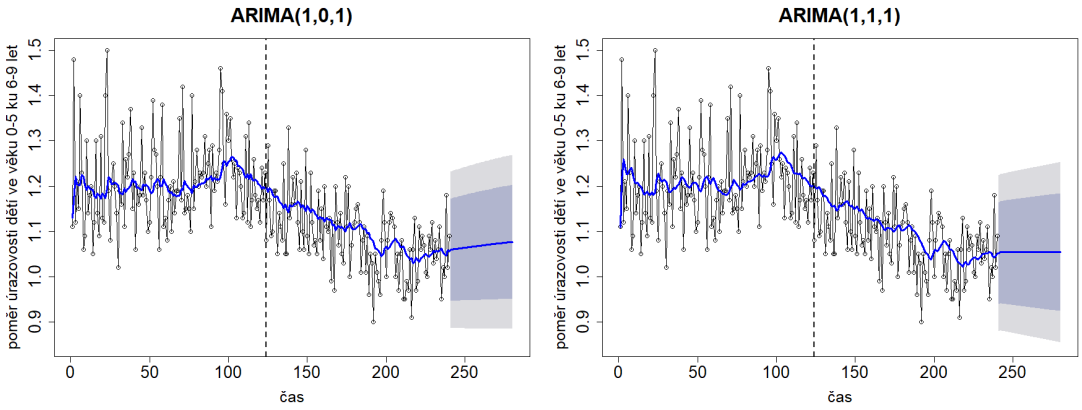


Obrázek 4.27: ACF a PACF pro původní data (nahoře) a pro diferencovaná (dole)

Na základě tvaru časové řady a existence výrazného bodu zlomu lze usoudit, že časová řada spíše není stacionární. Také z grafů ACF a PACF získáváme jasnější informace po diferencování časové řady. Z levého dolního grafu lze usoudit, že se v datech nachází autokorelace 1. řádu, což naznačuje přítomnost procesu MA(1). Z pravého dolního grafu odhadujeme parciální autokorelaci přibližně 4. řádu.

Tvorba modelu

Jak jsme mohli vidět, určení stupňů ARIMA modelů podle autokorelační a parciální autokorelační funkce nemusí být vždy jednoduché. Připomeňme, že hledáme model ARIMA(p,d,q), kde p představuje stupeň autoregresního procesu, d označuje stupeň diferencování a q značí stupeň modelu klouzavých součtů. V softwaru R budeme využívat funkci `Arima`, pomocí které vyzkoušíme různé modely, a na základě informačních kritérií rozhodneme o tom nejvhodnějším.



(a) ARIMA model (1,0,1)

(b) ARIMA model (1,1,1)

Obrázek 4.28: Časová řada proložená křivkou na základě ARIMA modelů s 95% predikčním intervalem (šedě) a 80% predikčním intervalem (modře)

Podle nejnižší hodnoty AIC je pro naši časovou řadu nejlepším modelem ARIMA(1,0,1) a podle BIC je to model ARIMA(1,1,1). Získali jsme tedy jeden model s diferencí a druhý bez ní. Proložení dat na základě těchto dvou modelů

včetně jejich predikcí vidíme na obrázku 4.28. Na levém grafu 4.28(a) pozorujeme odhadnutou křivku pomocí modelu ARIMA(1,0,1), kde je predikce budoucích hodnot mírně rostoucí. Na pravém grafu 4.28(b) vidíme proložení dat na základě modelu ARIMA(1,1,1) s využitím diferencování, kde jsou predikce téměř konstantní. Oba modely vykazují pro predikce poměrně široké intervaly spolehlivosti. Smíšený proces ARIMA(1,0,1) neboli ARMA(1,1) představuje posloupnost náhodných veličin $\{Y_t\}$ danou vztahem:

$$Y_t = \phi_1 Y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t = 0.99 Y_{t-1} - 0.91 \epsilon_{t-1} + \epsilon_t,$$

kde jsme odhady koeficientů ϕ_1 a θ_1 získali pomocí příkazu `summary` v R. Druhým modelem je integrovaný smíšený model ARIMA(1,1,1), který je vyjádřen posloupností diferencí $\{Y_t^d\}$ pro $d = 1$:

$$Y'_t = \phi_1 Y'_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t = -0.06 Y'_{t-1} - 0.9 \epsilon_{t-1} + \epsilon_t.$$

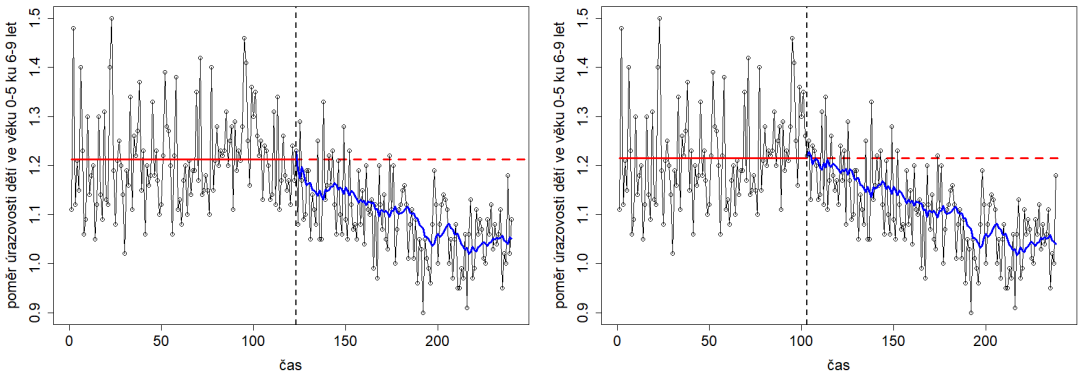
V obou případech je proložení dat velmi podobné a je zde viditelně patrná změna průběhu časové řady okolo bodu 100. Černou svislou čarou je na obrázcích 4.28 vyznačena reálná intervence v čase 124, kolem které vidíme, že se žádně zásadní změny nedějí. Můžeme tedy konstatovat, že modely ARIMA si s touto časovou řadou úspěšně poradily i bez zahrnutí bodu intervence. Také se na základě predikcí potvrdilo to, že po nějakém čase od intervence začaly poměry úrazovosti malých dětí opět mírně růst nebo se staly konstantní.

Odhad efektu intervence

Podobně jako v předchozí kapitole pomocí segmentové regrese, i zde se pokusíme odhadnout efekt intervence. Tentokrát však nebudeme prokládat segmenty lineárními funkcemi, ale použijeme ARIMA modely. Postup bude v zásadě stejný. Každý segment zvlášť proložíme odhadnutou křivkou a pro první segment záro-

veň provedeme predikci, jako kdyby žádná intervence nenastala. Zatímco výše jsme odhadli jeden model pro celou časovou řadu, nyní pomocí funkce `Arima` odhadneme stupně ARIMA modelů pro oba segmenty zvlášť.

Opět budeme nejprve pracovat s reálnou intervencí v čase 124. Pro první segment jsme získali odhadnutý model ARIMA(0,0,0), což odpovídá bílému šumu, a pro druhý segment dostáváme ARIMA(0,1,1). Z levého grafu 4.29(a) tedy vidíme, že predikce bude konstantní přímka, jelikož ARIMA modely nezaznamenaly v prvním segmentu žádný růst (stejně jako pomocí segmentové regrese). Překvapivě ani v druhé situaci, kdy za intervenci považujeme odhadnutý bod zlomu v čase 102, se stupně ARIMA modelů nezmění. Opět dostáváme bílý šum pro první segment a ARIMA(0,1,1) pro druhý segment (viz obrázek 4.29(b)).



(a) Reálná intervence v čase 124

(b) Odhadnutá intervence v čase 102

Obrázek 4.29: Proložení časové řady pomocí ARIMA modelů pro oba segmenty zvlášť

Tuto analýzu pro dva segmenty zvlášť můžeme využít k číselnému vyjádření efektu intervence a porovnat s výsledky na základě segmentové regrese. Opět se zaměříme na reálnou intervenci v čase 124 a vyjádříme rozdíl vyrovnaných hodnot v polovině druhého segmentu, tedy v čase 180 (56 měsíců po intervenci). Vyrovnanou hodnotu modelu po intervenci získáme pomocí modelu ARIMA(0,1,1), což odpovídá číslu 1.105 a predikovanou hodnotu prvního segmentu jako funkční hod-

notu červené přímky, tedy 1.212. Absolutní efekt intervence v čase 180 vychází následovně:

$$\widehat{Y}_{(s\ intervenci)} - \widehat{Y}_{(bez\ intervence)} = 1.105 - 1.212 = -0.107$$

a relativní efekt je roven -0.088. Oproti efektu intervence zjištěného pomocí segmentové regrese, kde jsme v čase 180 vypočítali pokles o 10.6 %, jsme nyní odhadli o něco menší efekt, a to pokles o 8.8 %. To je způsobeno tím, že ARIMA modely se více přizpůsobují datům a zachytí lokální změny. Pro případ s intervencí v čase 102 na obrázku 4.29(b) je tentokrát efekt v čase 180 velmi podobný (absolutní efekt je roven -0.109 a relativní pokles odpovídá 9.1 %). Efekt po 56 měsících na pravém grafu je dokonce o něco menší než na levém grafu, a to -0.081 (pokles o 6.7 %).

Na závěr můžeme konstatovat, že jak pomocí segmentové regrese, tak i pomocí ARIMA modelů jsme po změně legislativy v dubnu 2000 zaznamenali významný pokles poměru úrazovosti dětí ve věku 0–5 let oproti dětem ve věku 6–9 let. Zároveň jsme ale identifikovali nejvýraznější bod zlomu již dříve před zavedením této změny (přibližně červen 1998) a usoudili jsme, že tento fakt může být způsoben předčasnou opatrností před chystanou změnou zákona.

Závěr

Cílem této diplomové práce bylo seznámit čtenáře s problematikou přerušovaných časových řad. Hlavním úkolem analýzy těchto specifických časových řad je zejména detekce nejvýznamnějších bodů zlomu v časové řadě a následné vyjádření efektu těchto intervencí. Pokud je nám čas intervence znám, je našim úkolem časovou řadu také proložit a predikovat její budoucí průběh.

V teoretické části jsme se nejprve věnovali základnímu přístupu k přerušovaným časovým řadám, a to segmentové regresi. Popsali jsme situaci, kdy je pro nás čas nastání intervence znám, a ukázali jsme, jak pomocí této informace modelovat časovou řadu a vyjádřit efekt intervence. Dále jsme představili také algoritmus segmentové regrese, který umožňuje identifikaci bodů zlomu v případě, kdy čas intervence nemáme k dispozici. V druhé kapitole teoretické části jsme se věnovali modelům Boxovy-Jenkinsovy metodologie, které jsme později aplikovali také na proložení přerušované časové řady a vyjádření efektu intervence.

V praktické části práce jsme teoretické poznatky aplikovali na reálnou časovou řadu, týkající se měsíčních poměrů úrazovosti v automobilech u dětí vě věku 0–5 let ku dětem ve věku 6–9 let (v Japonsku, v letech 1990 – 2009). Za intervenci zde byla považována změna legislativy týkající se dětských zádržných systémů. Naším cílem bylo analyzoval vliv této změny a zjistit, zda došlo k prevenci dopravních úrazů u nejmladších dětí.

Největší bod zlomu jsme se nejprve pokusili detekovat, aniž bychom věděli, kde se reálná intervence nachází. V naší analýze jsme aplikovali tři metody, a to seg-

mentovou regresi, přístup pomocí výpočtu F-statistik (knihovna `strucchange`) a nakonec pomocí knihovny `EnvCpt`. Všechny tři přístupy dokázaly najít hledaný bod zlomu v podobném časovém úseku, konkrétně mezi březnem a srpnem roku 1998. Po odhalení skutečné intervence jsme ovšem zjistili, že reálná změna legislativy nastala až v dubnu roku 2000. Tento výsledek pravděpodobně naznačuje, že lidé začali dodržovat nový zákon preventivně již před jeho oficiálním vstupem v platnost. Díky této předčasné opatrnosti došlo k výraznému snížení počtu úrazů u malých dětí ještě před oficiální změnou legislativy. Zároveň jsme ale zjistili, že přibližně po 9 letech od této změny legislativy začal poměr úrazovosti dětí ve věku 0–5 ku dětem ve věku 6–9 let opět mírně stoupat.

Celý proces analýzy přerušovaných časových řad bychom mohli shrnout následovně. V případech, kdy máme k dispozici přesné informace o času intervence, je analýza relativně jednodušší a můžeme časovou řadu modelovat pomocí segmentové regrese nebo ARIMA modelů. V praktické části se však ukázalo, že je zároveň důležité zkusit intervenci detekovat, jelikož se tak může projevit předčasný či zpožděný efekt skutečné intervence. Přesný čas intervence pravděpodobně nedokážeme nikdy určit stoprocentně, ale pomocí zmíněných metod můžeme odhalit, kdy došlo v časové řadě k největší změně či změnám.

Během psaní této diplomové práce jsem získala mnoho nových poznatků v oblasti časových řad a osvojila si práci se softwarem R. Detailní zkoumání přerušovaných časových řad mě velmi bavilo. Byla jsem překvapena, jakých podrobných výsledků můžeme dosáhnout i ze zdánlivě jednoduchého zadání.

Literatura

[1] Bernal, J. L., Cummins, S., Gasparrini, A. (2017). *Interrupted time series regression for the evaluation of public health interventions: a tutorial.* International journal of epidemiology, 46(1), 348–355. Dostupné z: https://doi.org/10.1093/ije/dyw098.

[2] Box, G. E. P., Pierce, D. A. (1970) *Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models.* Journal of the American Statistical Association 65, no. 332: 1509–26. Dostupné z: https://doi.org/10.2307/2284333.

[3] Cipra, T. (1986). *Analýza časových řad s aplikacemi v ekonomii.* Státní nakladatelství technické literatury. Praha.

[4] D'Angelo, N., Priulla, A. (2020). *Estimating the number of changepoints in segmented regression models: comparative study and application.* Seas Working Papers. Dostupné z: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3724965

[5] Fišerová, E. (2015). *Lineární statistické modely.* 2. dopl. vydání, s. 83-84. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-4797-1.

[6] Fusi, F., LECY, J. (2020). *Interrupted time series.* [online], [cit. 2023-11-09]. Dostupné z: https://ds4ps.org/pe4ps-textbook/docs/p-020-time-series.html#fig:mixedeffect.

[7] Gimeno, R., Manchado, B., Mínguez, R. (1999). *Stationarity tests for financial time series*, Physica A: Statistical Mechanics and its Applications. Dostupné z: https://doi.org/10.1016/S0378-4371(99)00081-3

[8] Hobijn, B., Franses, P. H.; Ooms, M. (2004). *Generalizations of the KPSS-test for stationarity.* Statistica Neerlandica, 58.4: 483-502. Dostupné z: https://doi.org/10.1111/j.1467-9574.2004.00272.x

[9] Hron, K., Kunderová, P., Vencálek, O. (2018). *Základy počtu pravděpodobnosti a metod matematické statistiky*. 3. přepracované vydání. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-5398-9.

[10] Hyndman, R. J., Athanasopoulos, G. (2021). *Forecasting: principles and practice* (Third edition). OTexts.

[11] Jalil, A., Rao, N. H. (2019). *Time series analysis (stationarity, cointegration, and causality)*. Environmental Kuznets Curve (EKC). Academic Press, p. 85-99. Dostupné z: https://doi.org/10.1016/B978-0-12-816797-7.00008-4

[12] Killick R., Beaulieu C., Taylor S., Hullait H. (2021). *EnvCpt: Detection of structural changes in climate and environment time series*. R package version 1.1.3, Dostupné z: https://CRAN.R-project.org/package=EnvCpt

[13] Kwiatkowski, D., Phillips, P. C., Schmidt, P., Shin, Y. (1992). *Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?*. Journal of econometrics, 54(1-3), 159-178. Dostupné z: https://www.sciencedirect.com/science/article/pii/030440769290104Y

[14] Long, H., Yang, Y., Geng, X., Mao, Z., Mao, Z. (2022). *Changing characteristics of pharmaceutical prices in China under centralized procurement policy: a multi-intervention interrupted time series.* Frontiers in Pharmacology, 13, 944540. Dostupné z: https://doi.org/10.3389/fphar.2022.944540

[15] McDowall, D., McCleary, R., Bartos J. Bradley. (2019). *Interrupted Time Series Analysis*, Oxford University Press.

[16] Muggeo, V. M. (2003). *Estimating regression models with unknown break-points*. Statistics in medicine, 22(19), 3055–3071. Dostupné z: https://doi.org/10.1002/sim.1545

[17] Muggeo, V. M. (2008). *Segmented: an R package to fit regression models with broken-line relationships*. R news, 8.1: 20-25. Dostupné z: https://journal.r-project.org/articles/RN-2008-004/RN-2008-004.pdf

[18] Nakahara, S., Ichikawa, M., Nakajima, Y. (2015). *Effects of Increasing Child Restraint Use in Reducing Occupant Injuries Among Children Aged 0–5 Years in Japan, Traffic Injury Prevention*, 16:1, 55-61, Dostupné z: 10.1080/15389588.2014.897698

[19] Paparoditis, E., Politis, D. N. (2018). *The asymptotic size and power of the augmented Dickey–Fuller test for a unit root*. Econometric Reviews, 2018, 37.9: 955-973. Dostupné z: 10.1080/00927872.2016.1178887

[20] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Dostupné z https://www.R-project.org/

[21] Shumway, R. H., Stoffer, D. (2017). *Time series analysis and its applications: with R examples* (4th ed). Springer.

[22] Turner, S. L., Karahalios, A., Forbes, A. B., Taljaard, M., Grimshaw, J. M., McKenzie, J. E. (2021). *Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series*. BMC medical research methodology, 21(1), 134. Dostupné z: https://doi.org/10.1186/s12874-021-01306-w

[23] Wagner, A. K., Soumerai, S. B., Zhang, F., Ross-Degnan, D. (2002). *Segmented regression analysis of interrupted time series studies in medication use research*. Journal of clinical pharmacy and therapeutics, 27(4), 299–309. Dostupné z: https://doi.org/10.1046/j.1365-2710.2002.00430.x.

[24] Zeileis, A., Leisch, F., Hornik, K., Kleiber, Ch. (2002). *strucchange: An R package for testing for structural change in linear regression models.* Journal of statistical software, 7: 1-38. Dostupné z: 10.18637/jss.v007.i02

[25] Zhang, X., Wu, K., Pan, Y., Yin, R., Zhang, Y., Kong, D., Wang, Q., Chen, W. (2023). *Optimized segmented regression models for the transition period of intervention effects*. Global health research and policy, 8(1), 29. Dostupné z: https://doi.org/10.1186/s41256-023-00312-3.

[26] *Child car seat safety in Japan*. Plaza homes. [online], [cit. 2024-02-06]. Dostupné z: https://www.realestate-tokyo.com/living-in-tokyo/driving/child-seat-safety/