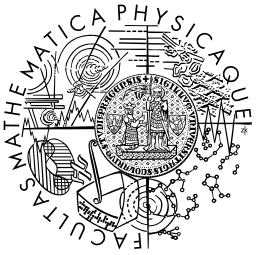


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Petra Pelikánová

Estimace dat s využitím intervalové analýzy

Katedra aplikované matematiky

Vedoucí bakalářské práce:	Mgr. Jaroslav Horáček
Studijní program:	Informatika
Studijní obor:	Obecná informatika

Praha 2017

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

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

V ..... dne ..... Podpis autora

Poděkování

Mé poděkování patří Mgr. Jaroslavovi Horáčkovi za odborné vedení, cenné rady, ochotu, trpělivost a veškerý věnovaný čas.

Název práce: Estimace dat s využitím intervalové analýzy

Autor: Petra Pelikánová

Katedra: Katedra aplikované matematiky

Vedoucí bakalářské práce: Mgr. Jaroslav Horáček, Katedra aplikované matematiky

Abstrakt: V práci popisujeme přístupy k modelování intervalových dat pomocí reálných a intervalových odhadů. Porovnáváme koncepty metody vnějšího modelu a vnitřního modelu, diskutujeme definice jejich slabých a silných variant. Vysvětlujeme postupy hledání lineární i nelineární estimace dat. Na reálných příkladech ilustrujeme toleranční metodu a odečítací přístup.

Klíčová slova: intervalová analýza, estimace dat, lineární regrese

Title: Estimating data with use of interval analysis

Author: Petra Pelikánová

Department: Department of applied mathematics

Supervisor: Mgr. Jaroslav Horáček, Department of applied mathematics

Abstract: This work is focused on estimating interval data by real functions and interval functions. It presents possibilistic and necessity models of interval regression and compares its strong and week formulations. Further we describe algorithms of linear and nonlinear estimation. The application part is based on demonstration of tolerance method and subtracting tolerance method analysing real cases.

Keywords: interval analysis, data estimation, linear regression

Obsah

Úvod

Cíle a struktura práce

Cílem práce je popsat známé přístupy k intervalové estimaci dat a následně je ilustrovat na reálných příkladech.

V úvodu je nastíněna motivace využití intervalové analýzy, která se zabývá počítáním s intervaly. Dále práce obsahuje část teoretickou a aplikační.

Teoretická část zahrnuje kapitolu 1 s definicemi intervalových pojmů využívaných v dalších kapitolách. Následně v kapitole 2 je připomenut běžný pojem lineární (reálné) regrese. Jsou zde představeny různé typy zpracovávaných dat (reálná až intervalová). Další kapitola 3 rozšiřuje reálnou estimaci na intervalovou. Popisuje a srovnává různé přístupy k modelování dat pomocí intervalové analýzy. Navazující kapitola 4 rozebírá různé postupy pro hledání řešení jednotlivých typů intervalové estimace. Poslední kapitola teoretické části (kap. 5) se pak zabývá intervalovou estimací nelineárních modelů.

Aplikační část v kapitole 6 ilustruje na reálných příkladech lineární a nelineární estimace toleranční přístup a odečítací metodu. Popisuje stručně implementaci a zvolený software.

Intervalová motivace

Běžně v životě potkáváme intervaly a vůbec se nad tím nepozastavujeme. Přirozeně porovnáváme věci a z toho nám vyplývá použití intervalů. Řekneme-li, že Barbora je vyšší než Adéla, ale menší než Cyril, vymezili jsme tím interval v němž se vyskytuje výška Barbory.

Geometrický význam reálného intervalu zavedl formálně už ve 4. století př. n. l. významný matematik Eukleides, když ve svém spisu definoval pojem úsečky. Na reálné ose můžeme uzavřený interval reprezentovat jako úsečku ohraničenou dvěma body, příkladem otevřeného intervalu pak může být celá přímka.

Další použití pojmu interval může být ve významu mezery, například časové. Opakující se událost, jako odjezd autobusu, může mít interval 15 minut. Jako vzdálenost mezi něčím je používán interval i ve významu hudebních intervalů.

Nastínili jsme tedy dva případy běžného použití a to buď ve významu že vymezená oblast něco obsahuje (výšku osoby), nebo naopak neobsahuje (příjezd autobusu). V této práci se budeme zabývat především prvním zmíněným případem.

Použití intervalů nám umožňuje vymezit prostor, ve kterém se s jistotou nachází důležitá konstanta π. Přesnost s jakou ji umíme určit se v průběhu času mění a zvětšuje. Od Archimédova odhadu z doby asi 200 let př. n. l. se interval, kterým dokážeme ohraničit toto číslo, zpřesňoval, až v roce 1973 přesáhl přesnost milion cifer. Tento ilustrativní příklad nám ukazuje, že u některých čísel neznáme jejich konkrétní hodnotu. Jiné konstanty ani nemají jedinou správnou hodnotu a přesto s nimi chceme počítat. Například tíhové zrychlení se liší na různých místech planety. Pokud bychom chtěli zahrnout ve výpočtu všechny možné hodnoty dané konstanty, můžeme je uzavřít do intervalu a počítat s celým intervalem.

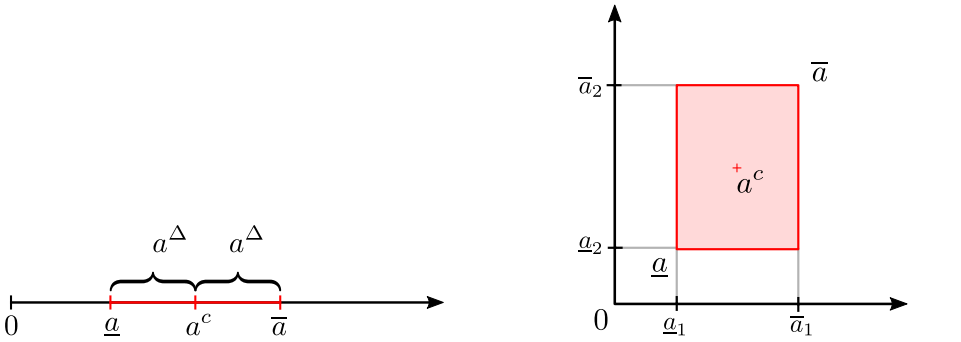
Dalším důvodem, který nás dovede k použití intervalové analýzy je nepřesnost. Jak bylo zmíněno, některé konstanty neumíme přesně reprezentovat, protože mají nekonečný desetinný rozvoj. Jiná čísla neznáme přesně z důvodu nepřesnosti měřících zařízení, ze kterých je získáváme. V obou těchto případech často můžeme určit interval, v němž se s jistotou nachází pro nás zajímavá hodnota.

Získaná nepřesná data potřebujeme umět uchopit a nějakým způsobem popsat. Ke zkoumání závislosti mezi proměnnými experimentu můžeme využít intervalové estimace. Pomocí regresních modelů lze popsat chování dat. Jestliže známe model dat, umožní nám to odhadovat, jaký může být vývoj dat pro hodnoty, které neznáme z měření.

1. Základy intervalového počítání

Tato kapitola obsahuje základní používané pojmy a značení. Většina základních definic je převzata z knihy Moore a kol. (2009).

Definice 1.1 (Reálný interval)**.** Nechť $\underline{a}, \overline{a} \in \mathbb{R} \cup \{-\infty, \infty\}$, $\underline{a} \leq \overline{a}$. *Reálný interval* je množina $\boldsymbol{a} := [\underline{a}, \overline{a}] = \{a \mid \underline{a} \leq a \leq \overline{a}\}$. Množinu všech reálných intervalů značíme $\mathbb{IR}$.
Střed intervalu $\boldsymbol{a}$ je reálné číslo $a^c := \frac{\overline{a}+\underline{a}}{2}$.
Poloměr intervalu je kladné reálné číslo $a^\Delta := \frac{\overline{a}-\underline{a}}{2}$.
Šířka intervalu je pak dvojnásobek poloměru.



(a) Reálný interval z $\mathbb{IR}$ (b) Dvourozměrný reálný interval

Obrázek 1.1: Geometrický význam reálného intervalu

Jinou alternativou definice intervalu je vyjádření pomocí středu a poloměru intervalu $\boldsymbol{a} = [a^c - a^\Delta, a^c + a^\Delta] = \{a : |a - a^c| \leq a^\Delta\}$, které nám dává lepší geometrickou představu (obr.1.1a pro $\mathbb{IR}$, pro dvourozměrný prostor $\mathbb{IR}^2$ pak 1.1b). Interval je množina bodů vzdálených od středu maximálně o poloměr intervalu.

Speciálním případem intervalu s nulovým poloměrem je *degenerovaný interval*, což je jednoprvková množina pro kterou platí $\underline{a} = \overline{a}$.

Definice 1.2 (Intervalová matice)**.** Nechť $\underline{A}, \overline{A} \in \mathbb{R}^{m \times n}$ jsou reálné matice a platí (po složkách) nerovnost $\underline{A} \leq \overline{A}$. *Intervalovou matici* definujeme

$$\boldsymbol{A} := \{A \mid \underline{A} \leq A \leq \overline{A}\}.$$

Protože intervalová matice je matice, jejíž prvky jsou reálné intervaly, je možné ji definovat také pomocí středové matice A^c a matice poloměrů A^Δ. Intervalové matice a vektory budou v textu vždy značeny tučně. Pro reálná čísla (v anglické literatuře též crisp) sdružená do vektorů či matic pak používáme běžný (netučný) řez písma.

Abychom mohli s intervaly pohodlně pracovat, definujme si nyní základní intervalové aritmetické operace. Umožní nám to zacházet s výrazy, v nichž proměnné nejsou pouze reálná čísla, ale mohou to být celé intervaly.

Při běžném dosazení proměnné $x \in \mathbb{R}$ do výrazu získáme vyhodnocením jeden výsledek. Pokud reálnou proměnnou vyměníme za interval $\boldsymbol{x} \in \mathbb{IR}$, můžeme si představit, že dosadíme do výrazu každý bod intervalu samostatně. Všechny získané výsledky jsou pak vyhodnocením výrazu pro celý interval. Zajímají nás tedy jako výsledek všechny možné realizace pro libovolné $x \in \boldsymbol{x}$.

Definice 1.3 (Aritmetické operace)**.** Pro reálné intervaly $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{IR}$ definujeme *výsledek aritmetické operace* $\circ$ následujícím způsobem

$$\boldsymbol{a} \circ \boldsymbol{b} := \{a \circ b \mid a \in \boldsymbol{a},\, b \in \boldsymbol{b}\}.$$

Navíc pro dělení budeme předpokládat $\{0\} \notin \boldsymbol{b}$[1]. Z definice plyne pro základní aritmetické operace

$$\begin{aligned}
&\boldsymbol{a} + \boldsymbol{b} := [\underline{a} + \underline{b},\, \overline{a} + \overline{b}] \\
&\boldsymbol{a} - \boldsymbol{b} := [\underline{a} - \overline{b},\, \overline{a} - \underline{b}] \\
&\boldsymbol{a} \cdot \boldsymbol{b} := [\min(M), \max(M)],\, M = \{\underline{a}\,\underline{b},\, \underline{a}\,\overline{b},\, \overline{a}\,\overline{b},\, \overline{a}\,\underline{b},\, \underline{a}\,\underline{b}\} \\
&\boldsymbol{a} \,/\, \boldsymbol{b} := [\min(M), \max(M)],\, M = \{\underline{a}/\underline{b},\, \underline{a}/\overline{b},\, \overline{a}/\overline{b},\, \overline{a}/\underline{b},\, \underline{a}/\underline{b}\}, \text{kde } \{0\} \notin \boldsymbol{b}.
\end{aligned}$$

Definice 1.4 (Intervalová soustava)**.** Nechť $\boldsymbol{A} \in \mathbb{IR}^{m \times n}$, $\boldsymbol{b} \in \mathbb{IR}^m$. *Intervalovou soustavu lineárních rovnic* $\boldsymbol{A}x = \boldsymbol{b}$ definujeme

$$\{Ax = b \mid A \in \boldsymbol{A},\, b \in \boldsymbol{b}\}.$$

Řešení x soustavy $Ax = b$ pro nějaká $A \in \boldsymbol{A}$, $b \in \boldsymbol{b}$ je řešením konkrétní realizace lineární soustavy. *Množinu řešení intervalové soustavy* definujeme jako

$$\Sigma := \{x \in \mathbb{R}^n \mid \exists A \in \boldsymbol{A},\, \exists b \in \boldsymbol{b} \text{ taková že } Ax = b\}.$$

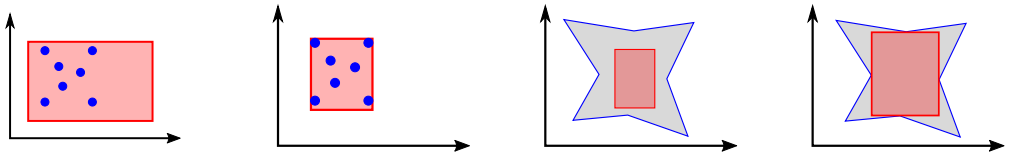
Definice 1.5 (Intervalová obálka)**.** Nechť $S \subset \mathbb{R}^n$ je množina. *(Vnější) intervalová obálka* (anglicky interval enclosure) množiny S je libovolný intervalový vektor $\boldsymbol{v} \in \mathbb{IR}^n$ takový, že obsahuje množinu S, tedy

$$S \subseteq \boldsymbol{v}.$$

Definice 1.6 (Intervalový obal)**.** Nechť $S \subset \mathbb{R}^n$ je množina. *(Vnější) intervalový obal* (anglicky interval hull) $\square S$ množiny S je nejtěsnější možný intervalový vektor obsahující S:

$$\square S := \bigcap_{\substack{\boldsymbol{v} \in \mathbb{IR}^n \\ S \subseteq \boldsymbol{v}}} \boldsymbol{v}.$$

Kromě vnější intervalové obálky a obalu můžeme definovat také vnitřní intervalovou obálku a obal (viz obr. 1.2).



(a) Vnější intervalová obálka

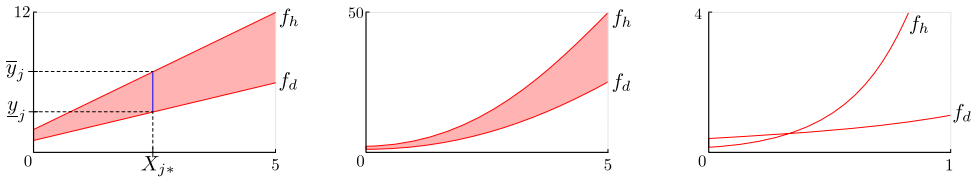
(b) Vnější intervalový obal

(c) Vnitřní intervalová obálka

(d) Vnitřní intervalový obal

Obrázek 1.2: Vnější a vnitřní obálka a obal (modře je znázorněna množina, červeně pak její obálka, resp. obal)

[1] Dělení lze definovat i bez tohoto požadavku.



(a) $x \in [0{,}5]$
$f_h = 2 + 2 \cdot x$
$f_d = 1 + 1 \cdot x$
$\boldsymbol{y} = \boldsymbol{a}_0 + \boldsymbol{a}_1 x$
$\boldsymbol{a} = ([1{,}2], [1{,}2])$

(b) $x \in [0{,}5]$
$f_h = 2 + 2 \cdot x^2$
$f_d = 1 + 2 \cdot x^2$
$\boldsymbol{y} = \boldsymbol{a}_0 + \boldsymbol{a}_1 x^2$
$\boldsymbol{a} = ([1{,}2],[1{,}2])$

(c) $x \in [0{,}1]$
$f_h = e^{4 \cdot (x - 0.5)}$
$f_d = e^{1 \cdot (x-1)}$
$\boldsymbol{y} = e^{\boldsymbol{a}_0 (x + \boldsymbol{a}_1)}$
$\boldsymbol{a} = ([1{,}4], [-1, -0.5])$

Obrázek 1.3: Intervalový pás

Definice 1.7 (Intervalový pás)**.** Nechť máme dvě reálné funkce $f_d, f_h : M \to \mathbb{R}$, definované pro všechna $x = (x_1, \ldots, x_n)$ z $M \subseteq \mathbb{R}^n$, pro které platí

- f_d, f_h jsou spojité na množině M,
- $\forall x \in M : f_d(x) \leq f_h(x)$.

Pak *intervalovým pásem* $P(x)$ nazýváme množinu bodů

$$P(x) := \{[x_1, \ldots, x_n, y] \in \mathbb{R}^{n+1} |\ (x_1, \ldots, x_n) \in M : f_d(x) \leq y \leq f_h(x)\}.$$

Funkci f_d nazýváme *dolní hranicí* pásu $P(x)$ a funkci f_h *horní hranicí*.

Intervalovým pásem tedy rozumíme plochu vymezenou hraničními funkcemi. První předpoklad definice, že hraniční funkce jsou spojité, nám zaručuje to, že intervalový pás je spojitá oblast. Druhý předpoklad pak neformálně řečeno vynucuje, že hraniční funkce se nekříží, takže dolní hranice je vždy dole a horní hranice vždy nahoře.

Budou-li hraniční funkce stejného tvaru, například obě to budou přímky, budeme hovořit o vymezeném pásu jako o *(parametrické) intervalové křivce*.

Na obrázku 1.3a vidíme příklad intervalové přímky. Její parametrický zápis je $\boldsymbol{y} = \boldsymbol{a}_0 + \boldsymbol{a}_1 x$. Příslušný intervalový pás zahrnuje všechny přímky, jejichž parametry a_0, a_1 jsou z intervalového vektoru parametrů $\boldsymbol{a} = ([1, 2], [1, 2])$. Příslušné hraniční funkce f_d, f_h odpovídají přímkám $f_d(x) = 1 + 1 \cdot x$, $f_h(x) = 2 + 2 \cdot x$. Díky spojitosti funkcí na intervalu $M = [0, 5]$ se každé $x \in M$ zobrazí na spojitý interval $\boldsymbol{y} = [\underline{y}, \overline{y}]$.

V druhé části obrázku 1.3b vidíme kvadratickou křivku. Poslední část obrázku 1.3c znázorňuje exponenciální funkci na definičním oboru [0,1], která s parametrickým vektorem $\boldsymbol{a} = ([1{,}4], [-1, -0.5])$ nedefinuje korektně intervalový pás. Porušuje totiž podmínku „křížení funkcí“, $f_d(x) \leq f_h(x)$ neplatí pro $x < 1/3$.

Stručně shrnuto můžeme intervalový pás v některých případech popsat parametrickou funkcí a hraniční funkce pásu získat dosazením krajních mezí parametrického vektoru. Při některých volbách funkcí a k nim příslušným parametrům musíme ale dávat pozor, abychom neporušili některé požadavky definice[2].

[2]To vychází z komplikací při vyhodnocování intervalových funkcí (Moore a kol., 2009).

2. Lineární (reálná) regrese

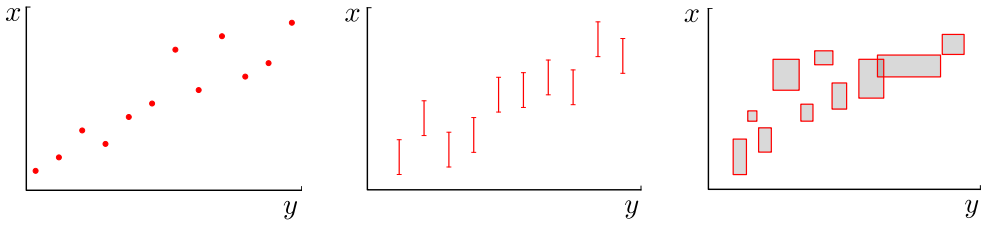
Cílem této kapitoly je připomenout pojem lineární regrese. Reálná regrese je zde chápána jako prokládání dat reálnou funkcí. Následující kapitola pak představuje její rozšíření na intervalovou lineární regresi pro reálná i intervalová data.

Definice 2.1 (Vstupní data)**.** Nechť $\boldsymbol{X} \in \mathbb{IR}^{p\times n}$ je matice, $\boldsymbol{y} \in \mathbb{IR}^{p}$ je vektor. Matici $\boldsymbol{X}$ a vektor $\boldsymbol{y}$ nazýváme souhrnně *vstupními daty*, značíme $(\boldsymbol{X}, \boldsymbol{y})$. Počet p řádků matice $\boldsymbol{X}$ je *počet pozorování*.

V následujících kapitolách budeme rozlišovat vstupní data na tři typy (tabulka 2.1, obrázek 2.1) . Nejjednodušší typ jsou *reálná vstupní data*, kdy všechny prvky matice X i vektoru y jsou pouze reálná čísla. Dále budeme přístupy k estimaci dat zobecňovat pro *reálně-intervalová* data, $X \in \mathbb{R}^{p\times n}$, $\boldsymbol{y} \in \mathbb{IR}^{p}$. Nejobecnější případ dat jsou *intervalová data*, $\boldsymbol{X} \in \mathbb{IR}^{p\times n}, \boldsymbol{y} \in \mathbb{IR}^{p}$, kde vstupní matice i vektor obsahují intervaly. (Intervalově-reálná data nepotřebujeme samostatně rozlišovat a můžeme je chápat jako speciální případ dat intervalových.)

$\boldsymbol{X}$	$\boldsymbol{y}$	Označení	Zkratka
$\mathbb{R}^{p\times n}$	$\mathbb{R}^{p}$	reálná data	$\mathbb{R}$-$\mathbb{R}$
$\mathbb{R}^{p\times n}$	$\mathbb{IR}^{p}$	reálně-intervalová data	$\mathbb{R}$-$\mathbb{I}$
$\mathbb{IR}^{p\times n}$	$\mathbb{IR}^{p}$	intervalová data	$\mathbb{I}$-$\mathbb{I}$

Tabulka 2.1: Typy vstupních dat



(a) $\mathbb{R}$-$\mathbb{R}$ (b) $\mathbb{R}$-$\mathbb{I}$ (c) $\mathbb{I}$-$\mathbb{I}$

Obrázek 2.1: Typy vstupních dat

2.1 Reálná data

Lineární regresí pomocí metody nejmenších čtverců rozumíme (pro dvourozměrná data) aproximaci vstupních dat přímkou. Pro reálná vstupní data (X, y), kde $X \in \mathbb{R}^{p\times n+1}, y \in \mathbb{R}^{p}$ tedy chceme najít nadrovinu $y = a_0 + x_1a_1 + \cdots + x_na_n$, která co nejlépe vystihuje trend vstupních dat. Naším cílem je nalézt přesné nebo alespoň přibližné řešení soustavy $y = Xa$, kde X je vstupní matice (nezávislých proměnných), y je vstupní vektor (závislé proměnné) a $a = (a_0, a_1, \ldots, a_n)$ je

neznámý vektor parametrů lineární regrese. Získáváme tak soustavu

$$\begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{p1} & \dots & x_{pn} \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix}. \tag{2.1}$$

Tato soustava často nemusí mít řešení, a proto hledáme její přibližné řešení. Metoda nejmenších čtverců [1] proto řeší soustavu

$$X^\top y = X^\top X a,$$

která má řešení vždy a navíc nalezené řešení $a = (X^\top X)^{-1} X^\top y$ minimalizuje chybu nalezené aproximující nadroviny ve smyslu euklidovské vzdálenosti levé a pravé strany původní rovnice (2.1).

Definice 2.2 (Lineární reálný regresní model)**.** Nechť (X,y), $X \in \mathbb{R}^{p\times n}$, $y \in \mathbb{R}^p$ jsou vstupní data. *Lineární model* pro daná vstupní data je $y(x) = x^\top a$, kde $a = (X^\top X)^{-1} X^\top y$.

2.2 Reálně-intervalová data

Zadaná data nezávislé proměnné jsou reálná čísla, $X \in \mathbb{R}^{p\times n}$, ale závislou proměnnou $\boldsymbol{y} \in \mathbb{IR}^p$ máme zadanou pomocí intervalů. Důvodem může být, že přesné hodnoty $\boldsymbol{y}$ nejsme schopni měřit, kvůli nepřesnostem měřícího zařízení.

Pro zjednodušenou představu budeme nyní uvažovat dvourozměrná data, tedy $n = 2$.

Stále chceme nalézt přímku, která aproximuje vstupní data. Pro reálná data jsme měli pro každý bod měření v soustavě $y = Xa$ jednu rovnici $y_j = a_1 + x_j a_2$, a hledali jsme pro danou soustavu parametr a, který minimalizovalo odchylky od přesného řešení každé rovnice.

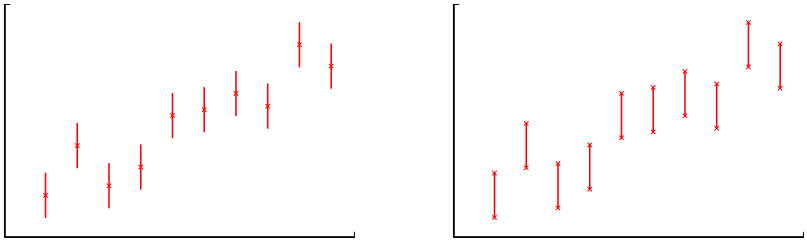
Analogicky intervalová soustava $\boldsymbol{y} = Xa$ v každém řádku reprezentuje intervalový výraz $\boldsymbol{y}_j = a_1 + x_j a_2$ odpovídající jednomu měření. Pro jednotlivé realizace $y_j \in \boldsymbol{y}_j$ můžeme nyní získat různá a_j splňující rovnici. Rozšířením levé strany na intervalový vektor jsme získali nekonečně mnoho reálných soustav $y = Xa$, $y \in \boldsymbol{y}$ s neznámou a.

Množina řešení intervalové soustavy $\boldsymbol{y} = Xa$ je množinou všech parametrů a odpovídajících libovolné realizaci soustavy $y = Xy$, $y \in \boldsymbol{y}$. Použitím metody nejmenších čtverců pro každou soustavu bychom dostali množinu přímek.

Co když ale stále chceme jednoduchou interpretaci dat pomocí jedné reálné přímky? V praxi může reálná funkce výrazně usnadnit pozorovateli orientaci ve vstupních datech. Dalším důvodem, proč hledat tuto aproximaci dat, je použití reálné funkce jako počátečního řešení pro intervalové algoritmy estimace dat.

Přímočarým řešením může být redukování problému na případ s pouze reálnými daty. Každý interval bychom mohli reprezentovat pouze jedním bodem, protože sice pro každé j máme dán interval $\boldsymbol{y}_j$ s nekonečně mnoha body, avšak předpokládáme, že reálná data jsou pro každé j pouze jeden bod uvnitř daného intervalu.

[1] Anděl (2007)



(a) Středy intervalů (b) Kraje intervalů

Obrázek 2.2: Prokládání dat středy intervalů nebo kraji

Logickou volbou reprezentanta intervalu může být jeho střed, pak můžeme nalézt řešení reálného problému se vstupními daty (X, y^c). Tato varianta však vůbec nepracuje se šířkou intervalů. Jako lepší volba by se tedy mohla zdát reprezentace dat kraji intervalů a řešit problém $(\tilde{X}, (\underline{y}, \overline{y}))$, čímž se nám zdvojnásobila délka vstupního vektoru. Matice $\tilde{X} \in \mathbb{R}^{2p \times n}$ značí $\tilde{X} = \begin{pmatrix} X \\ X \end{pmatrix}$. Následující věta shrnuje náš výsledek, že oba tyto přístupy dávají pro reálnou regresi (ve smyslu metody nejmenších čtverců) stejný odhad. Stačí tedy uvažovat jako reprezentanta intervalu opravdu pouze jeho střed.

Věta 2.1. Nechť $(X,\boldsymbol{y})$ jsou vstupní data, $X \in \mathbb{R}^{p \times n}, \boldsymbol{y} \in \mathbb{IR}^p$. Řešení lineární regrese ve smyslu metody nejmenších čtverců aplikované na kraje intervalů je stejné s regresí na středech intervalů.

Důkaz:

Podle metody nejmenších čtverců hledáme řešení intervalové soustavy

$$X^\top \boldsymbol{y} = X^\top X a, \tag{2.2}$$

kde hledaný vektor parametrů $a \in \mathbb{R}^n$ definuje regresní přímku pro daná vstupní data $(X, \boldsymbol{y})$. Sestavme pro jednotlivé případy regrese (na středech intervalů, resp. na krajích) odpovídající soustavy a vyjádřeme z nich neznámý vektor a.

Středy intervalů

Vyjádřením vektoru a z rovnice 2.2 definující metodu nejmenších čtverců získáme

$$a = (X^\top X)^{-1} X^\top y^c$$

$$\begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = \left(\begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} x_{11} & \dots & x_{1n} \\ \vdots & & \vdots \\ x_{p1} & \dots & x_{pn} \end{pmatrix} \right)^{-1} \cdot \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} y_1^c \\ \vdots \\ y_p^c \end{pmatrix}$$

Označme $Q := X^\top X$, pak prvky matice $Q \in \mathbb{R}^{n \times n}$ jsou

$$q_{ij} = \sum_{k=1}^{p} x_{ik} \cdot x_{kj}.$$

Dále označme $r := X^\top y^c$, vektor $r \in \mathbb{R}^n$ má prvky

$$r_i = \sum_{k=1}^{p} x_{ik} \cdot y_k^c.$$

Můžeme nyní hledaný vektor a vyjádřit jako

$$a = Q^{-1}r.$$

Kraje intervalů

Pro krajní body intervalů $\boldsymbol{y}_j$ můžeme opět vyjádřit vektor parametrů

$$\tilde{a} = (\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top \tilde{y}) \tag{2.3}$$

Nyní máme však dvojnásobný počet rovnic, které určují přímku prokládající vstupní data. Proto rozepsáním soustavy 2.3 získáváme následující rovnici

$$\tilde{a} = \left(\begin{pmatrix} X^\top & X^\top \end{pmatrix} \cdot \begin{pmatrix} X \\ X \end{pmatrix}\right)^{-1} \cdot \begin{pmatrix} X^\top & X^\top \end{pmatrix} \cdot \begin{pmatrix} \overline{y} \\ \underline{y} \end{pmatrix}$$

Analogicky jako v předchozím případě označme $\tilde{Q} := \tilde{X}^\top \tilde{X}$, $\tilde{r} = \tilde{X}^\top \tilde{y}$. Vektor parametrů určujících regresní nadrovinu pak lze zapsat jako součin matice a vektoru

$$\tilde{a} = \tilde{Q}^{-1}\tilde{r}.$$

Vyjádřením prvků matice $\tilde{Q}$ a prvků vektoru $\tilde{r}$ a ekvivalentními úpravami získáme

$$\tilde{q}_{ij} = \sum_{k=1}^{p} x_{ik}x_{kj} + \sum_{k=1}^{p} x_{ik}x_{kj} = 2\sum_{k=1}^{p} x_{ik}x_{kj} = 2 \cdot q_{ij}$$

$$\tilde{r}_i = \sum_{k=1}^{p} x_{ik} \cdot \overline{y}_k + \sum_{k=1}^{p} x_{ik} \cdot \underline{y}_k = \sum_{k=1}^{p}(x_{ik} \cdot \overline{y}_k + x_{ik} \cdot \underline{y}_k) = \sum_{k=1}^{p} x_{ik} \cdot (\overline{y}_k + \underline{y}_k).$$

Ze vzorce pro výpočet středu intervalu $y^c = \frac{\overline{y}+\underline{y}}{2}$ můžeme vyjádřit $2y^c = \overline{y} + \underline{y}$. Po dosazení do výrazu $\tilde{r}_i$ získáváme

$$\tilde{r}_i = \sum_{k=1}^{p} x_{ik} \cdot 2y_k^c = 2\sum_{k=1}^{p} x_{ik} \cdot y_k^c = 2 \cdot r_i.$$

Vrátíme-li se k maticovému zápisu vektoru $\tilde{a}$ máme

$$\tilde{a} = (2Q)^{-1} \cdot 2r = \frac{1}{2}Q^{-1} \cdot 2r = Q^{-1}r,$$

$$\tilde{a} = a.$$

Oba způsoby počítání přímky aproximující vstupní data (prokládání středů intervalů nebo krajů intervalů) definují stejnou křivku, čímž jsme dokončili důkaz věty.

□

Získání estimátoru pomocí odhadu chování středů (resp. krajů) intervalů nám dává aproximaci dat, která dobře vyhovuje podmínce centrality, tj. dobře popisuje středy intervalů zadaných dat. Toho lze využít i při intervalové estimaci dat, kdy hledáme intervalový pás popisující chování dat. Ovšem tato výhoda je zřetelná především u dat, kde všechny intervaly $\boldsymbol{y}_j$ mají stejnou šířku. Tento předpoklad je v praxi často splněn, protože nepřesnost pozorované veličiny y může být definována nepřesností měřícího přístroje.

2.3 Intervalová data

Nejobecnějším případem vstupních dat jsou intervalová data $(\boldsymbol{X}, \boldsymbol{y})$, kde $\boldsymbol{X} \in \mathbb{IR}^{p\times n}$, $\boldsymbol{y} \in \mathbb{IR}^p$. Taková vstupní data si lze představit jako n-rozměrné boxy (obr. 2.1). Trend jejich chování zjistíme nalezením přibližného řešení intervalové soustavy

$$\boldsymbol{y} = \boldsymbol{X}a,$$

k tomu můžeme využít intervalové formulace metody nejmenších čtverců. Ta nám ale opět dává více než jednu reálnou přímku popisující data.

Při hledání reálné funkce aproximující vstupní data jsme při přechodu od reálných dat k reálně-intervalovým datům pozorovali problém s nejednoznačností soustavy a problém jsme redukovali na řešení reálné soustavy lineárních rovnic. V tomto případě můžeme také reprezentovat intervalové boxy jejich středem. Pro nalezení reálné funkce popisující trend chování dat pak budeme řešit soustavu

$$(X^c)^\top y^c = (X^c)^\top X^c a.$$

Jako reálnou přímku prokládající vstupní data pak vezmeme $y(x) = x^\top a$, kde

$$a = \left((X^c)^\top X^c\right)^{-1} (X^c)^\top y^c.$$

Tímto jsme shrnuli prokládání různých typů dat (od reálných až po intervalová) nadrovinou v prostoru $\mathbb{R}^n$. Dále rozebereme způsoby popisu dat pomocí intervalového pásu. Rozšíříme reálnou regresi na intervalovou a přestaneme tedy vynucovat, že parametrický vektor a je pouze reálný.

3. Lineární intervalová estimace

V předchozí kapitole jsme připomněli pojem lineární (reálné) regrese, kdy modelujeme data pomocí reálné funkce. Tato kapitola představí obecnější koncept intervalové estimace dat. Potřeba hledat nové přístupy k estimaci vznikla s nutností zpracování takzvaných fuzzy dat[1] a snahou vypořádat se s nepřesností dat.

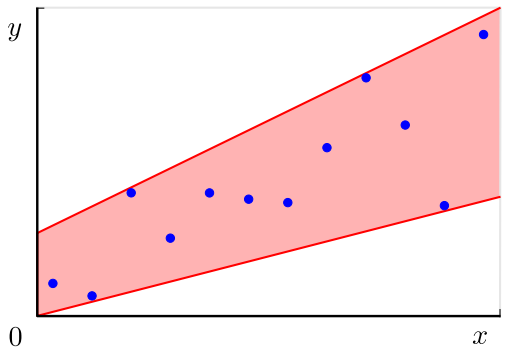
Model reálné lineární regrese $y(x) = a_0 + a_1x_1 + \cdots + a_nx_n$ nám umožňoval popsat data pomocí nadroviny (v dvourozměrném prostoru přímky). Nadrovinu jednoznačně určoval parametrický vektor a. Pokud povolíme, aby parametry modelu nebyla pouze reálná čísla, ale můžou to být intervaly, získáme intervalovou lineární regresi. Výraz $X\boldsymbol{a}$ pak nereprezentuje pouze jednu nadrovinu, ale celou množinu nadrovin. Pokud tyto nadroviny zapouzdříme intervalovým pásem, získáváme pás, který s jistotou obsahuje všechna vstupní data (obr. 3.1).

Zaveďme pojem intervalové lineární regrese v obecném znění, jak je uveden v článku od Hladík a Černý (2012a).

Modelem intervalové lineární regrese budeme rozumět

$$\boldsymbol{y}(x) = \boldsymbol{a}_0 + \boldsymbol{a}_1\boldsymbol{x}_1 + \cdots + \boldsymbol{a}_n\boldsymbol{x}_n. \tag{3.1}$$

Problémem lineární regrese stále zůstává nalézt vektor parametrů $\boldsymbol{a}$, což je nyní intervalový vektor, který určuje hraniční nadroviny intervalového pásu. Vstupní data tedy prokládáme intervalovým pásem, jehož hranice jsou u lineární regrese dvě nadroviny určené dolní a horní hranicí vektoru $\boldsymbol{a}$.



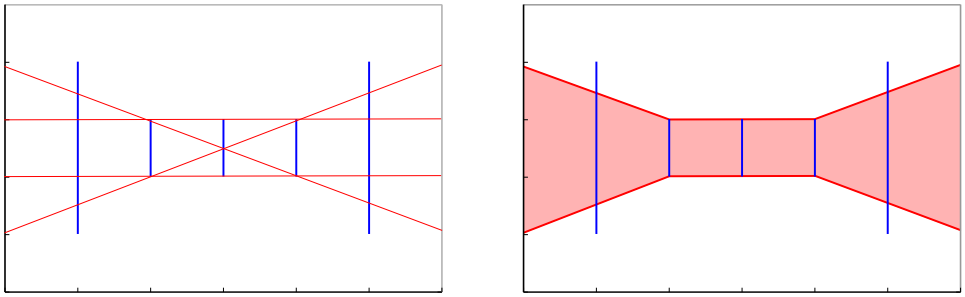
Obrázek 3.1: Intervalová lineární regrese zapouzdřující všechna vstupní data.

Zobecnění regresního modelu nám dává možnost různých přístupů k intervalové estimaci. Pro zpracování reálných dat dává nejlepší smysl hledat tzv. vnější model, tedy intervalový pás obsahující vstupní data. Pokud však zpracováváme data intervalová, lze hledat i pásy definované jinak. Nyní se podíváme na řešení intervalové soustavy definované vstupními daty, vysvětlíme souvislost s intervalovou metodou nejmenších čtverců a následně s vnější a vnitřní estimací.

3.1 Řešení soustavy

Pokud získáme z experimentu hodnoty proměnných $\boldsymbol{X'} \in \mathbb{IR}^{(n-1)\times p}$ a závislé proměnné $\boldsymbol{y} \in \mathbb{IR}^p$, chceme v prostoru $\mathbb{R}^n$ najít podprostor, který by dobře po-

[1]Fuzzy data nemají ostře definovanou vlastnost (ne)náležení do množiny. Neexistují zde pouze stavy patřit do množiny (obvykle ztotožněno s hodnotou 1) a nepatřit do množiny (obvykle ztotožněno s 0), ale stav náležení může nabývat i mezihodnot. Tedy čím víc prvek do množiny patří, tím více se blíží stav náležení hodnotě 1 a naopak, pokud prvek do množiny skoro nepaří, jde hodnota k 0.



(a) Extrémní řešení intervalové soustavy

(b) Oblast v níž leží řešení intervalové soustavy

Obrázek 3.2: Řešení intervalové soustavy se vstupními daty typu $\mathbb{R}$-$\mathbb{I}$

pisoval chování těchto dat. Pokud vztah mezi proměnnými je lineární, nabízí se hledat nadrovinu (pro data $\boldsymbol{X'} \in \mathbb{IR}^{1\times p}$ konkrétně přímku v $\mathbb{R}^2$). To odpovídá řešení soustavy sestavené z matice $\boldsymbol{X'}$ rozšířené o sloupec jedniček a z vektoru $\boldsymbol{y}$.

Rozšíříme-li matici $\boldsymbol{X'}$ o sloupec jedniček na matici $\boldsymbol{X} \in \mathbb{IR}^{n\times p}$, množina řešení soustavy $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a}$ definuje množinu nadrovin takovou, že pro každou hodnotu parametrů $a \in \boldsymbol{a}$ je jednoznačně určena nadrovina protínající všechna vstupní data. Pro dvourozměrný prostor tedy každá přímka definovaná řešením soustavy protíná všechny intervalové boxy definované j-tým pozorováním, tj. vektorem $(\boldsymbol{x}_j, \boldsymbol{y}_j)$. Samotná množina řešení definuje všechny směrnice těchto přímek.

Definice 3.1. Nechť máme vstupní data $(\boldsymbol{X}, \boldsymbol{y})$, $\boldsymbol{X} \in \mathbb{IR}^{n\times p}, \boldsymbol{y} \in \mathbb{IR}^p$. Máme systém intervalových lineárních rovnic $\mathbf{y} = \mathbf{X}a$, pro který definujeme množinu řešení

$$\Sigma = \{a \in \mathbb{R}^n \mid (\exists X \in \mathbf{X})(\exists y \in \mathbf{y})\ y = Xa\}$$

Na obrázku 3.2a můžeme vidět extrémní přímky, které jsou definované nějakým parametrickým vektorem $a \in \Sigma$. Tyto přímky vymezují oblast všech přímek odpovídajících nějakému možnému řešení intervalové soustavy (na obr. 3.2b).

Nevýhodou množiny řešení je, že množina může být prázdná, tedy nemusí existovat žádná nadrovina procházející data, což znamená, že soustava je neřešitelná[2]. Další nevýhodou je, že má nekonečně mnoho řešení, která nemusí být dobře uchopitelná a není tedy jednoduché je znázornit.

Estimace intervalových dat se většinou snaží znázornit graficky ne řešení soustavy, tedy vektoru $\boldsymbol{a}$ v tomto konkrétním případě, ale znázornit jemu odpovídající nadrovinu. Samotné řešení soustavy lze uzavřít do obálky či obalu, tedy zapouzdřit do intervalového boxu[3]. My budeme zkoumat, jak lze zapouzdřit námi hledané nadroviny pomocí nějakých hraničních funkcí a tím získat intervalový pás popisující vstupní data.

První zmíněný problém, neexistence řešení soustavy, můžeme řešit pomocí regresní metody nejmenších čtverců, jak bude ukázáno v následující podkapitole. Problém se složitostí popisu nadrovin pak budou řešit vnější a vnitřní modely, které budou definovány dále.

[2]Takovým příkladem mohou být přeurčené systémy, které mají více rovnic než proměnných, viz Horáček (2011).

[3]Toto téma je mimo rámec práce, je možné nalézt řadu publikací zabývajících se problémem efektivního hledání obálky řešení soustav. O zapouzdření množiny řešení mluví i Černý a kol. (2013).

3.2 Metoda nejmenších čtverců

Výše zmíněná množina Σ nám dává neprázdný výsledek jen v případě, že soustava $\mathbf{y} = \mathbf{X}a$ má nějaké řešení. Pokud však soustava žádné řešení nemá, často nás zajímá alespoň přibližné. V případě reálných dat je velmi používanou aproximací metoda nejmenších čtverců (anglicky Ordinary Least Squares), kterou lze rozšířit i pro intervalové soustavy (Černý a kol., 2013).

Definice 3.2. *OLS-množina* pro intervalovou soustavu $\mathbf{y} = \mathbf{Xa}$ je definována jako

$$\mathrm{OLS}(\mathbf{X}, \mathbf{y}) = \left\{ a \in \mathbb{R}^p \mid (\exists X \in \mathbf{X})(\exists y \in \mathbf{y})\ X^\top X a = X^\top y \right\}.$$

Vezmeme-li libovolnou možnou realizaci bodů ze vstupních dat (tedy z každého boxu odpovídajícího jednomu pozorování jeden bod), kterou proložíme metodou nejmenších čtverců, získáme nadrovinu, které odpovídá nějaký vektor parametrů $a \in \mathrm{OLS}(\boldsymbol{X}, \boldsymbol{y})$.

Je zřejmé, že OLS množina parametrů přímek obsahuje všechna řešení původní soustavy $\boldsymbol{y} = \boldsymbol{X}a$. Tedy platí $\Sigma \subseteq \mathrm{OLS}(\boldsymbol{X}, \boldsymbol{y})$. Přímky určené parametry z OLS jsou tedy nadmnožinou přímek určených řešením soustavy. Díky tomu, že soustava rovnic $X^\top X a = X^\top y$ má vždy řešení, je OLS množina vždy neprázdná.

Zmiňme, že při hledání aproximace OLS množiny, lze využít přepis soustavy

$$\boldsymbol{X}^\top \boldsymbol{X} a = \boldsymbol{X}^\top \boldsymbol{y},$$

který je vhodnější pro hledání intervalové obálky (viz Neumaier (1986), vektor b je zde přidaný vektor proměnných, jehož hodnoty se nevyužívají)

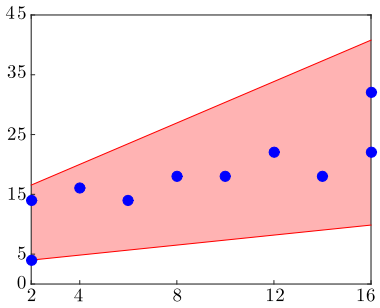
$$\begin{pmatrix} 0 & X^\top \\ X & I_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ \boldsymbol{y} \end{pmatrix}.$$

Přestože lze najít obal OLS množiny, stále přetrvává problém se složitostí popisu nadrovin odpovídajících parametrům z této množiny. Tvar oblasti, v níž se nachází všechny nadroviny definované OLS množinou, může být ještě o něco komplikovanější než oblast odpovídající množině řešení Σ.

Aby se nám pracovalo s řešením lépe, definujeme nyní vnější model intervalové regrese. Znamená to, že získáme jednodušší popis dat pomocí intervalového pásu. Důsledkem ale bude, že se oddálíme od původního statistického významu regrese jakožto aproximace dat. Budeme totiž vyžadovat verifikovanost nalezeného řešení, tedy stoprocentní správnost. To přibližuje intervalovou estimaci více k řešení problému hledání konvexního obalu a oddaluje ji od aproximace dat ve statistickém pojetí regrese.

3.3 Vnější model (Possibilistic model)

Při běžném zpracování naměřených dat nás může zajímat jednoduchý popis oblasti, ve které všechna naměřená data leží. Tento požadavek nás může dovést k hledání vnějšího modelu intervalové regrese (anglicky possibilistic model), Tanaka (1987). Pro lineární regresi tedy hledáme intervalový pás ohraničený lineární funkcí (nadrovinou) obsahující vstupní data (X, y), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^p$. Ukázku vnějšího lineárního modelu můžeme vidět na obrázku 3.3.



Obrázek 3.3: Vnější model s reálnými vstupními daty (tabulka 3.1).

j	1	2	3	4	5	6	7	8	9	10
x_j	2	4	6	8	10	12	14	16	2	16
y_j	14	16	14	18	18	22	18	22	4	32

Tabulka 3.1: Vstupní data pro vnější model, Hladík a Černý (2012a)

Podobnou snahu můžeme vidět při vytváření intervalových odhadů ve statistice. Statistický intervalový odhad nám dává interval spolehlivosti, v němž se zadanou (často 95% či 99%) pravděpodobností leží parametr regrese. My se budeme zabývat odhady intervalových pásů, které s jistotou obsahují hledaný parametr. Intervalový pás nám pro jednorozměrná data dává souměrný pás konstantní šířky kolem regresní přímky, což nemusí být pro řadu vstupních dat vhodné.

Reálná data

Hledání řešení lineární intervalové regrese pomocí tohoto modelu je analogií hledání vnější obálky řešení intervalového lineárního systému. Chceme nalézt intervalový pás (definovaný parametrem $\boldsymbol{a}$), který „obaluje“ všechna vstupní data (obr. 3.3).

Tento požadavek říká, že pro každé pozorování leží naměřená hodnota y_j v intervalu určeném intervalovým pásem v bodě x_j.

$$\forall j = 1, \dots p \qquad y_j \in \boldsymbol{y}(x_j) = X_{j*}\boldsymbol{a} = \boldsymbol{a}_0 + x_{j1}\boldsymbol{a}_1 + \cdots + x_{jn}\boldsymbol{a}_n$$

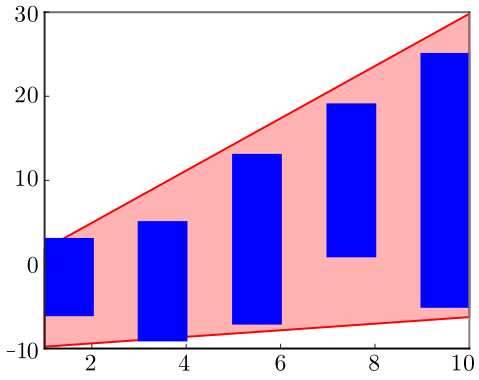
Nalezneme-li vektor regresních parametrů $\boldsymbol{a}$, můžeme pak s jistotou tvrdit, že pro j-té pozorování existuje přímka uvnitř intervalového pásu (určená nějakým konkrétním parametrem a z daného vektoru parametrů $\boldsymbol{a}$), na které leží pozorovaná hodnota y_j. Pro reálná data nám tuto informaci vyjadřuje vztah

$$\forall j = 1, \dots p \qquad \exists a \in \boldsymbol{a} : y_j = X_{j*}a.$$

Jeden z přirozených požadavků na hledaný intervalový pás je, aby byl co nejužší, a tedy obaloval data co nejtěsněji. Formulovaný problém nalezení vektoru $\boldsymbol{a}$ lze řešit pomocí lineárního programování (Hladík a Černý, 2012a), či kvadratického programování (Tanaka a Lee, 1998).

Reálně–intervalová data

V případě, že výstupní data jsou zatížena nějakou chybou (například omezení přesnosti měřícího zařízení), chceme využít obecnější model a to model pro reál-



Obrázek 3.4: Slabý vnější model intervalových dat nemusí obsahovat celý intervalový box (tabulka 3.2)

$\boldsymbol{x}_j$	[1, 2]	[3, 4]	[5, 6]	[7, 8]	[9, 10]
$\boldsymbol{y}_j$	[-6, 3]	[-9, 5]	[-7, 13]	[1, 19]	[-5, 25]

Tabulka 3.2: Vstupní data pro slabý vnější model

nou vstupní matici $X \in \mathbb{R}^{n\times p}$ a intervalový vektor $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p)^\top$. V takovém případě platí následující ekvivalentní vztahy

$$\begin{aligned} &\forall j = 1, \ldots, p \qquad \boldsymbol{y}_j \subseteq X_{j*}\boldsymbol{a}, \\ &\forall j = 1, \ldots, p \qquad \forall y_j \in \boldsymbol{y}_j \; \exists a \in \boldsymbol{a} : y_j = X_{j*}a. \end{aligned}$$

Interval z j-tého pozorování leží uvnitř intervalu daného intervalovým pásem regresního modelu v bodě x_j.

Chceme-li nalézt řešení reálně–intervalového modelu, můžeme využít redukce na případ s reálnými daty. Jako vstupní data označíme všechny krajní body jednotlivých intervalů. Tím se nám zdvojnásobí množství dat. Jestliže najdeme intervalový pás obsahující všechny krajní body, je zřejmé, že bude obsahovat i celé intervaly.

Pokud zeslabíme podmínku, že musí pro všechny hodnoty, kterých může nabývat y_j, existovat přímka (z intervalového pásu), na níž y_j leží, ale stačí nám, že pro j-té pozorování existuje alespoň jedna libovolná přímka z pásu, která prochází bodem z intervalu $\boldsymbol{y}_j$, dostaneme *slabý vnější model* (anglicky weak possibilistic model), tedy model, který má méně striktní podmínky a platí pro něj následující ekvivalentní vztahy

$$\begin{aligned} &\forall j = 1 \ldots p \qquad \boldsymbol{y}_j \cap X_{j*}\boldsymbol{a} \neq \emptyset, \\ &\forall j = 1, \ldots, p \qquad \exists y \in \boldsymbol{y}_j \; \exists a \in \boldsymbol{a} : y = X_{j*}a. \end{aligned}$$

Intervalová data

Pokud zobecníme vnější model pásu pro intervalová data $(\boldsymbol{X}, \boldsymbol{y})$, $\boldsymbol{X} \in \mathbb{IR}^{n\times p}$, $\boldsymbol{y} \in \mathbb{IR}^p$,získáváme ekvivalentní podmínky pro neznámý regresní parametr $\boldsymbol{a}$

$$\begin{aligned} &\forall j = 1, \ldots p \qquad \boldsymbol{y}_j \subseteq \boldsymbol{X}_{j*}\boldsymbol{a}, \\ &\forall j = 1, \ldots, p \qquad \forall y_j \in \boldsymbol{y}_j \; \exists X_{j*} \in \boldsymbol{X}_{j*} \; \exists a \in \boldsymbol{a} : y = X_{j*}a, \end{aligned}$$

které nám ale nezaručují, že regresní pás obsáhne celé intervalové boxy (obrázek 3.4). Takovou podmínku splňuje *silný vnější model* (anglicky strong possibilistic model).

Pro získání silné podmínky stačí změnit podmínku na existenci nějakého bodu $x_j \in \boldsymbol{x}_j$ pro který platí inkluze $\boldsymbol{y}_j \in \boldsymbol{y}(x_j)$ na podmínku, že inkluze musí platit pro všechny body z intervalu $\boldsymbol{x}_j$. Tím získáváme řešení, pro které platí následující ekvivalentní vztahy

$$\begin{aligned} &\forall j = 1, \ldots, p \qquad \forall X_{j*} \in \boldsymbol{X}_{j*} : \boldsymbol{y}_j \subseteq X_{j*}\boldsymbol{a}, \\ &\forall j = 1, \ldots, p \qquad \forall y_j \in \boldsymbol{y}_j \; \forall X_{j*} \in \boldsymbol{X}_{j*} \; \exists a \in \boldsymbol{a} : y = X_{j*}a. \end{aligned}$$

Hledání řešení tohoto modelu lze také redukovat na jednodušší případ. Intervalových dat se lze zbavit tím, že každý box nahradíme pouze jeho krajními vrcholy. Pak výsledný intervalový pás obsahující tyto krajní body musí nutně obsahovat i celý intervalový box.

3.4 Vnitřní model (Necessity model)

V předchozí podkapitole jsme uvažovali přístup, kdy máme intervalový pás, který se snažíme co nejvíce zúžit, aby stále obsahoval všechna data (může však obsahovat i něco navíc). Nyní si představíme trochu protikladný přístup. Vezmeme pás, který prochází všemi naměřenými daty a ten se budeme snažit co nejvíce rozšiřovat, aby stále procházel skrze všechna měření.

Tento přístup se může uplatnit při potvrzování hypotéz o chování dat. Předpokládáme, že nějaká data jsou lineárně závislá, a pokud tomu tak opravdu je, pak existuje tento *vnitřní model*.

Když se vrátíme k definici množiny řešení intervalové soustavy Σ (definice 3.1) definované vstupními daty $(\boldsymbol{X}, \boldsymbol{y})$, můžeme nahlédnout, že množina nadrovin popsaná touto množinou je vždy nadmnožinou vnitřního modelu.

Reálná data

Formálně tedy vnitřní model můžeme pro reálná data popsat jako soustavu lineárních rovnic

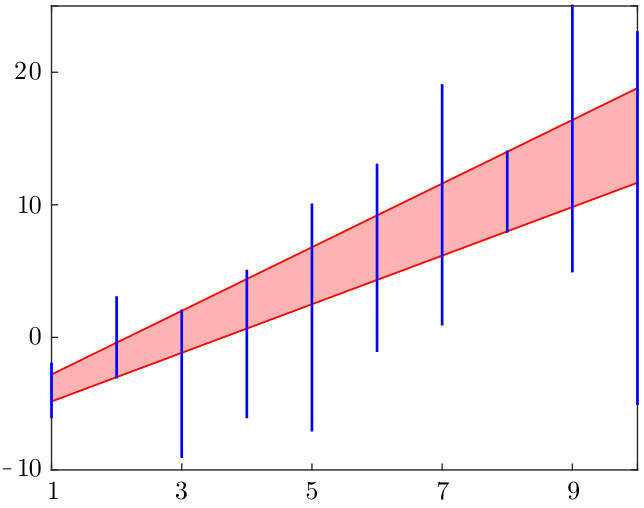
$$\forall j = 1, \ldots, p \qquad y_j = a_0 + x_{j1}a_1 + \cdots + x_{jn}a_n = X_{j*}a.$$

Protože vstupní data jsou reálná a vynucujeme rovnost, je vektor parametrů degenerovaný a také pouze reálný. Řešení existuje jen v případě, že všechna data leží v jedné nadrovině (resp. na přímce).

Reálně–intervalová data

Již zobecnění pro reálnou matici X a intervalový vektor $\mathbf{y}$ začíná být vnitřní přístup zajímavější (obr. 3.5). Hledání intervalového pásu lze popsat následujícími ekvivalentními vztahy

$$\begin{aligned} &\forall j = 1, \ldots, p \qquad \mathbf{y}_j \supseteq X_{j*}\mathbf{a}, \\ &\forall j = 1, \ldots, p \qquad \forall a \in \mathbf{a} \; \exists y \in \mathbf{y}_j : y_j = X_{j*}a. \end{aligned}$$



Obrázek 3.5: Vnitřní lineární model s reálně-intervalovými daty (tab. 3.3)

x_j	1	2	3	4	5
$\boldsymbol{y}_j$	[-6, -2]	[-3, 3]	[-9, 2]	[-6, 5]	[-7, 10]
x_j	6	7	8	9	10
$\boldsymbol{y}_j$	[-1, 13]	[1, 19]	[8, 14]	[5, 25]	[-5, 23]

Tabulka 3.3: Vstupní data pro vnitřní model

Je tedy požadováno, aby každá nadrovina z vymezeného intervalového pásu protínala všechna měření.

Intervalová data

Zobecnění na intervalová data nám umožňuje dvě různé formulace vnitřního modelu. U *silného vnitřního* modelu musí intervalový pás procházet celou šířkou každého boxu. Formální zápis

$$\forall j = 1, \ldots, p: \ \forall x_j \in \mathbf{X}_{j*}\ \mathbf{y}_j \supseteq \mathbf{y}(x_j).$$

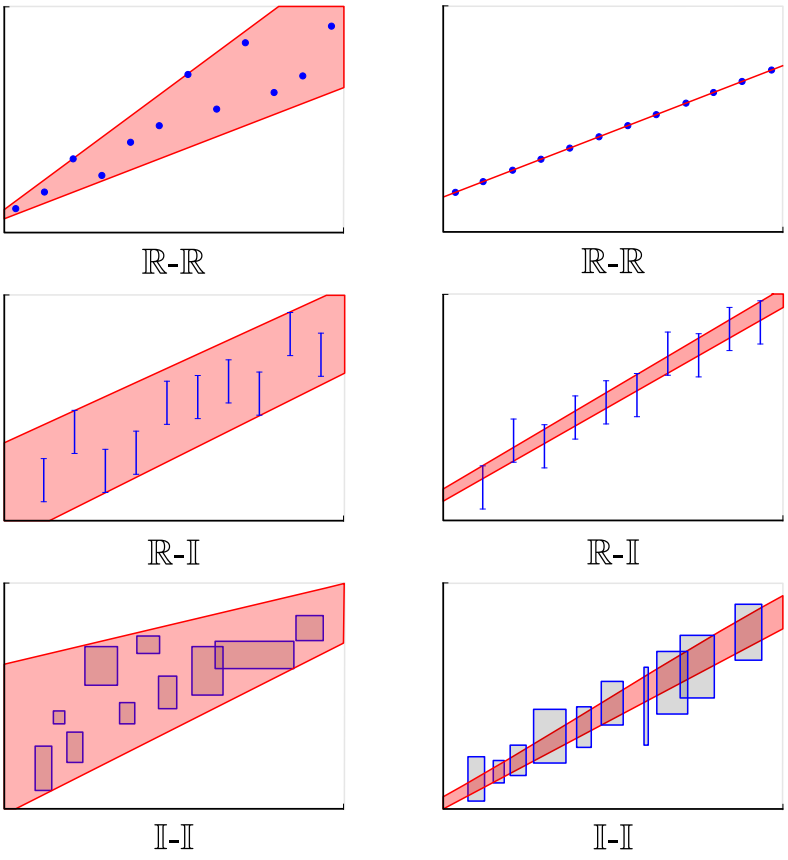
Slabý vnitřní model můžeme formulovat z představy, že pro každý box musí existovat interval $y_j \in \mathbf{y}_j$, který je protnut oběma krajními přímkami intervalového pásu (tj. krajní přímky se mohou dotýkat boxu pouze v rozích, avšak tyto rohy musí být nad sebou).

$$\forall j = 1, \ldots, p: \ \exists x_j \in \mathbf{X}_{j*}\ \mathbf{y}_j \supseteq \mathbf{y}(x_j)$$

3.5 Srovnání typů lineární intervalové estimace

Představili jsme několik přístupů k lineární intervalové regresi. Nyní porovnáme jednotlivé definice. Budeme diskutovat o extrémních řešeních jednotlivých modelů a vše bude ilustrováno názornými obrázky.

Při řešení problému lineární intervalové regrese předpokládáme lineární vztah závislé proměnné $\boldsymbol{y}$ (pozorované veličiny) na nezávislé proměnné $\boldsymbol{X}$ (vstupních



(a) Vnější model (b) Vnitřní model

Obrázek 3.6: Srovnání vnější a vnitřní estimace

parametrech experimentu). Podle typu vstupních dat se tedy regrese vždy snaží nalézt co nelepší aproximaci řešení některé následující intervalové soustavy.

$$\begin{aligned} \mathbb{R}\text{-}\mathbb{R}: &\quad X \in \mathbb{R}^{n\times p}, y \in \mathbb{R}^p: &\quad y = X\boldsymbol{a} \\ \mathbb{R}\text{-}\mathbb{I}: &\quad X \in \mathbb{R}^{n\times p}, \boldsymbol{y} \in \mathbb{IR}^p: &\quad \boldsymbol{y} = X\boldsymbol{a} \\ \mathbb{I}\text{-}\mathbb{I}: &\quad \boldsymbol{X} \in \mathbb{IR}^{n\times p}, \boldsymbol{y} \in \mathbb{IR}^p: &\quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} \end{aligned}$$

Geometrická představa základních typů vnějšího a vnitřního přístupu (pro vstupní data, kde $n = 2$) je znázorněna na obrázku 3.6.

Pomocí množinových operací můžeme definice zapsat vztahy z tabulky 3.4. Můžeme snadno vidět, že hledání vnějšího intervalového pásu se s hledáním vnitřního pásu pro všechny typy dat liší obrácením inkluze, resp. znaménka náležení. Pro vnitřní model bychom u reálných dat mohli rovnost též nahradit vztahem náležení a dostali bychom ekvivalentní vyjádření modelu. Kvůli tomu, že y je reálný vektor, podmínka vynucuje degenerovanost vektoru regresních parametrů a a model tedy nemá řešení, pokud zde nenastane přímo rovnost.

Data	Vnější pás	Vnitřní pás
$\mathbb{R}$-$\mathbb{R}$	$\forall j:\ y_j \in X_{j*}\mathbf{a}$	$\forall j:\ y_j = X_{j*}a$
$\mathbb{R}$-$\mathbb{I}$	$\forall j: \mathbf{y}_j \subseteq X_{j*}\mathbf{a}$	$\forall j: \mathbf{y}_j \supseteq X_{j*}\mathbf{a}$
$\mathbb{I}$-$\mathbb{I}$	$\forall j\ \forall X_{j*} \in \mathbf{X}_{j*}: \mathbf{y}_j \subseteq X_{j*}\mathbf{a}$	$\forall j\ \forall X_{j*} \in \mathbf{X}_{j*}: \mathbf{y}_j \supseteq X_{j*}\mathbf{a}$

Tabulka 3.4: Porovnání vnější a vnitřní estimace pro různé typy vstupních dat (pro $\mathbb{R}$-$\mathbb{I}$ a $\mathbb{I}$-$\mathbb{I}$ data zde vidíme silnou variantu modelů)

Nyní se podívejme na všechny definice vnějších modelů najednou (v tabulce 3.5) a srovnejme si je. Na ilustraci jednotlivých variant (obr. 3.7), můžeme vidět rozdíly mezi definicí silné a slabé formulace modelu.

Pro reálná data existuje vždy řešení vnějšího modelu, protože požadavek na uzavření všech dat v intervalovém pásu lze splnit volbou dostatečně vzdálených hraničních přímek. Hranice lze k sobě libovolně přibližovat, dokud nenarazí na body vstupních dat. Pokud všechna data leží na přímce, může intervalový pás degenerovat až do reálné přímky.

Data	Varianta	Množinový zápis	Alternativní zápis
$\mathbb{R}$-$\mathbb{R}$	–	$\forall j: y_j \in X_{j*}\boldsymbol{a}$	$\forall j\ \exists a \in \boldsymbol{a}: y_j = X_{j*}a$
$\mathbb{R}$-$\mathbb{I}$	silná	$\forall j: \boldsymbol{y}_j \subseteq X_{j*}\boldsymbol{a}$	$\forall j\ \forall y_j \in \boldsymbol{y}_j\ \exists a \in \boldsymbol{a}: y_j = X_{j*}a$
$\mathbb{R}$-$\mathbb{I}$	slabá	$\forall j: \boldsymbol{y}_j \cap X_{j*}\boldsymbol{a} \neq \emptyset$	$\forall j\ \exists y_j \in \boldsymbol{y}_j\ \exists a \in \boldsymbol{a}: y_j = X_{j*}a$
$\mathbb{I}$-$\mathbb{I}$	silná	$\forall j\ \forall X_{j*} \in \boldsymbol{X}_{j*}: \boldsymbol{y}_j \subseteq X_{j*}\boldsymbol{a}$	$\forall j\ \forall y_j \in \boldsymbol{y}_j\ \forall X_{j*} \in \boldsymbol{X}_{j*}\ \exists a \in \boldsymbol{a}: y_j = X_{j*}a$
$\mathbb{I}$-$\mathbb{I}$	slabá	$\forall j: \boldsymbol{y}_j \subseteq \boldsymbol{X}_{j*}\boldsymbol{a}$	$\forall j\ \forall y_j \in \boldsymbol{y}_j\ \exists X_{j*} \in \boldsymbol{X}_{j*}\ \exists a \in \boldsymbol{a}: y_j = X_{j*}a$

Tabulka 3.5: Přehled variant vnějšího modelu

Pro reálně-intervalová data zde máme dvě definice, silnou a slabou. Z obrázků je snadno vidět, že pokud nalezneme co nejtěsnější řešení, je řešením i libovolné rozšíření tohoto pásu (ve směru šipek). Silná formulace vynucuje, že všechny intervaly musí být celé obsažené uvnitř nalezeného řešení. Slabá varianta vynucuje pouze neprázdný průnik pásu s každým intervalem, může tedy degenerovat až na reálnou přímku. Přesněji může degenerovat až na řešení vnitřního modelu.

Posledním typem dat jsou data nejobecnější, intervalová. Silná formulace opět vynucuje, že všechny data musí být uvnitř pásu. Celý box tedy musí ležet uvnitř, pokud mají všechny boxy stejnou výšku a jejich středy leží na přímce, můžeme nalézt extrémní řešení, které těsně zapouzdřuje všechna data.

Slabá varianta intervalového vnějšího modelu nemusí nutně obsahovat každý box celý. Může degenerovat dokonce až na reálnou přímku. Tato přímka ale nutně musí protínat každý box v celé jeho výšce, tedy musí existovat úsečka spojující horní hraniční interval se spodním hraničním intervalem, která celá leží uvnitř nalezeného pásu. Pro dostatečně široký pás je zřejmé, že je tato podmínka splněna.

Data	Varianta	Množinový zápis	Alternativní zápis
$\mathbb{R}$-$\mathbb{R}$	–	$\forall j : y_j \ni X_{j*}a$	$\forall j : y_j = X_{j*}a$
$\mathbb{R}$-$\mathbb{I}$	–	$\forall j : \boldsymbol{y}_j \supseteq X_{j*}\boldsymbol{a}$	$\forall j\ \forall a \in \boldsymbol{a}\ \exists y_j \in \boldsymbol{y}_j : y_j = X_{j*}a$
$\mathbb{I}$-$\mathbb{I}$	silná	$\forall j\ \forall X_{j*} \in \boldsymbol{X}_{j*} : \boldsymbol{y}_j \supseteq X_{j*}\boldsymbol{a}$	$\forall j\ \forall a \in \boldsymbol{a}\ \forall X_{j*} \in \boldsymbol{X}_{j*}\ \exists y_j \in \boldsymbol{y}_j : y_j = X_{j*}a$
$\mathbb{I}$-$\mathbb{I}$	slabá	$\forall j\ \exists X_{j*} \in \boldsymbol{X}_{j*} : \boldsymbol{y}_j \supseteq X_{j*}\boldsymbol{a}$	$\forall j\ \forall a \in \boldsymbol{a}\ \exists X_{j*} \in \boldsymbol{X}_{j*}\ \exists y_j \in \boldsymbol{y}_j : y_j = X_{j*}a$

Tabulka 3.6: Přehled variant vnitřního modelu

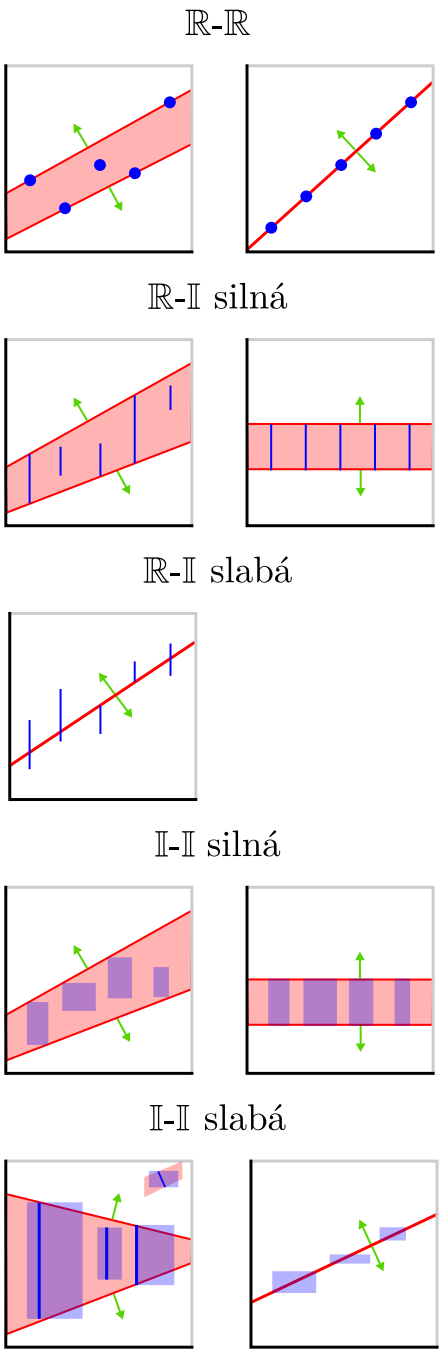
Přehled definic všech vnitřních modelů je v tabulce 3.6, vizualizace jednotlivých definic je pak na obrázku 3.8. Hlavní myšlenkou hledání tohoto modelu je nalézt intervalový pás, který je co nejširší a zároveň prochází všemi daty, tedy každá přímka z tohoto pásu protíná všechna vstupní data (v celé jejich šířce).

Nejjednodušším případem vstupních dat jsou data reálná. Avšak hned ta mohou být neřešitelným problémem, protože pokud všechna neleží na jedné přímce, neexistuje řešení vnitřního modelu. Pro reálná data má tedy tento model, pokud existuje, vždy tvar reálné přímky (resp. nadroviny).

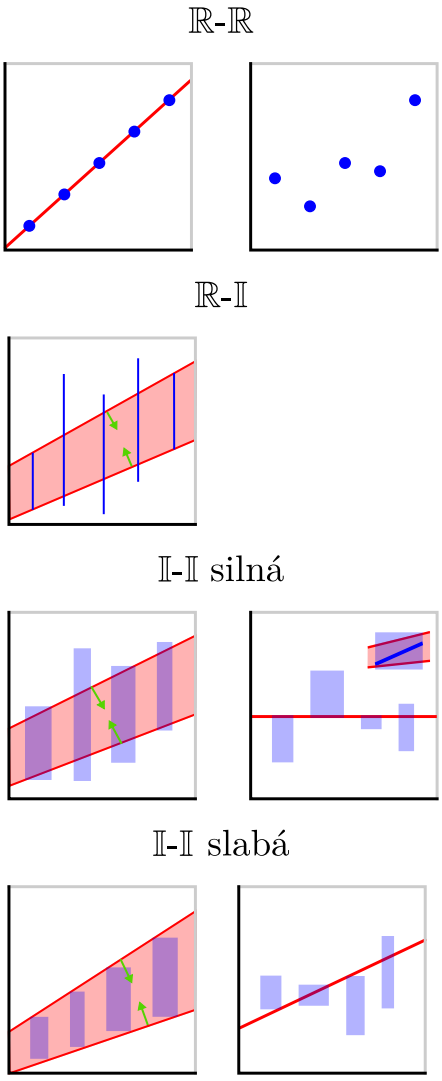
Podíváme-li se na data reálně-intervalová, opět nám definice modelu dává podmínku, že každá přímka pásu musí procházet všemi daty. Pokud jsou data příliš roztroušená, znamená to, že vnitřní model nemusí vůbec existovat, což může znamenat, že modelované proměnné mezi sebou nemají lineární vztah a je nutné zvolit jiný model. Jistým ukazatelem toho, jak moc vzdáleni jsme od existence řešení vnitřního modelu, může být řešení slabé formulace vnějšího modelu. Čím užší je nalezený pás vnějšího slabého modelu, tím blíže jsme k řešení vnitřního modelu. Pokud vnitřní model existuje, pak vnější slabý model může degenerovat až na reálnou přímku.

Jak se chová vnitřní model pro intervalová data? Opět je jistým protikladem vnějšího modelu. U slabé (i silné) formulace problému vynucoval vnější model to, že musí existovat přímka protínající celou „výšku“ boxu. Silná varianta vniřního modelu vynucuje to, že každá přímka z intervalového pásu modelu musí protínat celou „šířku“ každého boxu (na obrázku jsou vidět ukázky nejextrémnější varianty šířky pásu).

Slabá varianta vnitřního modelu je volnější a říká, že pro každý box stačí jeden svislý interval pásu, který je podmnožinou intervalu odpovídajícího boxu. Pokud je tento svislý interval degenerovaný, odpovídá tomu i pouhý jediný bod, tedy stačí, pokud pás protne každý box v alespoň jednom bodě.



Obrázek 3.7: Vnější model přehled



Obrázek 3.8: Vnitřní model přehled

4. Postupy lineární estimace

V této kapitole shrneme možné přístupy k hledání lineární intervalové estimace dat. Nejdříve popíšeme, jak lze využít lineárního a kvadratického programování, poté představíme toleranční přístup a na závěr odečítací metodu.

Nyní zmíníme užitečnou formulaci intervalu $\boldsymbol{y}(x_j)$ (Tanaka a Lee (1998)) pro intervalový lineární model. Tato formulace vyjadřuje v jakém intervalu se nachází naměřená hodnota y_j pro vstupní bod x_j.

$$\boldsymbol{y}(x_j) = \left[X_{j*}a^c - |X|_{j*}a^\Delta, X_{j*}a^c + |X|_{j*}a^\Delta\right] \tag{4.1}$$

Tvar intervalu vychází z intervalové aritmetiky a získáme ho rozepsáním modelu intervalové lineární regrese $\boldsymbol{y}(x_j)$.

$$\begin{gathered}\boldsymbol{y}(x_j) = X_{j*}\boldsymbol{a} = \boldsymbol{a}_0 + x_{j1}\boldsymbol{a}_1 + \cdots + x_{jn}\boldsymbol{a}_n = \\ = \left[a_0^c - a_0^\Delta, a_0^c + a_0^\Delta\right] + x_{j1}\left[a_1^c - a_1^\Delta, a_1^c + a_1^\Delta\right] + \cdots + x_{jn}\left[a_n^c - a_n^\Delta, a_n^c + a_n^\Delta\right] = \\ = \left[X_{j*}a^c - |X|_{j*}a^\Delta, X_{j*}a^c + |X|_{j*}a^\Delta\right]\end{gathered}$$

□

4.1 Lineární programování

Přístup lineárního programování(Peters, 1994) vychází z výše zmíněné formulace intervalu 4.1. Popíšeme význam jednotlivých částí lineárního programu (LP) a ilustrujeme geometrický význam na obrázku.

4.1.1 Reálná data

Vnější model

Mějme vstupní data (X, y), kde $X \in \mathbb{R}^{p\times n}$, $y \in \mathbb{R}^p$. Uvažujme model lineární regrese $\boldsymbol{y}(x) = x^\top \boldsymbol{a}$ a vektor parametrů ve tvaru $\boldsymbol{a} = \left[a^c - a^\Delta, a^c + a^\Delta\right]$.

Zformulujeme-li problém hledání (co nejmenší) vnější obálky vstupních dat, získáme přímočaře formulaci lineárního programu (ilustrační obrázek 4.1).

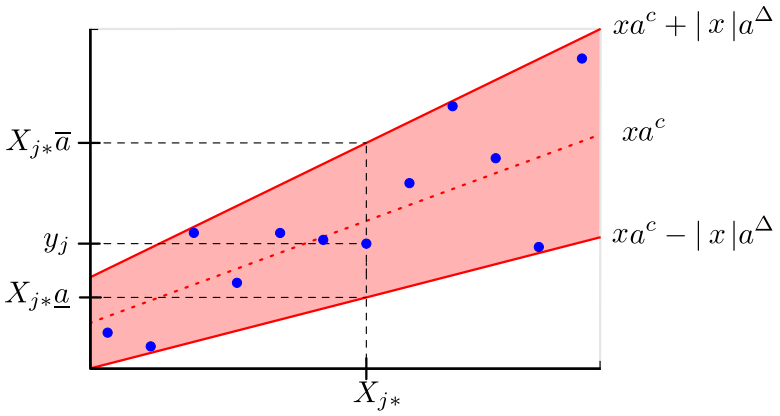
$$\min_{a^c, a^\Delta} \sum_{j=1}^{p} |X|_{j*}a^\Delta \tag{4.2}$$

$$\forall j = 1, \ldots, p \qquad y_j \le X_{j*}a^c + |X|_{j*}a^\Delta \tag{4.3}$$

$$\forall j = 1, \ldots, p \qquad y_j \ge X_{j*}a^c - |X|_{j*}a^\Delta \tag{4.4}$$

$$a^\Delta \ge 0 \tag{4.5}$$

Hledáme intervalovou přímku $P(x) = x\boldsymbol{a}$ určenou vektorem parametrů $\boldsymbol{a}$, která obsahuje všechna vstupní data. Geometrický význam některých částí lineárního programu je znázorněn na obrázku 4.1.Pro každé pozorování každý bod



Obrázek 4.1: Lineární program pro reálná data

x_j je poloměr vektoru parametrů pásu $|X_{j*}|a^\Delta$. Abychom získali co nejtěsnější obálku dat, chceme minimalizovat vzdálenost hraničních přímek od vstupních dat, což vyjadřuje účelová funkce (4.2) minimalizující součet poloměrů přímky v bodech jednotlivých pozorování.

Aby byla splněna podmínka, že intervalová přímka obsahuje vstupní data, musí být horní hraniční přímka

$$P_h(x) = x\overline{a} = xa^c + |x|a^\Delta$$

ležet „nad“ daty, formálně tedy nabývá pro každé x_j hodnotu alespoň y_j. Splnění této podmínky pro přípustné řešení lineárního programu zajišťuje první sada nerovností (4.3). Analogicky hraniční přímka intervalového pásu určující dolní mez $P_d(x) = x\underline{a} = xa^c - |x|a^\Delta$ musí být „pod“ daty a to je splněno díky druhé sadě podmínek (4.4). V neposlední řadě musí být splněno, že poloměr intervalové přímky musí být nezáporný, což zaručuje nerovnost (4.5).

Vnitřní model

Podíváme-li se na definici vnitřního modelu pro reálná vstupní data (X, y), v definici máme oproti vnějšímu modelu obrácen vztah náležení.

$$y_j \ni X\boldsymbol{a}$$

Reálnost vstupního vektoru y vynucuje to, že přípustné řešení modelu musí mít degenerovaný intervalový vektor parametrů, tedy $a \in \mathbb{R}^n$. Pro nalezení řešení modelu, pokud existuje, proto stačí pomocí libovolné metody lineární algebry vyřešit reálnou soustavu rovnic $y = Xa$.

4.1.2 Reálně-intervalová data

Vnější model

Přejdeme-li od modelu s reálnými vstupními daty k reálně-intervalovým datům $(X, \boldsymbol{y})$, formulace lineárního programu se výrazně nezmění. Stačí poupravit podmínky pro jednotlivé hraniční přímky a to tak, aby horní hranice ležela „nad“ horními kraji intervalů $\boldsymbol{y}_j$ (nerovnice (4.7)) a dolní hranice ležela „pod“ dolními

hranicemi intervalů (nerovnice (4.8)). Účelová funkce a pravé části nerovností zůstávají nezměněny, pouze se vhodně nahradí y_j za krajní body (červeně zvýrazněno). Podmínka (4.9) na nezápornost poloměru vektoru parametrů regrese zůstává také stejná.

$$\min_{a^c, a^\Delta} \sum_{j=1}^{p} |X|_{j*} a^\Delta \tag{4.6}$$

$$\forall j = 1, \ldots, p \qquad \overline{y}_j \leq X_{j*} a^c + |X|_{j*} a^\Delta \tag{4.7}$$

$$\forall j = 1, \ldots, p \qquad \underline{y}_j \geq X_{j*} a^c - |X|_{j*} a^\Delta \tag{4.8}$$

$$a^\Delta \geq 0 \tag{4.9}$$

Vnitřní model

Vstupní data máme opět $(X, \boldsymbol{y})$ avšak zásadně změníme podmínku na hledanou intervalovou přímku. Chceme, aby procházela všemi měřeními a byla co nejširší. To se v lineárním programu promítne tím, že se otočí nerovnosti pro hraniční přímky, viz (4.11) a (4.12), a místo minimalizace poloměrů pásu je budeme maximalizovat (4.10).

$$\max_{a^c, a^\Delta} \sum_{j=1}^{p} |X|_{j*} a^\Delta \tag{4.10}$$

$$\forall j = 1, \ldots, p \qquad \overline{y}_j \geq X_{j*} a^c + |X|_{j*} a^\Delta \tag{4.11}$$

$$\forall j = 1, \ldots, p \qquad \underline{y}_j \leq X_{j*} a^c - |X|_{j*} a^\Delta \tag{4.12}$$

$$c \geq 0 \tag{4.13}$$

4.1.3 Intervalová data

Zobecněný problém pro intervalová data $(\boldsymbol{X}, \boldsymbol{y})$ jde převést na reálný problém. Uvažujme silné varianty vnějšího modelu. Nahradíme vstupní data reálnými body, které odpovídají vrcholům jednotlivých intervalových boxů (pro každé pozorování). Získáním vnějšího pásu zapouzdřujícího tyto body, získáme i korektní vnější pás zapouzdřující vstupní data. Bohužel se tím významně zvětší počet vstupních dat, protože je třeba vzít všechny kombinace krajních hodnot intervalů $\boldsymbol{X}_{j*}$, $\boldsymbol{y}_j$. Lineární program by tedy byl shodný s programem pro reálná data, ale s větším počtem podmínek.

4.1.4 Problémy lineárního programování

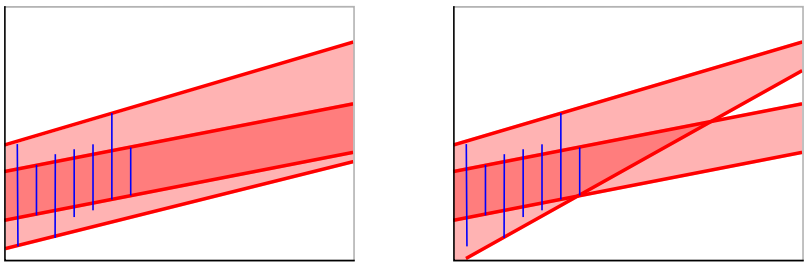
Nyní zmíníme několik špatných vlastností řešení nalezených pomocí představených lineárních programů. V následujících kapitolách představíme přístupy k řešení jednotlivých problémů.

Řešení nalezená pomocí zformulovaných lineárních programů jsou velmi citlivá na chyby. Pokud vstupní data obsahují i jen malé množství chybně naměřených dat, která jsou velmi vzdálená od správných dat, výstupní intervalový pás bude velmi široký a nebude dobře vypovídat o trendu chování dat (Peters, 1994).

Hledání vnějšího $P_{vnější}$ a vnitřního pásu $P_{vnitřní}$ prokládající vstupní data pomocí zmíněných dvou lineárních programů nemusí najít řešení, které splňuje podmínku inkluze (Tanaka a Lee, 1998), že vnější pás obsahuje vnitřní.

$$\text{Tanaka-Lee:} \qquad P_{vnitřní} \subseteq P_{vnější}. \tag{4.14}$$

Podmínky formulovaných programů zajišťují, že inkluze platí pouze pro oblast vstupních dat, ale při predikci chování dat se mohou modely rozbíhat, viz obrázek 4.2.



(a) Inkluze je splněna (b) Inkluze neplatí

Obrázek 4.2: Tanaka-Lee podmínka inkluze

Tento problém lze řešit pomocí formulace lineárního programu, který řeší oba problémy najednou (Ishibuchi, 1993).

Dalším problémem nastíněným v článku (Hladík a Černý, 2012a) je špatná centralita nalezeného řešení. Jako vylepšení přístupu, který řeší problém podmínky inkluze (4.14) a zároveň zohledňuje problém centrality, Tanaka vytvořil řešení lineární intervalové estimace pomocí kvadratického programování. Ten představíme v následující části práce.

Posledním významným problémem je tendence k degeneraci některých parametrů. Intervalový parametr je pouze reálné číslo, což může způsobit, že jiné parametry jsou pak příliš široké a model opět není vyhovující.

4.2 Kvadratické programování

Nejdříve uvedeme základní formulaci kvadratického přístupu (Tanaka a Lee, 1998) k nalezení vnější obálky, dále uvedeme kvadratický program zohledňující centralitu nalezeného estimátoru. Jako řešení problému inkluze vnějšího a vnitřního odhadu uvedeme kvadratický program řešící oba odhady zároveň.

V této podkapitole tedy nerozdělujeme jednotlivé problémy pro různé typy dat zvlášť na hledání vnějšího a vnitřního modelu. Pro reálná data je řešením vnitřního modelu pouze řešení soustavy lineárních rovnic, proto se zabýváme pouze vnějším modelem. U dalších typů dat pak řešíme oba problémy najednou, pomocí jednoho kvadratického programu.

4.2.1 Reálná data

Pro reálná vstupní data (X, y) předpokládejme, že mohou být popsána lineárním regresním modelem $\boldsymbol{y}(x) = x^\top \boldsymbol{a}$. Hledání vnější obálky nám dává podmínku

$$\forall j = 1, \ldots, p: \;\; y_j \in \boldsymbol{y}(x_{j*}),$$

čímž získáváme nerovnosti určující horní (4.16) a dolní (4.17) meze pro kvadratický program shodné s LP pro hledání vnější intervalové obálky.

Účelová funkce (4.15) má za úkol minimalizovat poloměr pásu, a proto je definována jako součet čtverců vzdálenosti krajů pásu v jednotlivých bodech X_{j*}. Oproti LP je zde navíc přidán člen zohledňující hodnotu středu parametrického vektoru $(a^c)^\top a^c$ přenásobený malým kladným číslem ε.

$$\min_{a^c, a^\Delta} \sum_{j=1}^{p} (|X|_{j*} a^\Delta)^2 + \varepsilon (a^c)^\top a^c \tag{4.15}$$

$$\forall j = 1, \ldots, p \qquad y_j \leq X_{j*} a^c + |X|_{j*} a^\Delta \tag{4.16}$$

$$\forall j = 1, \ldots, p \qquad y_j \geq X_{j*} a^c - |X|_{j*} a^\Delta \tag{4.17}$$

$$a^\Delta \geq 0 \tag{4.18}$$

Formulovaný kvadratický program má menší tendenci k degeneraci vektoru parametrů díky tomu, že a^Δ nemá sklon být reálný.

Stále však není řešen problém, že středová přímka nalezeného intervalového pásu špatně prokládá data. To lze vyřešit přidáním k účelové funkci výrazu, který splňuje odhad pomocí nejmenších čtverců. Ten minimalizuje sumu čtverců vzdáleností dat od přímky, kterou data prokládáme.

$$\sum_{j=1}^{p} (y_j - \boldsymbol{a}^\top X_{j*})^2$$

Stačí tedy použít jako účelovou funkci kombinaci obojího

$$\min_{a^c, a^\Delta} k_1 \sum_{j=1}^{p} \left(|X|_{j*} a^\Delta \right)^2 + k_2 \sum_{j=1}^{p} \left(y_j - \boldsymbol{a}^\top X_{j*} \right)^2 . \tag{4.19}$$

Koeficienty k_1, k_2 pak dynamicky mohou určovat, jak velkou váhu chceme přikládat centralitě hledaného modelu lineární regrese.

4.2.2 Reálně-intervalová data

Vstupní data řešeného problému jsou reálná matice $X \in \mathbb{R}^{p \times n}$ a intervalový vektor $\boldsymbol{y} \in \mathbb{IR}^p$. Uvažme dva regresní modely. První model nám má dát vnitřní intervalový pás prokládající vstupní data. Druhý pak vnější intervalový pás obsahující všechna data. Předpokládejme, že oba modely mají stejnou středovou funkci a liší se pouze poloměrem. Proto vyjádříme vektory parametrů pro vnitřní a vnější model následujícím způsobem:

$$\boldsymbol{a}_{vnitřní} = (a^c, a^\Delta_{vnitřní})$$

$$\boldsymbol{a}_{vnější} = (a^c, a^\Delta_{vnější}) = (a^c, a^\Delta_{vnitřní} + d)$$

Pro vnitřní model chceme maximalizovat poloměr hledaného pásu, pro vnější model ho chceme minimalizovat. To vyjadřují následující funkce.

$$\max \sum_{j=1}^{p} |X_{j*}| a^\Delta_{vnitřní}$$

$$\min \sum_{j=1}^{p} |X_{j*}| a^{\Delta}_{vnější}$$

Sjednocení těchto požadavků do jedné účelové funkce získáme tak, že odečteme funkci pro vnitřní pás od účelové funkce pro vnější pás. Následným zjednodušením získáme výraz (4.20).

$$\min \left(\sum_{j=1}^{p} |X_{j*}| a^{\Delta}_{vnější} - \sum_{j=1}^{p} |X_{j*}| a^{\Delta}_{vnitřní} \right)$$

$$\min \sum_{j=1}^{p} |X_{j*}| \left(a^{\Delta}_{vnější} - a^{\Delta}_{vnitřní} \right)$$

$$\min \sum_{j=1}^{p} |X_{j*}| (a^{\Delta}_{vnitřní} + d - a^{\Delta}_{vnitřní})$$

$$\min \sum_{j=1}^{p} |X_{j*}| d \tag{4.20}$$

Formulace problému kvadratického programování řešící oba modely a navíc splňující podmínku inkluze (4.14) vypadá takto:

$$\min_{a^c, a^{\Delta}_{vnitřní}, d} \sum_{j=1}^{p} (|X_{j*}|d)^2 + \varepsilon[(a^c)^2 + (a^{\Delta}_{vnitřní})^2] \tag{4.21}$$

$$j = 1, \ldots, p \tag{4.22}$$

$$\overline{y_j} \leq X_{j*} a^c + |X_{j*}| a^{\Delta}_{vnitřní} + |X_{j*}| d \tag{4.23}$$

$$\underline{y_j} \geq X_{j*} a^c - |X_{j*}| a^{\Delta}_{vnitřní} - |X_{j*}| d \tag{4.24}$$

$$\overline{y_j} \geq X_{j*} a^c + |X_{j*}| a^{\Delta}_{vnitřní} \tag{4.25}$$

$$\underline{y_j} \leq X_{j*} a^c - |X_{j*}| a^{\Delta}_{vnitřní} \tag{4.26}$$

$$a^{\Delta}_{vnitřní} \geq 0 \tag{4.27}$$

$$d \geq 0 \tag{4.28}$$

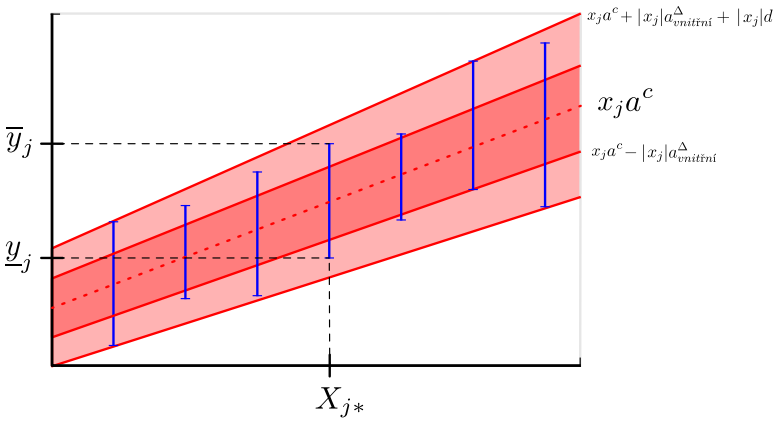
(Na obrázku 4.3 je vidět geometrický význam některých částí kvadratického programu.) První dvě sady nerovností (4.23, 4.24) udávají podmínku, že vnější pás musí ležet vně všech $\boldsymbol{y}_j$, a následující dvě sady nerovností (4.25, 4.26) korespondují s tím, že vnitřní pás musí ležet uvnitř krajních bodů jednotlivých $\boldsymbol{y}_j$.

Podívejme se na hledané přípustné řešení daného kvadratického programu. Z podmínky, že vnější obálka obsahuje všechna data $\boldsymbol{y}_j \subseteq X_{j*}\boldsymbol{a}$, plyne, že vždy existuje přípustné řešení tohoto problému. Stačí zvolit dostatečně velkou šířku intervalového pásu. Řešení vnitřního lineárního modelu však existovat nemusí.

4.3 Toleranční přístup

V této podkapitole navážeme tolerančním přístupem, který nám dává postup k nalezení intervalového vektoru parametrů vnějšího a vnitřního modelu. Tento přístup byl publikován v článku Hladík a Černý (2012a).

Toleranční přístup má dva hlavní kroky:



Obrázek 4.3: Ilustrační obrázek ke kvadratickému programu pro nalezení vnějšího a vnitřního intervalového lineárního modelu

1. Nalezení centrálního vektoru parametrů přímky pomocí běžných statistických aproximací dat.

2. Spočítání nejvhodnějšího poloměru vektoru parametrů modelu.

4.3.1 Reálná data

Vnější model

Pro reálná vstupní data (X, y) a vnější model chceme nalézt vektor parametrů regrese $\boldsymbol{a}$ takový, aby výsledný intervalový pás určený tímto parametrem obsahoval všechna data, tedy

$$\forall j = 1, \ldots, p: \; y_j \in X_{j*}\boldsymbol{a}.$$

Pro tento účel využijeme formulaci vektoru parametrů $\boldsymbol{a}$ pomocí poloměru. Samotný poloměr vyjádříme jako součin tolerančního vektoru c a koeficientu δ, jejichž přesný význam bude vzápětí vysvětlen.

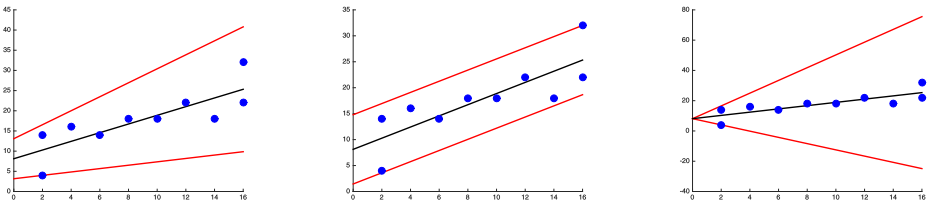
$$\boldsymbol{a} = [a^c - a^{\Delta}, a^c + a^{\Delta}] = [a^c - \delta c, a^c + \delta c]$$

Vektor a^c je středem intervalu $\boldsymbol{a}$. Je určen pomocí libovolného odhadu, který dobře aproximuje vstupní data. Můžeme použít například metodu nejmenších čtverců, tedy $a^c = (X^\top X)^{-1} X^\top y$, nebo libovolný jiný statistický přístup pro reálná data. Tento postup odpovídá prvnímu kroku tolerančního přístupu.

Vektor c je daný toleranční vektor určující míru vlivu jednotlivých parametrů na šířku výsledného intervalového pásu. Vektor c je nezáporný. Lze ho definovat jako vektor jedniček, pokud chceme absolutní toleranci, či jako $c = |a^c|$ pro relativní toleranci. Další možností je pro nějakou zvolenou indexovou množinu I definovat

$$c_i = \begin{cases} |a_i^c| & \text{pro } i \in I, \\ 0 & \text{pro } i \notin I. \end{cases}$$

Na obrázku 4.4 můžeme vidět vliv hodnot tolerančního vektoru na chování estimátoru. Volbou vhodné indexační množiny lze docílit, že pás bude obsahovat



(a) $c = |a^c| = (|a_0^c|, |a_1^c|)$ (b) $c = (|a_0^c|, 0)$ (c) $c = (0, |a_1^c|)$

Obrázek 4.4: Vliv tolerančního vektoru na tvar nalezeného řešení. Pás je intervalová přímka $y = \boldsymbol{a}_0 + \boldsymbol{a}_1 x$. (Středová přímka je znázorněna černě, hraniční přímky červeně, data viz 3.1.)

pouze přímky rovnoběžné se středovou (obr. 4.4b), případně přímky lišící se pouze sklonem (obr. 4.4c).

Abychom získali výsledný intervalový vektor $\boldsymbol{a}$ parametrů určující intervalovou přímku obsahující všechna vstupní data, potřebujeme určit poloměr pásu a to pomocí určení hodnoty tolerančního koeficientu δ. Aby byl splněn přirozený požadavek na co nejužší výsledný pás, je třeba určit δ jako co nejmenší. Konstantu určíme podle následující věty 4.1 jako $\delta := \delta^*$ tak, aby odpovídala těsně bodu nejvzdálenějšímu od středové přímky. Tedy všechny body včetně nejvzdálenějšího, budou obsaženy uvnitř intervalového pásu.

Toleranční koeficient δ lze interpretovat jako ukazatel toho, jak dobře nalezený model odpovídá vstupním datům. Pokud je $\delta^* = 0$, znamená to, že data leží přímo na nalezené přímce, tj. $y = Xa$. Čím větší je toleranční koeficient, tím hůře kopíruje model chování dat.

Věta 4.1. Pokud pro nějaké $j \in \{1, \ldots, p\}$ platí, že $|X|_{j*}c = 0$ a $y_j \neq X_{j*}a^c$, pak neexistuje žádná δ splňující:

$$\forall j \in \{1, \ldots, p\}\ \exists a' \in [a^c - \delta c, a^c + \delta c] : y_j = X_{j*}a'.$$

Jinak existuje

$$\delta^* := \max_{j:|X|_{j*}c>0} \frac{|y_j - X_{j*}a^c|}{|X|_{j*}c},$$

kde z definice $\max \emptyset = 0$. Pak δ^* je nejmenší toleranční koeficient.

Celý důkaz zmíněné věty je možné najít v Hladík a Černý (2010).

4.3.2 Reálně-intervalová data

Toleranční přístup pro reálně-intervalová data $(X, \boldsymbol{y})$ vychází opět ze stejné formulace vektoru parametrů pomocí středového vektoru, tolerančního vektoru a tolerančního koeficientu

$$\boldsymbol{a} = [a^c - \delta c, a^c + \delta c].$$

Středový vektor a^c je opět určen vhodnou reálnou aproximací, toleranční vektor c je určen podle stejných požadavků jako pro reálná data. Poslední neznámou částí potřebnou k vytvoření výsledného intervalového pásu je toleranční koeficient δ, který získáme z redukce problému na pomocný reálný problém s daty (X, y').

Vnější model

K nalezení vnějšího intervalového pásu vyřešíme pomocný problém (X, y') s danými a^c, c. Nalezením δ^* získáme parametr δ pro původní problém $(X, \boldsymbol{y})$. Pomocný reálný vektor y' vytvoříme tak, že vezmeme z každého $\boldsymbol{y}_j$ jeden krajní bod, a to ten vzdálenější od středové přímky Xa^c.

$$y'_j = \begin{cases} \underline{y}_j & \text{pokud } |\underline{y}_j - X_{j*}a| \geq |\overline{y}_j - X_{j*}a| \\ \overline{y}_j & \text{jinak} \end{cases}$$

Určením δ^* pro (X, y') podle věty 4.1 pro reálná data, získáme minimální δ pro původní data $(X,\boldsymbol{y})$. Tím získáme i nejužší možný pás obsahující všechna data určený hraničními přímkami $x\overline{a}$, $x\underline{a}$.

Vnitřní model

Pokud reálná regrese určená středovým parametrickým vektorem a^c protíná všechna vstupní data, můžeme určit i vnitřní model dat. Nyní však musíme zvolit nejmenší δ přes všechna pozorování $j \in 1, \ldots, p$, abychom zaručili, že každá přímka v nalezeném pásu protíná každý interval $\boldsymbol{y}_j$.

Opět provedeme redukci na reálná data (X, y'), nyní však z každého intervalu vezmeme jako y'_j vždy krajní bod, který je bližší ke středovému odhadu.

$$y'_j = \begin{cases} \underline{y}_j & \text{pokud } |\underline{y}_j - X_{j*}a| \leq |\overline{y}_j - X_{j*}a| \\ \overline{y}_j & \text{jinak} \end{cases}$$

Věta 4.2. Pokud $X_{j*}a \notin \boldsymbol{y}_j$ pro nějaké $j \in \{1, \ldots, p\}$, pak vnitřní model nemá řešení. Jinak

$$\delta^* := \min_{j:|X|_{j*}c>0} \frac{|y'_j - X_{j*}a^c|}{|X|_{j*}c}$$

je největší možná $\delta \geq 0$ splňující podmínku (silného) vnitřního modelu

$$\forall j = 1, \ldots, p : \ \boldsymbol{y}_j \supseteq X_{j*}a.$$

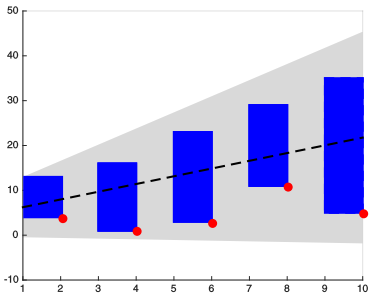
4.3.3 Intervalová data

Vnější model

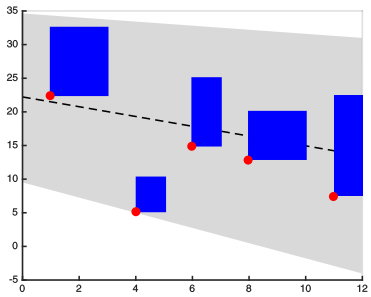
Máme-li intervalová data a chceme nalézt jejich silnou vnější obálku, použijeme stejně jako u reálně-intervalových dat redukci na reálný případ. Obálku našich vstupních dat $\boldsymbol{X} \in \mathbb{IR}^{m\times n}, \boldsymbol{y} \in \mathbb{IR}^m$ najdeme pomocí řešení dvou pomocných problémů s daty (X^d, y^d) a (X^h, y^h), která jsou pouze reálná. Opět budeme hledat co nejmenší koeficient δ, avšak zvlášť pro dolní a horní hranice intervalů $\boldsymbol{y}_j$.

Pro dané dva reálné problémy najdeme δ^*_d a δ^*_h. Výsledný hledaný parametrický vektor $\boldsymbol{a}$ určíme za pomoci většího z parametrů.

$$\delta^* = \max\{\delta^*_d, \delta^*_h\}$$
$$\boldsymbol{a} = [a^c - \delta^* c, a^c + \delta^* c]$$

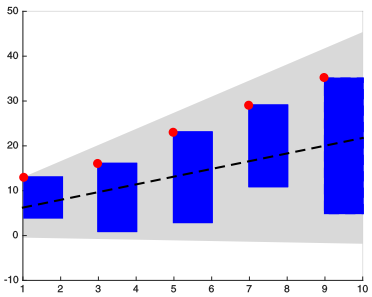


(a) Volba reálného x^d_{ji} pro dolní meze $\underline{y}_j$ pro rostoucí estimátor

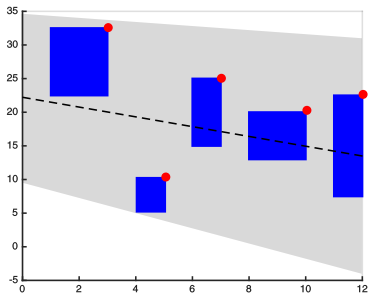


(b) Volba reálného x^d_{ji} pro dolní meze $\underline{y}_j$ pro klesající estimátor

Obrázek 4.5: První pomocný problém (X^d, y^d)



(a) Volba reálného x_{ji} pro horní meze $\overline{y}_j$ pro rostoucí estimátor



(b) Volba reálného x_{ji} pro horní meze $\overline{y}_j$ pro klesající estimátor

Obrázek 4.6: Druhý pomocný problém (X^h, y^h)

Redukci na pomocné problémy provedeme tak, že y^d_j, y^h_j zvolíme jako dolní, resp. horní hranici intervalu $\boldsymbol{y}_j$, reálné x^d_{ji}, x^h_{ji} bude zvoleno opět jako bod intervalového boxu, který je co nejvzdálenější od reálné středové aproximační přímky. Zjednodušeně řečeno, pokud v daném bodě středová přímka pásu roste či klesá, zvolíme vždy vzdálenější hranici intervalu $\boldsymbol{X}_{j*}$.

1. redukce

$$y^d_j = \underline{y}_j \quad x^d_{ji} = \begin{cases} \overline{x}_{ji} & a^c_i \geq 0 \\ \underline{x}_{ji} & \text{jinak} \end{cases}$$

2. redukce

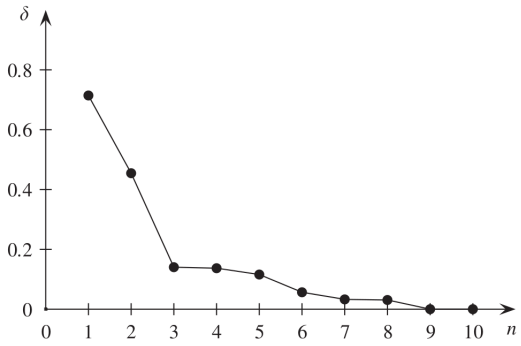
$$y^h_j = \overline{y}_j \quad x^h_{ji} = \begin{cases} \underline{x}_{ji} & a^c_i \geq 0 \\ \overline{x}_{ji} & \text{jinak} \end{cases}$$

Tímto jsme získali nejužší možnou dolní hraniční přímku určenou vektorem $a^d = a^c - \delta^*_d c$ a horní hranici určenou pomocí $a^h = a^c + \delta^*_h c$. Pro výsledný pás vybereme toleranční koeficient δ^* odpovídající většímu poloměru.

4.3.4 Chyby měření

Nyní si představíme ještě jedno využití interpretace tolerančního koeficientu. Jak pomocí něj odstranit hrubé náhodné chyby měření (anglicky outliers).

Toleranční koeficient má přímý vztah k poloměru nalezeného intervalového pásu při modelování dat. Pro každé pozorování, tedy každé $j = 1, \ldots, p$ nám



Obrázek 4.7: Sestupně setřízené toleranční koeficienty pro data tab. 3.1, obr. 3.3. Z grafu vidíme, že měření odpovídající prvním dvěma koeficientům jsou pravděpodobně chybná.

parametr δ určuje minimální poloměr pásu takový, aby hraniční přímka procházela bodem y_j, resp. jeho odpovídajícím nejvzdálenějším krajním bodem. Chybná měření lze poznat jako body, které tento poloměr nepřiměřeně rozšiřují.

Setřídíme-li všechny koeficienty $\delta_j = \frac{|y_j - X_{j*}a^c|}{|X_{j*}c|}$ spočítané podle vzorce ve větě 4.1 (o počítání delty), můžeme nahlédnout, zda jsou některé koeficienty výrazně větší než zbývající (viz obrázek 4.7). Měření odpovídající příliš velkému tolerančnímu koeficientu δ_j pak můžeme prohlásit za chyby měření a jim odpovídající pozorování odstranit ze vstupních dat. Následné přepočítání regresního modelu nám dá lepší výsledný intervalový pás.

4.3.5 Výhody postupu

Hlavní problémy, které toleranční přístup řeší lépe než například přístup pomocí lineárního programování, jsou:

- Dobrá centralita nalezeného řešení. Středová přímka nalezeného intervalového pásu dobře aproximuje data.
- Citlivost k chybám. Toleranční přístup umožňuje snadnou detekci hrubých chyb měření, jejichž odstraněním lze získat lepší regresní model.
- Vyváženost. Díky tolerančnímu vektoru lze zamezit nežádoucímu efektu, že některé parametry regresního modelu degenerují a jiné jsou na jejich úkor příliš široké.

Díky těmto výhodám se intervalová regrese zbavuje špatných vlastností, které ji podle Kim a kol. (1996) znevýhodňovaly vůči běžným statistickým metodám.

4.4 Odečítací přístup

Posledním přístupem, který je vhodný pro hledání vnějšího modelu vstupních dat, je odečítací metoda. Ta je založená na transformaci dat a jejich posunu odečtením reálného regresního estimátoru. Její hlavní výhodou je snadná úprava pro řešení nelineárních regresních úloh.

Postup:

1. Reálná lineární regrese

   Nalezení reálné lineární regresní křivky $y(x)$.

2. Transformace vstupních dat

   Posun dat, tj. odečtení reálné regresní křivky z 1. kroku od hodnot vstupních dat. V případě reálně-intervalových dat a intervalových dat dojde ještě k redukci vstupních dat na data reálná. Pro vstup ($\boldsymbol{X}$,$\boldsymbol{y}$) jsou vytvořena data $(\tilde{X}, \tilde{y})$.

3. Intervalový pás

   Nalezení hraničních přímek intervalového pásu zapouzdřující data $(\tilde{X}, \tilde{y})$. Ve dvou krocích se najde nejdřív horní hranice $\tilde{f}_h$ a následně dolní hranice $\tilde{f}_d$. Každá z hraničních přímek je hledána nezávisle na druhé. Výsledkem je tedy intervalový pás $\tilde{P}(x)$ definovaný funkcemi $\tilde{f}_d, \tilde{f}_h$.

4. Transformace řešení podproblému na řešení původní úlohy

   Posun intervalového pásu z 3. kroku přičtením reálné regresní křivky z 1. kroku. Je tedy nalezen pás $P(x)$ zapouzdřující vstupní data ($\boldsymbol{X}$,$\boldsymbol{y}$), který je definovaný funkcemi $f_d = \tilde{f}_d + y$, $f_h = \tilde{f}_h + y$.

Pokud jsou vstupními daty algoritmu přímo reálná data, můžeme postupovat přesně podle uvedeného postupu. Pro intervalově-reálný vstup proběhne v 2. kroku transformace dat i jejich redukce na reálná data. Vstupní data podproblému jsou určena jako $(\tilde{X}, \tilde{y})$, kde

$$\tilde{X} := \begin{pmatrix} X \\ X \end{pmatrix},$$

$$y_{posun} := \boldsymbol{y} - y(X),$$

$$\tilde{\boldsymbol{y}} := \begin{pmatrix} \overline{y}_{posun} \\ \underline{y}_{posun} \end{pmatrix}.$$

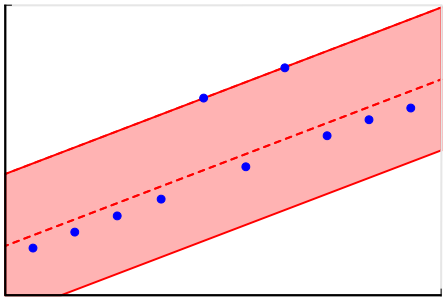
U obecných intervalových dat je nutné zahrnout do $(\tilde{X}, \tilde{y})$ všechny vrcholy vstupních boxů.

Výhody a nevýhody

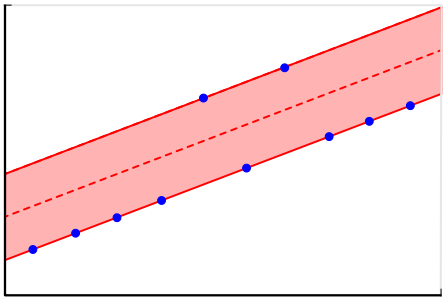
Jako výhody tohoto postupu můžeme uvést:

- V řadě případů můžeme získat užší pás než použitím tolerančního přístupu. A to díky nevyžadování symetrie pásu podle centrální nadroviny. Toto ilustruje obrázek 4.8.
- Snadná rozšiřitelnost postupu na nelineární regresi.
- Možnost využít k řešení reálné regrese nebo intervalové. V kroku tři lze využít přímo metod pro nalezení intervalové lineární regrese.
- Zjednodušení intervalových problémů na problém s reálnými daty.

Obrázek 4.8: Centralita na úkor šířky pásu



(a) Odečítání estimátoru s příliš širokým pásem (dobrá centralita na úkor šířky)



(b) Odečtení estimátoru (užší pás splňující požadavky vnějšího intervalového modelu)

Nevýhodou je horší centralita nalezeného řešení. Tedy že středová nadrovina nemusí příliš dobře popisovat data.

Pokud však chceme tuto vlastnost zachovat, můžeme v kroku, kde dochází k transformaci dat, po posunutí vzít absolutní hodnotu daného čísla, pak nalézt pouze horní hraniční přímku a jako dolní hraniční přímku použít přímku zrcadlově otočenou kolem osy x. Pokud bychom zachovali pravidla určování hranice využívané u tolerančního přístupu, pak dávají tyto postupy stejný výsledek.

5. Nelineární intervalová estimace

Předchozí část práce popisovala pouze lineární přístupy k intervalové estimaci, následující část se bude věnovat případům, kdy pracujeme s daty, u kterých předpokládáme jinou než lineární závislost a chceme je proložit jinou křivkou než přímkou (resp. nadrovinou). V první části kapitoly uvedeme způsoby řešení pomocí linearizace funkce, kterou chceme data prokládat, v druhé části pak zobecníme odečítací metodu použitou pro lineární estimaci pro nelineární data a na závěr teoretické části práce zmíníme iterativní toleranční metodu.

5.1 Linearizace

Řadu funkcí, které mohou vyjadřovat závislost dat, můžeme převést pomocí transformace parametrů na problém, který lze řešit řešením lineární soustavy rovnic. Tento způsob je velice užitečný a může využít libovolných existujících řešičů intervalových soustav. Bohužel ne všechny funkce lze linearizovat, takže použití této metody je omezené.

Příklady jednotlivých funkcí, včetně převodu na lineární formu, jsou v souhrnné tabulce 5.1, sestavené podle Sit a kol. (1994). Pro danou funkci vytvoříme lineární tvar, který je řešitelný lineární soustavou a následně převedeme získané parametry pomocné funkce na parametry původní nelineární funkce.

Funkce	Lineární forma	Parametry
$Y = a + bX$	—	—
$Y = a + bX + cX^2$		
$Y = a + bX + cX^2 + dX^3$		
$Y = \frac{X}{a+bX}$	$\frac{1}{Y} = \frac{A}{X} + B$	$a = A, b = B$
$Y = \frac{X}{a+bX+cX^2}$	$\frac{1}{Y} = \frac{A}{X} + B + CX$	$a = A, b = B, c = C$
$Y = \frac{X}{a+bX+cX^2+dX^3}$	$\frac{1}{Y} = \frac{A}{X} + B + CX + DX^2$	$a = A, b = B, c = C, d = D$
$Y = \frac{a}{X} + bX + c$	$Y = A + BX + CX^{-1}$	$a = C, b = B, c = A$
$Y = ae^{bX}$	$\ln(Y) = A + BX$	$a = e^A, b = B$
$Y = e^{a-bX}$	$\ln(Y) = A + BX$	$a = A, b = -B$
$Y = ae^{b/X}$	$\ln(Y) = A + \frac{B}{X}$	$a = e^A, b = B$
$Y = ab^X - ae^{X\ln(b)}$	$\ln(Y) = A + BX$	$a = e^A, b = e^B$
$Y = ab^{(X-c)^2}$	$\ln(Y) = A + BX + CX^2$	$a = e^{A-\frac{B^2}{4C}}, b = e^C, c = \frac{-B}{2C}$
$Y = e^{(a+b/X)}$	$\ln(Y) = A + \frac{B}{X}$	$a = A, b = B$
$Y = 1 - e^{-aX^b}$	$\ln\left[\frac{-1}{a}\ln(Y)\right] = BX$	$b = B$
$Y = a(1 - e^{-bX})^c$	$\ln\left[1 - (\frac{Y}{a})^{1/c}\right] = BX$	$b = -B$
$Y = \frac{a}{d+e^{b-cX}}$	$\ln(\frac{a}{Y} - d) = B + CX$	$b = B, c = -C$
$Y = \frac{a}{1+e^{b-cX}}$	$\ln(\frac{a}{Y} - 1) = B + CX$	$b = B, c = -C$
$Y = ae^{-e^{b-cX}}$	$\ln\left[-\ln(\frac{Y}{a})\right] = B + CX$	$b = B, c = -C$
$Y = aX^b$	$\ln(Y) = A + B\ln(X)$	$a = e^A, b = B$
$Y = aX^b c^X$	$\ln(Y) = A + B\ln(X) + CX$	$a = e^A, b = B, c = e^C$
$Y = aX^b e^{cX}$	$\ln(Y) = A + B\ln(X) + CX$	$a = e^A, b = B, c = C$
$Y = a + b\ln(X)$	—	—

Tabulka 5.1: Linearizace funkcí

Pro ilustraci rozebereme pár ukázkových převodů podrobněji.

Exponenciální model

Máme-li data, která chceme aproximovat exponenciální funkcí, hledáme popis exponenciální křivky

$$y = a \cdot b^x,$$

kde x, y jsou vstupní data a a, b jsou hledané parametry určující tvar exponenciály. Pro převod problému na lineární regresi zlogaritmujeme danou rovnici a získáme

$$\ln(y) = \ln(ab^x),$$

$$\ln(y) = \ln(a) + x \ln(b).$$

Přeznačením jednotlivých parametrů dostaneme

$$y' = A + Bx.$$

Nyní stačí nalézt pro data x, $y' = \ln(y)$ přímku $y' = A + Bx$ prokládající již lineární model a z jejích koeficientů dopočítáme koeficienty exponenciály aproximující původní data x, y.

$$a = e^A$$

$$b = e^B$$

Mocninný model

Rovnice křivky mocninné funkce je $y = a \cdot x^b$. Zopakujeme trik použití zlogaritmování

$$\ln(y) = \ln(ax^b)$$

$$\ln(y) = \ln(a) + b \ln(x)$$

$$y' = A + B \cdot x'.$$

Dostaneme problém lineární, avšak s jednou komplikací navíc. Oproti exponenciálním datům musíme zlogaritmovat nejen hodnoty y ($y' := \ln(y)$), ale i vstupní hodnoty x ($x' = \ln(x)$). Zjednodušilo se nám však dopočítávání koeficientů. Po nalezení přímky $y' = A + B \cdot x'$ aproximující zlinearizovaná data nám stačí pro určení mocninné funkce aproximující původní data pouze spočítat parametr a, parametr b přepočítávat nemusíme.

$$a = e^A$$

$$b = B$$

5.2 Odečítací přístup

Jak se vypořádat s nelineárními daty, která chceme proložit intervalovou křivkou, a zároveň hledaný model nemůžeme linearizovat, budeme zkoumat v této podkapitole. Metoda využívá hledání lineární intervalové regrese.

Předpokládejme, že máme vstupní data $(X, \boldsymbol{y})$, $X \in \mathbb{R}^{p \times n}, y \in \mathbb{IR}^p$. Potřebujeme najít vnější model dat. Tedy hledáme intervalový pás $P(x)$, který obsahuje vstupní data. Předpokládejme, že známe reálnou funkci $f : \mathbb{R}^n \to \mathbb{R}$, která popisuje trend chování vstupních dat. Tuto funkci lze získat běžnými (reálnými) statistickými postupy nebo třeba pomocí numerické interpolace dat.

Nalezení intervalového pásu $P(x)$:

1. Odečtení funkce f od vstupních dat. Tím vytvoříme podproblém s daty $(X, \tilde{\boldsymbol{y}})$.
$$\forall j \in \{1, \ldots, p\} : \tilde{\boldsymbol{y}}_j = \boldsymbol{y}_j - f(X_{j*})$$

2. Vyřešíme estimaci dat $(X, \tilde{\boldsymbol{y}})$ pomocí lineární estimace. Získáme tedy lineární intervalový pás s horní hraniční přímkou $\tilde{f}_h$ a dolní hraniční přímkou $\tilde{f}_d$.

3. Určíme $P(x)$ jako intervalový pás s hraničními křivkami
$$f_h = f + \tilde{f}_h,$$
$$f_d = f + \tilde{f}_d.$$

Neurčíme-li každou hraniční přímku zvlášť, ale určíme je například pomocí tolerančního přístupu, získáme tímto postupem intervalovou křivku jejíž obě hraniční křivky jsou popsány funkcí stejného tvaru, lišící se pouze svými parametry. Křivka je lineární kombinací původní funkce s lineární funkcí. Nalezený vektor parametrů zajišťuje, že intervalový pás je „souměrný“ podle funkce $f(x)$, v každém bodě X_{j*} je $f(X_{j*})$ středem intervalu $P(X_{j*})$. Díky tomu má intervalový pás $P(x)$ dobrou centralitu (středová křivka dobře popisuje chování dat). Avšak šířka nalezeného pásu může být zbytečně velká. Tuto nevýhodu postupu ilustrují obrázky 4.8a, 4.8b.

Další variantou řešení je hledání jednotlivých posunů $f(x)$ (nahoru a dolů) zvlášť. V 2. kroku chceme najít zvlášť horní hraniční přímku a zvlášť dolní hraniční přímku. To lze řešit například pomocí lineárního programování. Nechť $(X,\boldsymbol{y})$ jsou vstupní data a $(X, \tilde{\boldsymbol{y}})$ jsou vstupní data posunutá o funkci $f(x)$. Pak přímka $\tilde{f}_d(x) = aX$ je řešením
$$\min_a \sum_j |X_{j*}a - \underline{\tilde{y}}_j|$$
$$\underline{\tilde{y}}_j \geq X_{j*}a$$
a určuje spodní hranici intervalového pásu $P(x)$. Horní hranice je pak určena přímkou $\tilde{f}_h(x) = aX$ pro
$$\min_a \sum_j |X_{j*}a - \overline{\tilde{y}}_j|$$
$$\overline{\tilde{y}}_j \leq X_{j*}a.$$

Pokud bychom chtěli postup aplikovat na intervalová data, problém se zkomplikuje. Pokud funkce určující středovou křivku intervalového pásu není monotónní ve svých proměnných, je komplikované nalézt obraz intervalu $\boldsymbol{X}_{j*}$ a tedy určit horní, resp. dolní hraniční funkci.

5.3 Toleranční přístup

Pro některé třídy funkcí lze použít toleranční přístup publikovaný v Hladík a Černý (2012b). Jde o iterativní metodu hledající vnější model. Pro nalezení aproximační parametrické funkce s vektorem parametrů $\boldsymbol{a}$ v proměnné x se postupuje následujícím způsobem:

1. Nalezne se reálná aproximace pomocí tradičních reálných nelineárních regresních postupů, čím je určen středový vektor parametrů a^c. Je dán toleranční vektor c (běžně jako $c = |a^c|$ či $c = (1, \ldots, 1)$.

2. Je nalezen koeficient δ. Nejdříve je inicializován $\delta = 1$ a následně se pro zvolený počet iterací opakuje: Pro každé $j = 1, \ldots, p$, pokud $\boldsymbol{y}_j$ je uvnitř obrazu intervalu $\boldsymbol{x}_j$, zmenši δ, jinak zvětši.

Pokud je δ zvětšována či zmenšována půlením, pak metoda konverguje exponenciálně rychle k optimu. Pro praktické využití proto stačí použít přibližně 10 iterací pro nalezení aproximace dat. V jistých případech algoritmus nekonverguje k řešení, příkladem může být Gompertzova křivka

$$f(x) = e^{-e^{x-a_0}},\ x \in \mathbb{R} \cup \{-\infty, \infty\},$$

pokud vstupní data obsahují bod $x = 0$, $y = 1$.

Metoda je vhodná pro funkce, které jsou analyticky vyjádřitelné jako výrazy složené z operací $+, -, \times, \div$ a základních (spojitých reálných) funkcí, které lze vyhodnotit na intervalu. Navíc jsou monotónní v každém svém parametru, případně i v každé proměnné a každý parametr, případně proměnná, se ve výrazu objeví nejvýše jednou.

6. Aplikace metod

6.1 Použitý software

V roce 2015 vyšel standard intervalové aritmetiky[1], který definuje, jak přesně má být korektně implementována intervalová aritmetika v počítačích. Tento standard je v současné době implementován a bezplatně přístupný v knihovně pro C++ a v intervalovém balíku pro GNU Octave.

GNU Octave je prostředí s jazykem zaměřující se na matematická využití. Je to multiplatformní free software s možností vizualizace dat. Další výhodou je částečná kompatibilita zdrojových kódů s komerčním systémem Matlab. Původně začaly metody vznikat jako kód právě pro Matlab.

Naimplementované metody jsou součástí intervalového balíku LIME pro GNU Octave, který je vyvíjen studenty Matematicko-fyzikální fakulty Univerzity Karlovy. Původně vznikal pro použití v Matlabu a využíval balíky Intlab a Versoft. Nyní je však primárně určen pro GNU Octave a využívá jeho balík pro intervalové počítání.

Při vytváření této práce bylo použito GNU Octave verze 4.0.0 a intervalový balík verze 2.1.0.

Některé další dostupné softwarové možnosti pro intervalové počítání zmiňuje Horáček (2011).

6.2 Implementace

Název	Vzorec
'lin'	$Y = a + bX$
'quad'	$Y = a + bX + cX^2$
'cub'	$Y = a + bX + cX^2 + dX^3$
'exptype1'	$Y = ae^{bX}$
'exptype2'	$Y = e^{a-bX}$
'exptype3'	$Y = ae^{b/X}$
'exptype4'	$Y = ab^X$
'pow'	$Y = aX^b$
'exppowtype1'	$Y = aX^b c^X$
'exppowtype2'	$Y = aX^b e^{cX}$
'log'	$Y = a + b\ln(X)$
'loglin'	$Y = a + bX + c\ln(X)$
'explin'	$Y = a + bX + ce^X$
'eplintype2'	$Y = a + bX + ce^{(-X)}$

Tabulka 6.1: Flagy pro jednotlivé typy funkcí

Součástí elektronické přílohy této práce je implementace toleranční metody pro lineární regresi a odečítacího přístupu pro nelineární estimaci. Dále jsou zde pomocné funkce, několik demonstračních skriptů a vstupní data ukázkových příkladů. Ve složce `doc` se nachází html dokumentace v níž jsou popsány jednotlivé

[1] The Institute of Electrical and Electronics Engineers (2015)

funkce. Je zde uživatelská dokumentace popisující vstupní a výstupní parametry funkcí a programátorská dokumentace, která obsahuje poznámky k vlastní implementaci.

Funkce hledající vnější model podle tolerančního přístupu (funkce `linregtol`) je naprogramována podle článku Hladík a Černý (2012a). Výstupem metody jsou intervalové parametry křivky modelující data.

Odečítací funkce využívá k vyřešení lineárního problému toleranční metodu. Funkce `estimationoutersep` hledá intervalový pás pomocí odečítacího postupu popsaného v kapitole 5.2, který určí každou hraniční funkci pásu zvlášť. Pás symetrický podle reálného estimátoru hledá funkce `estimationouter`.

Modely, které lze pomocí naprogramovaných funkcí hledat, jsou v tabulce 6.1.

6.3 Lineární intervalová estimace

Pro ilustraci lineární estimace použijeme příklad s reálnými vstupními daty (X, y). Hodnoty nezávislé proměnné X jsou teplota p-Xylenu, který se využívá k výrobě polymerů, v kelvinech. Měřené hodnoty y jsou pak rychlost šíření ultrazvukových vln v metrech za sekundu v této kapalině. Vstupní data (Bank, Dortmund Data, 2017) měření můžeme vidět v tabulce 6.2.

Předpokládáme lineární závislost rychlosti šíření vln na teplotě. Pro estimaci dat jsme použili lineární toleranční přístup (obr. 6.1a) a odečítací metodu (obr. 6.1b). Toleranční přístup nejdříve odhadne pomocí regresní metody nejmenších čtverců středovou přímku (znázorněnou čárkovaně)

$$y = 2304.071 - 3.3208x,$$

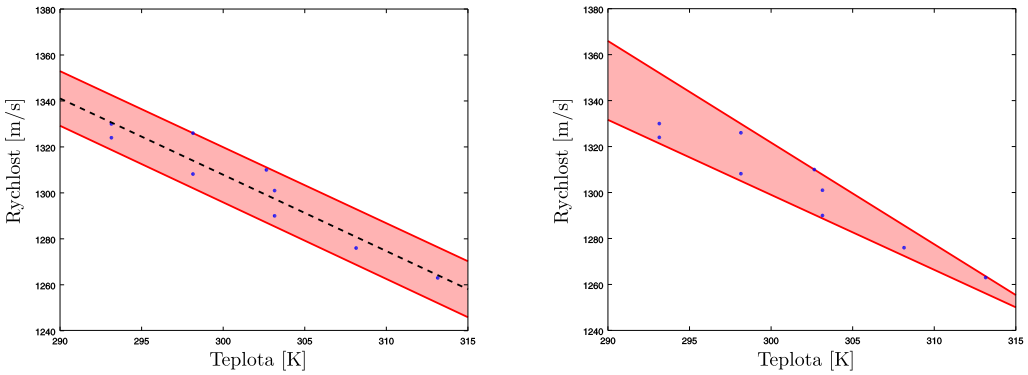
jejíž parametry jsou středovým vektorem parametrického vektoru $\boldsymbol{a}$, tedy $a^c = (2304.071, -3.3208)^\top$. Použili jsme zde relativní toleranci, kde toleranční vektor je absolutní hodnotou středového vektoru $c = (2304.071, 3.3208)^\top$. Po spočítání všech tolerančních koeficientů δ byl určen poloměr vektoru parametrů a výsledný intervalový pás je definovaný intervalovou přímkou

$$\boldsymbol{y} = [2295.6, 2312.5] + [-3.3328, -3.3085]x.$$

Další použitou metodou byl odečítací přístup. Ten využívá stejné středové přímky y. Následně od vstupních dat odečte tento středový estimátor a vytvoří posunutá data $y'_j = y_j - y(x_j)$. Pro tyto data určí samostatně (pomocí tolerančního přístupu) horní a dolní hraniční funkci. Výsledný pás je pak ohraničen funkcemi

$$f_h = y + 344.47 - 1.102x,$$
$$f_d = y - 26.023 + 0.057243x.$$

Na obrázku 6.1 vidíme obě použité metody. Oba přístupy nám dávají verifikovaný pás, který zapouzdřuje všechna vstupní data. U pásu určeného toleranční metodou 6.1a vidíme, že centrální přímka pásu dobře popisuje data. Důraz na centralitu a z ní vyplývající symetrie však způsobují, že intervalový pás není nejtěsnější možný. Na obrázku 6.1b vidíme použití odečítací metody, která také charakterizuje trend dat. Hlavní zřetelnou výhodou je, že lépe zohledňuje, jak se chovají horní a dolní okraje dat.



(a) Toleranční přístup (b) Odečítací přístup

Obrázek 6.1: Závislost rychlosti šíření ultrazvukového vlnění na teplotě kapaliny

x_j	293.15	293.15	298.15	298.15	302.65	303.15	303.15	308.15	313.15
y_j	1324.0	1330.0	1308.2	1326.0	1310.0	1290.0	1301.0	1276.0	1263.0

Tabulka 6.2: Vstupní data pro lineární estimaci

6.4 Nelineární intervalová estimace

Příklad nelineární estimace ilustrujeme na intervalově-reálných datech pocházejících z diagnostiky plicních funkcí – konkrétně na příkladu vyplavování dusíku z plic čistým kyslíkem. Anonymizovaná data byla naměřena a předzpracována v rámci grantu GAUK č. 174815, na kterém spolupracujeme s 2. lékařskou fakultou UK. Námi vybraný datový soubor obsahuje koncentrace dusíku (v procentech) měřené na konci jednotlivých dechů. Intervaly na ose y jsou určeny podle chyb jednotlivých měřících senzorů, ze kterých se počítá výsledná koncentrace dusíku.

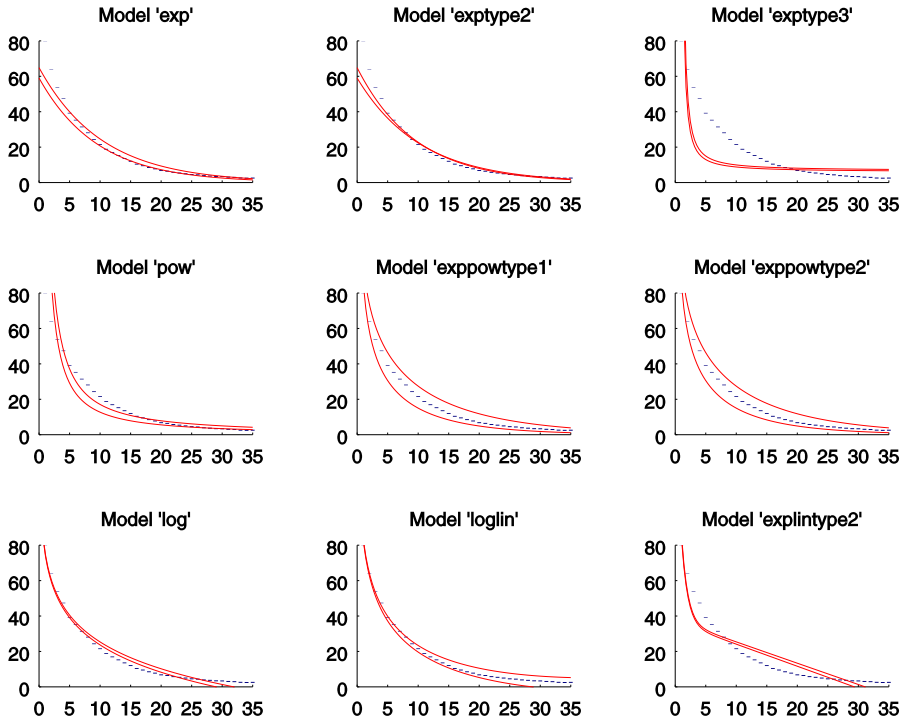
Na vstupní data aplikujeme intervalovou metodu nejmenších čtverců, toleranční přístup a odečítací přístup.

6.4.1 Metoda nejmenších čtverců

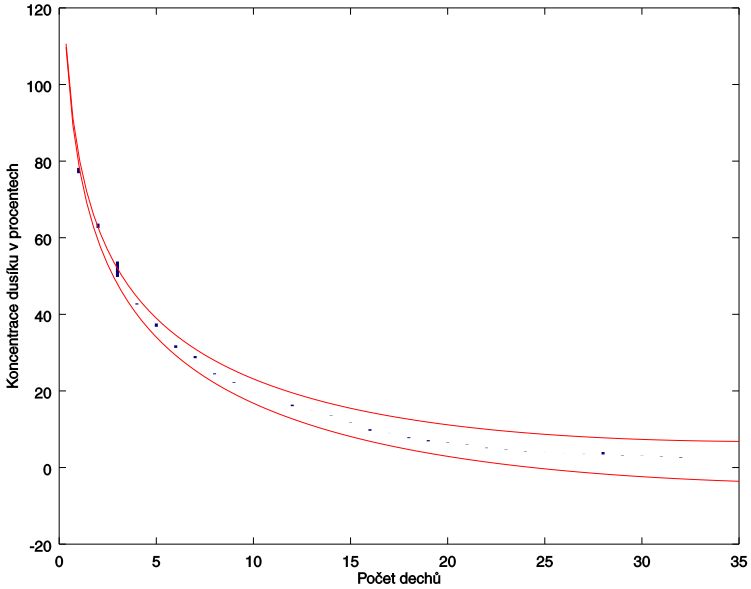
Pokud nevíme nic o chybách vstupních dat a nemůžeme předpokládat nějaké rozdělení chyby na daných intervalech, nabízí se použít intervalovou metodu nejmenších čtverců. Na obrázku 6.2 jsou znázorněny výsledky této metody s použitím různých typů modelů. Znázorněný intervalový pás vždy zapouzdřuje všechny křivky, které vzniknou použitím reálné metody nejmenších čtverců pro libovolnou realizaci dat.

Z přehledového obrázku můžeme vidět, že některé použité modely vstupním datům vůbec neodpovídají ('exptype3', 'exptype4'). Další modely založené na exponenciální křivce špatně kopírují začátek dat, i když pozvolný konec aproximují vcelku dobře ('exp', 'exptype2'). Mocninná funkce hůře popisuje střední část dat ('pow', 'exppowtype1', 'exppowtype2'). Křivky založené na logaritmické funkci dobře kopírují začátek dat ('log', 'loglin'). Dobrým způsobem popisuje celá data funkce, která je součtem logaritmické křivky s lineární, 'loglin' model. Pás se však na konci rozevírá, což může být nevhodné pro predikci dat.

Detailní obrázek modelu 'loglin' 6.3 nám ukazuje, že vstupní data nemusí být zahrnuta vždy uvnitř pásu určeného metodou nejmenších čtverců.



Obrázek 6.2: Intervalová metoda nejmenších čtverců



Obrázek 6.3: Intervalová metoda nejmenších čtverců, model 'loglin'

6.4.2 Toleranční metoda a odečítací metoda

Chceme-li získat verifikovaný pás uzavírající data, není metoda nejmenších čtverců vyhovující a je třeba použít jiný přístup k intervalové estimaci. Vnější pásy zapouzdřující všechny intervaly vidíme na obrázcích 6.4 a 6.5. V obou případech je použit model 'loglin'.

Obě metody využívají jako inicializační řešení reálnou estimaci pomocí metody nejmenších čtverců aplikovanou na středy intervalů, ta prokládá data reálnou funkcí

$$y(x) = 79.207 + 0.6864 \cdot x - 28.612 \cdot \log(x).$$

Analytickou linearizací modelující funkce a použitím lineárního tolerančního přístupu s relativní tolerancí (toleranční vektor je volen jako absolutní hodnota vektoru parametrů funkce $y(x)$) získáme intervalový pás popsaný funkcí

$$y = [76.164, 83.227] + [0.69139, 0.75551]x + [-30.373, -27.795]\log(x).$$

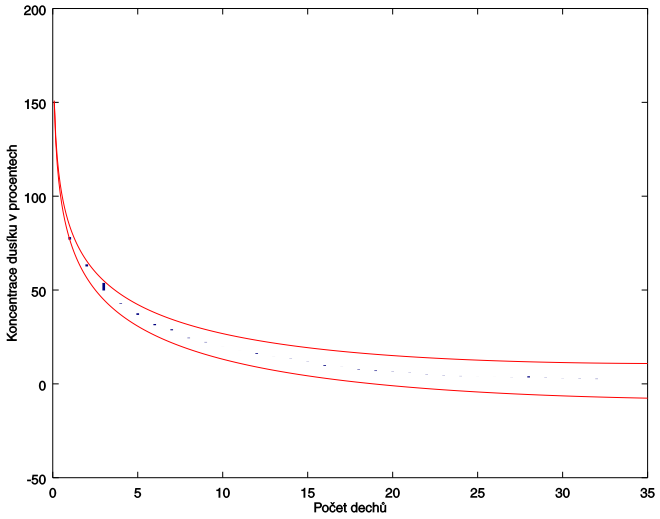
Tento pás vidíme na obrázku 6.4. Nevýhodou tohoto postupu linearizace a zpětného převodu je, že funkce určená středovým vektorem parametrů nyní neleží ve středu intervalového pásu. Z obrázku také můžeme vidět, že intervalová křivka se na konci rozevírá, což nemusí být žádoucí. Proto může být vhodnější použít odečítací metodu.

Na vstupní data jsme aplikovali nejdříve posun odečtením reálné estimační funkce $y(x)$, a poté jsme použili lineární toleranční přístup k určení přímek zapouzdřující posunutá data (s absolutní tolerancí povolující jen posun). Každá mezní funkce byla určena samostatně. Tím jsme získali intervalový pás, jehož horní a dolní hraniční funkce jsou

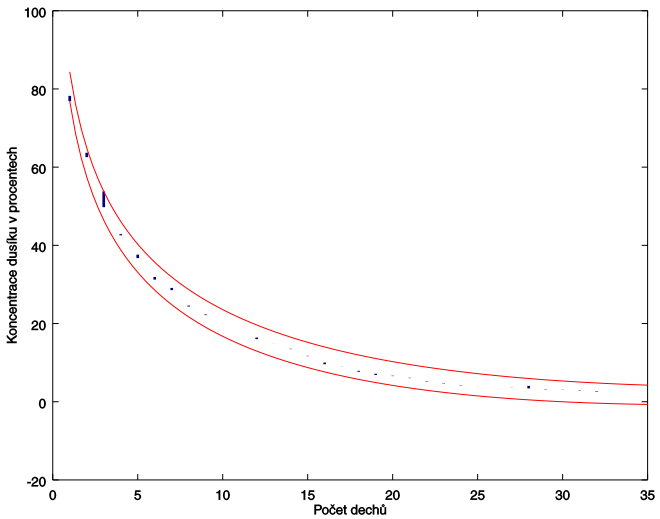
$$f_h(x) = y(x) + 3.9928 - 0.038873 \cdot x,$$

$$f_d(x) = y(x) - 3.6005 + 0.037183 \cdot x.$$

Vykreslené funkce vidíme na obrázku 6.5. Protože výsledný pás vznikl přičtením klesající přímky (u horní hranice) a rostoucí (u dolní hranice) vidíme, že intervalový pás se na konci zužuje. Zároveň stále velice úzce kopíruje začátek dat.



Obrázek 6.4: Toleranční metoda pro nelinární data, model 'loglin'



Obrázek 6.5: Odečítací metoda pro nelinární data, model 'loglin'

Závěr

Práce podává ucelený souhrn přístupů k intervalové estimaci. Nejprve jsme zmínili základní pojmy intervalové analýzy a zavedli jsme pojem intervalového pásu.

Popsali jsme koncept reálné regrese pro různé typy vstupních dat (reálná, reálně-intervalová, intervalová). Dokázali jsme ekvivalenci regresní metody nejmenších čtverců aplikované na středy intervalově-reálných dat s aplikací na krajní body intervalů.

Systematizovali jsme přístupy k intervalové lineární estimaci. Popsali jsme vztahy mezi řešením intervalové lineární soustavy metodou nejmenších čtverců, vnějším modelem (possibilistic model) a vnitřním modelem (necessity model). Porovnali jsme formulace vnějšího a vnitřního konceptu na základě typu vstupních dat a diskutovali jsme jejich silné a slabé formulace. Silná formulace vnějšího modelu je v praxi užitečná, protože dává verifikovaný intervalový pás obsahující vstupní data. Slabé varianty formulací vnějšího a vnitřního konceptu mají spíše teoretický význam.

V rámci kapitoly o postupech lineární estimace jsme popsali hledání modelů dat pomocí lineárního programování, kvadratického programování, tolerančního přístupu a představili jsme odečítací přístup. Lineární programy pro určení vnějšího a vnitřního modelu nemusí vždy splňovat podmínku inkluze, že řešení vnitřního modelu je podmnožinou řešení vnějšího. Tento problém řeší formulace společného kvadratického programu hledající oba modely zároveň. S kvadratickým programováním lze také dát větší důraz na centralitu řešení, aby středová přímka nalezeného intervalového pásu dobře aproximovala data ve smyslu metody nejmenších čtverců. Toleranční přístup vychází z reálné estimace a proto vykazuje dobrou centralitu. Může však kvůli vynucené symetrii podle středového estimátoru najít příliš široký pás. Tuto nevýhodu může odstranit odečítací metoda, jejíž výhodou je také snadné rozšíření pro nelineární modely dat.

Shrnuli jsme možnosti nelineární estimace za pomoci linearizace modelu, použití odečítacího nebo iteračního tolerančního přístupu.

Poslední část práce ukazuje aplikaci lineární i nelineární intervalové estimace na reálných příkladech vstupních dat. Použili jsme metodu nejmenších čtverců, toleranční metodu a odečítací metodu. Nevýhodou intervalové metody nejmenších čtverců je, že nenajde verifikovaný pás zapouzdřující vstupní data. Metoda je užitečná při hledání modelu intervalových dat. Verifikovaný intervalový pás lze pak nalézt pomocí tolerančního přístupu. Odečítací metoda oproti tolerančnímu přístupu lépe zohledňuje chování krajů vstupních dat.

Seznam použité literatury

Anděl, J. (2007). *Statistické metody*. Čtvrté přepracované vydání. Matfyzpress, Praha. ISBN 80-7378-003-8.

Bank, Dortmund Data (2017). version 2017. *DDBST Software and Separation Technology GmbH: Oldenburg, Germany*.

Hladík, M. a Černý, M. (2010). New approach to interval linear regression. *International Conference 24th Mini EURO Conference „Continuous Optimization and Information-Based Technologies in the Financial Sector“*.

Hladík, M. a Černý, M. (2012a). Interval regression by tolerance analysis approach. *Fuzzy Sets and Systems*, **193**, 85–107.

Hladík, M. a Černý, M. (2012b). On the tolerance approach to possibilistic nonlinear regression over interval data. *5th International Conference on Reliable Engineering Computing*, pages 183–195.

Horáček, J. (2011). Přeurčené soustavy intervalových lineárních rovnic. Diplomová práce, Matematicko-fyzikální fakulta University Karlovy v Praze.

Ishibuchi, H. (1993). A unified approach to possibility and necessity regression analysis with interval regression models. In *Proc. Of the Fifth IFSA World Congress, Seoul, Korea, 1993*.

Kim, K. J., Moskowitz, H. a Koksalan, M. (1996). Fuzzy versus statistical linear regression. *European Journal of Operational Research*, **92**(2), 417 – 434. ISSN 0377-2217. doi: http://dx.doi.org/10.1016/0377-2217(94)00352-1. URL `http://www.sciencedirect.com/science/article/pii/0377221794003521`.

Moore, R. E., Kearfott, R. B. a Cloud, M. J. (2009). *Introduction to interval analysis*. SIAM.

Neumaier, A. (1986). Linear interval equations. *Interval Mathematics 1985*, pages 109–120.

Peters, G. (1994). Fuzzy linear regression with fuzzy intervals. *Fuzzy Sets and Systems*, **63**(1), 45 – 55. ISSN 0165-0114. doi: http://dx.doi.org/10.1016/0165-0114(94)90144-9. URL `http://www.sciencedirect.com/science/article/pii/0165011494901449`.

Sit, V., Poulin-Costello, M. a Bergerud, W. (1994). *Catalogue of curves for curve fitting*. Forest Science Research Branch, Ministry of Forests.

Tanaka, H. (1987). Fuzzy data analysis by possibilistic linear models. *Fuzzy sets and systems*, **24**(3), 363–375.

Tanaka, H. a Lee, H. (1998). Interval regression analysis by quadratic programming approach. *Fuzzy Systems, IEEE Transactions on*, **6**(4), 473–481.

The Institute of Electrical and Electronics Engineers (2015). *IEEE Standard for Interval Arithmetic*. IEEE Computer Society.

Černý, M., Antoch, J. a Hladík, M. (2013). On the possibilistic approach to linear regression models involving uncertain, indeterminate or interval data. *Information Sciences*, **244**, 26–47.

Seznam obrázků

Značení

$\mathbb{R}$	množina reálných čísel
$\mathbb{R}^*$	$\mathbb{R} \cup \{-\infty, \infty\}$
$\mathbb{IR}$	množina všech reálných intervalů
$\boldsymbol{A}$, $\boldsymbol{a}$	intervalová matice, intervalový vektor
$\underline{a}$, $\inf(\boldsymbol{a})$	dolní mez intervalu $\boldsymbol{a}$
$\overline{a}$, $\sup(\boldsymbol{a})$	horní mez intervalu $\boldsymbol{a}$
a^c	střed intervalu
a^Δ	poloměr intervalu
x_j	$x_j = X_{j*} = (x_{j1}, \ldots, x_{jn})$
$(\boldsymbol{X}, \boldsymbol{y})$	vstupní data
I_n	jednotková matice řádu $n \times n$
LP	lineární program
QP	kvadratický program