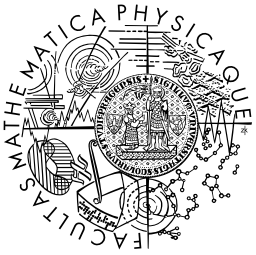


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Adéla Zavřelová

Konfidenční pásy pro regresní křivky

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Zdeněk Hlávka, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2017

Prohlašuji, že jsem tuto bakalářskou práci vypracovala 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

Název práce: Konfidenční pásy pro regresní křivky

Autor: Adéla Zavřelová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zdeněk Hlávka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá sestrojením pásu spolehlivosti pro lineární regresní model. Jsou zde předloženy základní vlastnosti lineárního modelu a popsána konstrukce různých pásů spolehlivosti pro modely, kde je závislost určena funkcí jedné proměnné. Zvláště je popsána konstrukce pásů pro polynomiální model.

Klíčová slova: lineární regresní model, polynomiální regrese, pás spolehlivosti.

Title: Confidence bands for regression curves

Author: Adéla Zavřelová

Department: Department of Probability and Mathematical Statistic

Supervisor: doc. RNDr. Zdeněk Hlávka, Ph.D., Department of Probability and Mathematical Statistic

Abstract: This thesis deals with the constructions of the confidence band for a linear regression model. Basic characteristics of a linear model are given and constructions of different confidence bands are described for models, where the relationship is set by a one variable function. The main focus is on bands of polynomial models.

Keywords: linear regression model, polynomial regression, confidence band.

Na tomto místě bych ráda poděkovala panu doc. RNDr. Zdeňku Hlávkovi, Ph.D., za vedení, konzultace a cennou pomoc při psaní mé bakalářské práce.

Obsah

Úvod

V této práci se budeme zabývat různými metodami sestrojení pásu spolehlivosti pro regresní model. V první kapitole zavedeme značení a seznámíme se se základními vlastnostmi lineárního modelu.

V první části druhé kapitoly uvedeme interval spolehlivosti, který dodržuje pravděpodobnost pokrytí střední hodnoty regresního modelu v jednom daném bodě. V druhé části sestrojíme predikční interval pro budoucí pozorování v pevném bodě. Ve třetí části předvedeme konstrukci pásu spolehlivosti, který pokrývá střední hodnotu regresního modelu pro všechny body daného intervalu zároveň. Všechny tyto postupy vycházejí z knihy Zvára (2008). Získaný pás spolehlivosti je totožný s pásem sestrojeným Scheffého metodou, kterou popsal Naiman (1986). Tento postup konstrukce pásu je velmi rozšířený a často používaný. V poslední části druhé kapitoly předvedeme uvedené postupy na příkladu polynomiálního modelu.

Ve třetí kapitole popíšeme postup sestrojení přesného pásu spolehlivost pro polynomiální regresní model založený na textu Wynn (1984). Výsledný pás bude po částech polynomiální stejného stupně. Předvedeme postup na příkladu a porovnáme získaný pás s pásem sestojeným podle druhé kapitoly.

Jedná se stále o aktivní téma, například srovnáním pásů spolehlivosti pro polynomiální modely se zabývá Lin (2016). Porovnáním metod pro sestrojení pásu spolehlivosti pro kvadratickou závislost na omezeném intervalu a jejich přesností se zabývá Spurier (1993).

1. Model lineární regrese

Budeme se zabývat závislostí náhodné veličiny Y na nenáhodných proměnných $x_1, \ldots, x_p$, $p \geq 1$, uvedeme definici a základní vlastnosti lineárního modelu. Budeme vycházet z knih Anděl (2007) a Zvára (2008). Nechť $Y_1, \ldots, Y_n$ jsou náhodné veličiny, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ a $\boldsymbol{X} = (x_{ij})$ je matice daných čísel typu $n \times p$, kde $p < n$. Model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{e}$$

nazveme lineárním modelem, kde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ je vektor neznámých parametrů a $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$ je vektor náhodných veličin splňující $\mathsf{E}\ \boldsymbol{e} = \boldsymbol{0}$, $\operatorname{var} \boldsymbol{e} = \sigma^2 \boldsymbol{I_n}$, kde $\sigma^2 > 0$ je neznámý parametr.

Dále budeme požadovat, aby matice $\boldsymbol{X}$ měla lineárně nezávislé sloupce. Potom $h(\boldsymbol{X}) = p$ a $\boldsymbol{X}^\top \boldsymbol{X}$ je regulární matice typu $p \times p$, neboť předpodkládáme, že $p < n$.

Parametry $\theta_1, \ldots, \theta_p$ se odhadují metodou nejmenších čtverců, tedy hledáme minimum $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})$ jakožto funkce $\boldsymbol{\theta}$.

Věta 1. *Nechť $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{e}$ je lineární model. Odhad $\hat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$ metodou nejmenších čtverců je dán vztahem*

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Důkaz. Pro vektor $\hat{\boldsymbol{\theta}}$ platí

$$\boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) = \boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = 0.$$

Odtud dostáváme

$$\begin{aligned}
&(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}) \\
&\qquad = [(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) + (\boldsymbol{X}\hat{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta})]^\top [(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) + (\boldsymbol{X}\hat{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta})] \\
&\qquad = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top (\boldsymbol{X}^\top \boldsymbol{X}) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
&\qquad \geq (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}).
\end{aligned}$$

Matice $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ je regulární a pozitivně definitní, proto nastává rovnost právě tehdy, když $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

□

Věta 2. *Nechť $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{e}$ je lineární model, $\hat{\boldsymbol{\theta}}$ je odhad $\boldsymbol{\theta}$ metodou nejmenších čtverců. Potom platí*

$$\begin{aligned}
\mathsf{E}\, \hat{\boldsymbol{\theta}} &= \boldsymbol{\theta}, \\
\operatorname{var} \hat{\boldsymbol{\theta}} &= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.
\end{aligned}$$

Důkaz. Odhad $\hat{\boldsymbol{\theta}}$ splňuje $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ a proto

$$\begin{aligned}
\mathsf{E}\ \hat{\boldsymbol{\theta}} &= \mathsf{E}\, (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top\, \mathsf{E}\ \boldsymbol{Y} = \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{\theta},
\end{aligned}$$

$$\begin{aligned}\operatorname{var}\hat{\boldsymbol{\theta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(var\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.\end{aligned}$$

□

Označme $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$ odhad vektoru $\boldsymbol{Y}$. Veličinu $S_e = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ nazveme reziduální součet čtverců.

Následující tvrzení uvedeme bez důkazů. Ty lze najít např. v Anděl (2007).

Věta 3. *Nechť $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{e}$ je lineární model a S_e je reziduální součet čtverců. Náhodná veličina $s^2 = S_e/(n-p)$ je nestranným odhadem parametru σ^2.*

Dále budeme předpokládat, že $\boldsymbol{e} \sim \mathsf{N}(\boldsymbol{0},\sigma^2\boldsymbol{I})$ a tedy $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$. V tomto případě hovoříme o normálním lineárním modelu. Uvedeme některé vlastnosti normálního lineárního modelu, které později použijeme k sestrojení pásů spolehlivosti.

Věta 4. *Nechť $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$ je normální lineární model, $\hat{\boldsymbol{\theta}}$ je odhad $\boldsymbol{\theta}$ metodou nejmenších čtverců a S_e je reziduální součet čtverců. Potom platí*

$$\begin{aligned}&\hat{\boldsymbol{\theta}} \sim \mathsf{N}(\boldsymbol{\theta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}),\\ &S_e/\sigma^2 \sim \chi^2_{n-p}.\end{aligned}$$

Věta 5. *Nechť $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$ je normální lineární model. Odhad $\hat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$ metodou nejmenších čtverců a $s^2 = S_e/(n-p)$ jsou nezávislé.*

Věta 6. *Nechť $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$ je normální lineární model, $\hat{\boldsymbol{\theta}}$ je odhad $\boldsymbol{\theta}$ metodou nejmenších čtverců a $s^2 = S_e/(n-p)$. Označme v_{ij} prvky matice $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. Pro $j = 1, \ldots, p$ platí*

$$T_j = \frac{\theta_j - \hat{\theta}_j}{s\sqrt{v_{jj}}} \sim t_{n-p}.$$

Důsledek. Interval

$$(\hat{\theta}_j - s\sqrt{v_{jj}}t_{n-p}(1-\alpha/2),\ \hat{\theta}_j + s\sqrt{v_{jj}}t_{n-p}(1-\alpha/2))$$

tvoří interval spolehlivosti pro θ_j s pravěpodobností pokrytí $1-\alpha$, kde $t_{n-p}(1-\alpha/2)$ značí $1-\alpha/2$ kvantil t rozdělení s $n-p$ stupni volnosti.

Věta 7 (Zvára, 2008)**.** *Nechť $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$ je normální lineární model, $\hat{\boldsymbol{\theta}}$ je odhad $\boldsymbol{\theta}$ metodou nejmenších čtverců a $s^2 = S_e/(n-p)$. Množina*

$$K_2 = \left\{\boldsymbol{\theta} \in \mathbb{R}^p : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) < ps^2F_{p,n-p}(1-\alpha)\right\}$$

tvoří množinu spolehlivosti pro parametr $\boldsymbol{\theta}$ se spolehlivostí $1-\alpha$, kde $F_{p,n-p}(1-\alpha)$ značí $1-\alpha$ kvantil F rozdělení s p a $n-p$ stupni volnosti.

Příklad 1. Určíme odhady neznámých parametrů v příkladu polynomiální regrese. Uvažujme regresní polynom stupně 3.

$$Y_{x_i} = 1.6x_i - 0.2x_i^2 + 0.01x_i^3 + e_i,$$

kde $e_i \sim \mathsf{N}(0,1)$, $i = 1, \ldots, n$ a zvolme $n = 12$. Nasimulujeme pozorování v bodech $x_i = i$ pro $i = 1, \ldots, 12$ v programu R (viz obrázek 2.1).

x_i	1	2	3	4	5	6	7	8	9	10	11	12
Y_{x_i}	2,36	2	3,16	4,79	4,84	5,06	6,01	4,02	5,19	6,92	5,56	6

Máme tedy

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \\ 1 & 5 & 25 & 125 \\ 1 & 6 & 36 & 216 \\ 1 & 7 & 49 & 343 \\ 1 & 8 & 64 & 512 \\ 1 & 9 & 81 & 729 \\ 1 & 10 & 100 & 1000 \\ 1 & 11 & 121 & 1331 \\ 1 & 12 & 144 & 1728 \end{pmatrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} 2{,}36 \\ 2{,}00 \\ 3{,}16 \\ 4{,}79 \\ 4{,}84 \\ 5{,}06 \\ 6{,}01 \\ 4{,}02 \\ 5{,}19 \\ 6{,}92 \\ 5{,}56 \\ 6{,}00 \end{pmatrix},$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} 2{,}677 & -1{,}584 & 0{,}253 & -0{,}012 \\ -1{,}584 & 1{,}095 & -0{,}187 & 0{,}009 \\ 0{,}253 & -0{,}187 & 0{,}034 & -0{,}002 \\ -0{,}012 & 0{,}009 & -0{,}002 & 0{,}0001 \end{pmatrix}.$$

Odhad $\hat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$ metodou nejmenších čtverců je

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = (0{,}72;\ 1{,}25;\ -0{,}11;\ 0{,}004)^\top. \tag{1.1}$$

Odhad $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$ vektoru $\mathsf{E}\ \boldsymbol{Y}$

$$\hat{\boldsymbol{Y}} = (1{,}86\ \ 2{,}80\ \ 3{,}55\ \ 4{,}16\ \ 4{,}62\ \ 4{,}98\ \ 5{,}24\ \ 5{,}44\ \ 5{,}60\ \ 5{,}74\ \ 5{,}88\ \ 6{,}04)^\top.$$

Reziduální součet čtverců $S_e = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = 5{,}79$. Odhad s^2 parametru σ^2 je $s^2 = S_e/(n-p) = 0{,}72$.

2. Pásy spolehlivosti

V této kapitole budeme vycházet z knihy Zvára (2008). Budeme se zabývat pouze normálními lineárními modely, kde lze popsat závislost Y na x pomocí funkcí jedné proměnné.

Uvažujme regresní závislost $Y_{x_i} = \boldsymbol{x}^\top(x_i)\boldsymbol{\theta} + e_i$, kde $\boldsymbol{x}(x)$ je vektor známých spojitých funkcí, $x_i \in \mathbb{R}$ a $e_i \sim \mathsf{N}(0,\sigma^2)$ pro $i = 1,\ldots,n$. Označme vektor $\boldsymbol{Y} = (Y_{x_1},\ldots,Y_{x_n})^\top$ a matici $\boldsymbol{X}$, jejíž i-tý řádek je $\boldsymbol{x}^\top(x_i)$. Požadujeme, aby sloupce matice $\boldsymbol{X}$ byly lineárně nezávislé.

2.1 Interval spolehlivosti pro pevné x

Uvažujme nyní jedinou pevnou hodnotu $x_0 \in \mathbb{R}$. Cílem bude najít interval spolehlivosti pro $\mathsf{E}\, Y_{x_0} = \boldsymbol{x}^\top(x_0)\boldsymbol{\theta}$. Bodovým odhadem je statistika $\boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}}$ s rozptylem

$$\sigma^2\boldsymbol{x}^\top(x_0)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}(x_0) = \sigma^2 d^2(x_0),$$

kde označíme $d^2(x) = \boldsymbol{x}^\top(x)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}(x)$.

S využitím Věty 4 a Věty 5 víme, že

$$T_1 = \frac{\boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}} - \boldsymbol{x}^\top(x_0)\boldsymbol{\theta}}{\sigma d(x_0)} \sim \mathsf{N}(0,1),$$

$$T_2 = \frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}$$

a T_1 a T_2 jsou nezávislé. Potom

$$\frac{\boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}} - \boldsymbol{x}^\top(x_0)\boldsymbol{\theta}}{s\ d(x_0)} = \frac{T_1}{\sqrt{T_2/(n-p)}} \sim t_{n-p}. \tag{2.1}$$

Získané poznatky jsou popsány v následující větě.

Věta 8. *Nechť $Y_{x_i} = \boldsymbol{x}^\top(x_i)\boldsymbol{\theta} + e_i,\ i = 1,\ldots,n$, je normální lineární model, kde $\boldsymbol{x}^\top(x)$ je vektor spojitých funkcí a nechť $x_0 \in \mathbb{R}$. Potom platí*

$$P\left(\left|\frac{\boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}} - \boldsymbol{x}^\top(x_0)\boldsymbol{\theta}}{s\ d(x_0)}\right| < t_{n-p}\left(1-\frac{\alpha}{2}\right)\right) = 1-\alpha$$

a interval spolehlivosti pro $\mathsf{E}\, Y_{x_0} = \boldsymbol{x}^\top(x_0)\boldsymbol{\theta}$ je

$$(\boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}} - t_{n-p}(1-\alpha/2)s\ d(x_0), \boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}} + t_{n-p}(1-\alpha/2)s\ d(x_0)). \tag{2.2}$$

Příklad 2. Speciálně pro regresní přímku $Y_{x_i} = \theta_0 + \theta_1 x_i + e_i$ platí

$$\begin{aligned}
\boldsymbol{x}(x_i) &= (1\ x_i),\\
\boldsymbol{x}^\top(x)\hat{\boldsymbol{\theta}} &= \hat{\theta}_0 + \hat{\theta}_1 x,\\
\boldsymbol{X}^\top\boldsymbol{X} &= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix},\\
(\boldsymbol{X}^\top\boldsymbol{X})^{-1} &= \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}\\
&= \frac{1}{\sum x_i^2 - n\bar{x}^2}\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix},
\end{aligned}$$

kde $\bar{x} = \sum_{i=1}^{n} x_i/n$.

$$\begin{aligned} d^2(x) &= (1\ x)(\boldsymbol{X}^\top \boldsymbol{X})^{-1}(1\ x)^\top \\ &= \frac{1}{\sum x_i^2 - n\bar{x}^2}\left(\frac{1}{n}\sum x_i^2 - x\bar{x} - x\bar{x} + x^2\right) \\ &= \frac{1}{\sum x_i^2 - n\bar{x}^2}\left[\frac{1}{n}\left(\sum x_i^2 - n\bar{x}^2\right) + \bar{x}^2 - 2x\bar{x} + x^2\right] \\ &= \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)}. \end{aligned}$$

Dosazením do vzorce (2.2) dostaneme, pro pevné $x \in \mathbb{R}$, interval spolehlivosti s krajními body

$$\left(\hat{\theta}_0 + \hat{\theta}_1 x \pm s\ t_{n-2}(1-\alpha/2)\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)}}\right). \tag{2.3}$$

Intervaly spolehlivosti pro skutečná data jsou zobrazeny na obrázku 2.1.

2.2 Predikční interval pro pevné x

Sestrojme nyní predikční interval takový, že s danou pravděpodobností $1-\alpha$ obsahuje nezávislé budoucí pozorování Y_{x_0}, kde $x_0 \in \mathbb{R}$ je předem pevně dané. Zajímá nás tedy $Y_{x_0} = \boldsymbol{x}^\top(x_0)\boldsymbol{\theta} + e$, kde $e \sim \mathsf{N}(0,\sigma^2)$. Bodovým odhadem je opět statistika $\boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}}$, ale rozdíl $Y_{x_0} - \boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}}$ má rozptyl $\sigma^2(1+d^2(x_0))$, protože Y_{x_0} a $\boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}}$ jsou nezávislé náhodné veličiny. Obdobně jako v (2.1) má statistika

$$\frac{Y_{x_0} - \boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}}}{s\sqrt{1 + d^2(x_0)}}$$

t rozdělení s $n-p$ stupni volnosti a pro x_0 platí následující věta.

Věta 9. *Nechť $Y_{x_i} = \boldsymbol{x}^\top(x_i)\boldsymbol{\theta} + e_i,\ i = 1,\dots,n$, je normální lineární model, kde $\boldsymbol{x}^\top(x)$ je vektor spojitých funkcí a nechť $x_0 \in \mathbb{R}$. Potom platí*

$$P\left(\left|\frac{Y_{x_0} - \boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}}}{s\sqrt{1 + d^2(x_0)}}\right| < t_{n-p}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

a predikční interval pro $Y_{x_0} = \boldsymbol{x}^\top(x_0)\boldsymbol{\theta} + e$ je

$$\left(\boldsymbol{x}^\top(x_0)\hat{\boldsymbol{\theta}} \pm t_{n-p}(1-\alpha/2)s\sqrt{1 + d^2(x_0)}\right). \tag{2.4}$$

Příklad 3. Stejně jako v příkladu 2 dostaneme pro regresní přímku $Y_{x_i} = \theta_0 + \theta_1 x_i + e_i$

$$\begin{aligned} \boldsymbol{x}^\top(x)\hat{\boldsymbol{\theta}} &= \hat{\theta}_0 + \hat{\theta}_1 x, \\ d^2(x) &= \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)}. \end{aligned}$$

Dosazením do vzorce (2.4) získáme predikční interval pro přímku

$$\left(\theta_0 + \theta_1 x \pm s\ t_{n-2}(1-\alpha/2)\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i^2-\bar{x}^2)}}\right). \tag{2.5}$$

Predikční interval je širší, než interval spolehlivosti, neboť bodový odhad $Y_{x_0} = \boldsymbol{x}^\top(x_0)\boldsymbol{\theta} + e$ má větší rozptyl, než bodový odhad $\mathsf{E}\, Y_{x_0} = \boldsymbol{x}^\top(x_0)\boldsymbol{\theta}$ (viz obrázek 2.1).

2.3 Pás spolehlivosti

Následující postup vychází z knihy Zvára (2008). Budeme se zabývat průběhem $\mathsf{E}\, Y_x = \boldsymbol{x}^\top(x)\boldsymbol{\theta}$ pro všechna $x \in \mathbb{R}$ a sestavením pásu spolehlivosti s pravděpodobností pokrytí alespoň $1-\alpha$.

Nechť K je množina spolehlivosti pro parametr $\boldsymbol{\theta}$. Zavedeme funkce

$$L(x) = \inf_{\theta\in K} \boldsymbol{x}^\top(x)\boldsymbol{\theta} \qquad U(x) = \sup_{\theta\in K} \boldsymbol{x}^\top(x)\boldsymbol{\theta}. \tag{2.6}$$

Pás spolehlivosti na intervalu $T \subset \mathbb{R}$ sestrojíme jako množinu

$$M = \left\{(x,y)^\top : L(x) < y < U(x), x \in T\right\}.$$

Je-li spolehlivost K rovna $1-\alpha$ pak má pás M pravděpodobnost pokrytí alespoň $1-\alpha$ pro všechna $x \in T$ současně. Neboli

$$\mathsf{P}\left(\{(x, \boldsymbol{x}^\top(x)\boldsymbol{\theta})^\top, x \in T\} \subset M\right) \geq \mathsf{P}\left(\boldsymbol{\theta} \in K\right),$$

neboť je-li $\boldsymbol{\theta} \in K$ pak podle (2.6) pro všechny $x \in T$ platí nerovnost $L(x) \leq \boldsymbol{x}(x)^\top\boldsymbol{\theta} \leq U(x)$.

Uvažujme nyní normální lineární model. Jako výchozí množinu spolehlivosti K využijeme množinu K_2 z Věty 7

$$K_2 = \left\{\boldsymbol{\theta} \in \mathbb{R}^p : (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \leq ps^2 F_{p,n-p}(1-\alpha)\right\}.$$

Množina K_2 má tvar elipsoidu a tedy extrémy udávající $L(x)$ a $U(x)$ nastanou v jejích krajních bodech. Hledáme tedy extrémy funkce $\boldsymbol{x}^\top(x)\boldsymbol{\theta}$ za podmínky, že $(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) = c$, kde $c = ps^2F_{p,n-p}(1-\alpha)$. Můžeme použít metodu Lagrangeových multiplikátorů k jejich nalezení

$$\varphi(\boldsymbol{\theta},\lambda) = \boldsymbol{x}^\top(x)\boldsymbol{\theta} - \frac{\lambda}{2}((\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) - c).$$

Derivace podle $\boldsymbol{\theta}$ jsou nulové pro $\boldsymbol{x}(x) = \lambda\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$. Extrém tedy nastává v bodě

$$\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \frac{1}{\lambda}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}(x).$$

Dosazením do podmínky získáme hodnotu $1/\lambda$

$$\frac{1}{\lambda}\boldsymbol{x}^\top(x)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{X})\frac{1}{\lambda}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}(x) = c$$

$$\frac{1}{\lambda} = \pm\frac{\sqrt{c}}{d(x)},$$

kde $d(x) = \sqrt{\boldsymbol{x}^\top(x)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}(x)}$ je nezáporná funkce. Odtud dostaneme extrém $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} \pm \sqrt{c}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}(x)/d(x)$. Extrémní hodnota je

$$\boldsymbol{x}^\top(x)\tilde{\boldsymbol{\theta}} = \boldsymbol{x}^\top(x)\hat{\boldsymbol{\theta}} \pm \frac{\sqrt{c}}{d(x)}\boldsymbol{x}^\top(x)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}(x) = \boldsymbol{x}^\top(x)\hat{\boldsymbol{\theta}} \pm \sqrt{c}d(x).$$

Získané meze pásu jsou popsány v následující větě.

Věta 10. *Nechť $Y_{x_i} = \boldsymbol{x}^\top(x_i)\boldsymbol{\theta} + e_i,\ i = 1, \ldots, n$, je normální lineární model, kde $\boldsymbol{x}^\top(x)$ je vektor spojitých funkcí a nechť $T \subset \mathbb{R}$. Potom platí*

$$\mathsf{P}\Big(\boldsymbol{x}^\top(x)\boldsymbol{\theta} \in (L(x),U(x));\ x \in T\Big) \geq 1 - \alpha,$$

kde

$$L(x) = \boldsymbol{x}^\top(x)\hat{\boldsymbol{\theta}} - s\ d(x)\sqrt{pF_{p,n-p}(1-\alpha)}, \tag{2.7}$$

$$U(x) = \boldsymbol{x}^\top(x)\hat{\boldsymbol{\theta}} + s\ d(x)\sqrt{pF_{p,n-p}(1-\alpha)} \tag{2.8}$$

jsou meze pásu spolehlivosti.

Stejný pás spolehlivosti lze získat také Scheffého metodou, kde vycházíme z následující věty.

Věta 11 (Scheffého)**.** *Nechť $\boldsymbol{W} = (W_1, \ldots, W_p) \sim \mathsf{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{V})$, kde $\boldsymbol{V}$ je známá pozitivně definitní matice a $\sigma^2 > 0$ neznámý parametr. Nechť $\mathbb{A}$ je $t-$rozměrný podprostor prostoru $\mathbb{R}^p$ a nechť s^2 je nezávislý odhad σ^2 splňující $\nu s^2/\sigma^2 \sim \chi^2_\nu$, kde $\nu \in \mathbb{N}$ a platí, že s^2 a $\boldsymbol{W}$ jsou nezávislé. pak*

$$\mathsf{P}\Big(|\boldsymbol{a}^\top\boldsymbol{W} - \boldsymbol{a}^\top\boldsymbol{\mu}| \leq \sqrt{ts^2F_{t,\nu}(1-\alpha)\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{a}};\ \boldsymbol{a} \in \mathbb{A}\Big) = 1 - \alpha.$$

Důkaz. Lze nalézt např. v Anděl, 2007 (str. 206). □

Pro normální lineární model $Y_{x_i} = \boldsymbol{x}^\top(x_i)\boldsymbol{\theta} + e_i,\ i = 1, \ldots, n$, víme, z věty 4, že platí $\hat{\boldsymbol{\theta}} \sim \mathsf{N}(\boldsymbol{\theta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$ a $(n-p)s^2/\sigma^2 \sim \chi^2_{n-p}$. Položme $\mathbb{A} = \mathbb{R}^p$. Potom podle věty 11 platí

$$\mathsf{P}\left(|\boldsymbol{a}^\top\hat{\boldsymbol{\theta}} - \boldsymbol{a}^\top\boldsymbol{\theta}| \leq \sqrt{ps^2F_{p,n-p}(1-\alpha)\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}};\ \boldsymbol{a} \in \mathbb{R}^p\right) = 1 - \alpha.$$

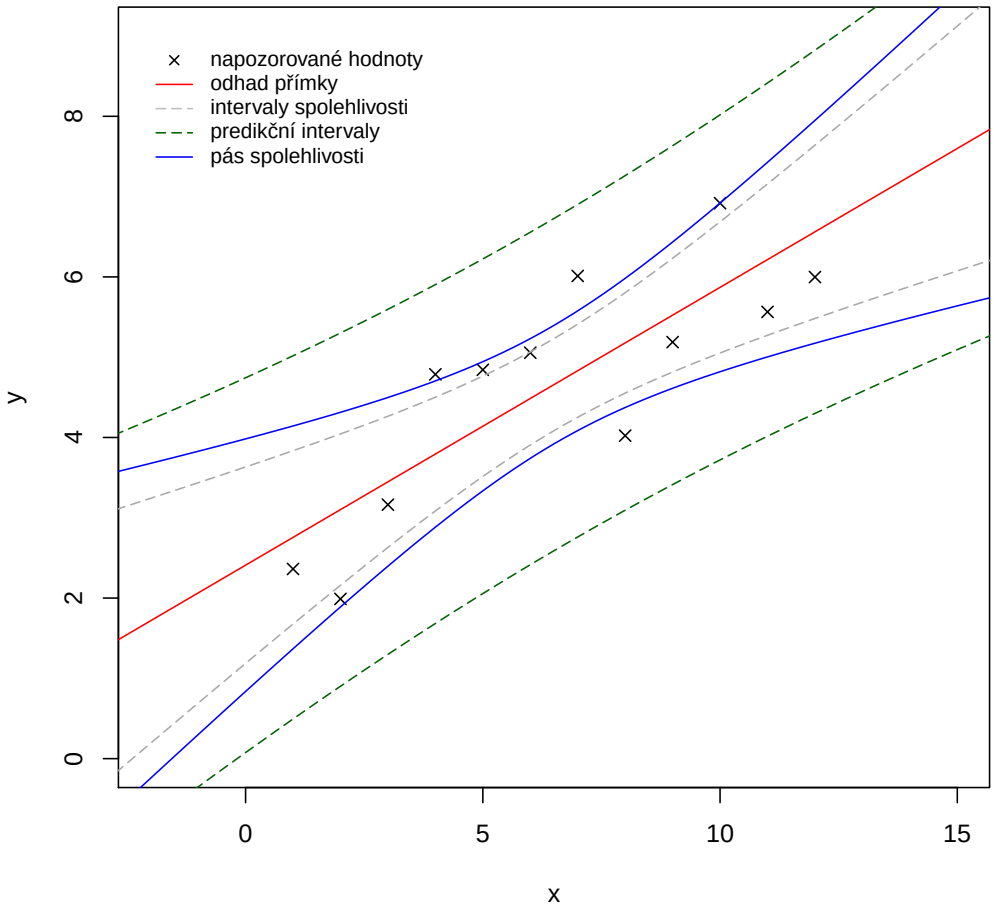
Označme $B = \{\boldsymbol{x}(x) \in \mathbb{R}^p : x \in \mathbb{R}\} \subset \mathbb{R}^p$, dostaneme

$$\mathsf{P}\left(|\boldsymbol{a}^\top\hat{\boldsymbol{\theta}} - \boldsymbol{a}^\top\boldsymbol{\theta}| \leq \sqrt{ps^2F_{p,n-p}(1-\alpha)\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}};\ \boldsymbol{a} \in B\right) \geq 1 - \alpha,$$

dosazením $a = \boldsymbol{x}(x)$ získáme pás z věty 10.

Příklad 4. Speciálně pro regresní přímku $Y_{x_i} = \theta_0 + \theta_1 x_i + e_i$, dostaneme jako v příkladu 2,

$$\boldsymbol{x}^\top(x)\hat{\boldsymbol{\theta}} = \hat{\theta}_0 + \hat{\theta}_1 x,$$
$$d^2(x) = \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i^2 - \bar{x}^2)}.$$



Obrázek 2.1: Pás spolehlivosti, interval spolehlivosti a predikční interval předpokládáme-li lineární závislost.

Meze pásu spolehlivosti v lineárně závislém modelu jsou

$$L(x) = \hat{\theta}_0 + \hat{\theta}_1 x - s\sqrt{\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)}\right) 2F_{2,n-2}(1-\alpha)},$$

$$U(x) = \hat{\theta}_0 + \hat{\theta}_1 x + s\sqrt{\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)}\right) 2F_{2,n-2}(1-\alpha)}. \qquad (2.9)$$

Pás spolehlivosti je širší, než pás sestrojený z interalů spolehlivosti, neboť dodržuje pravděpodobnost pokrytí pro všechna $x \in \mathbb{R}$ zároveň, zatímco interval spolehlivosti dodržuje přesnost pouze pro jediné pevné x (viz obrázek 2.1).

Příklad 5. V praxi obvykle neznáme stupeň polynomu. Předpokládáme-li lineární závislost dat, z příkladu 1, dostaneme ze vzorců (2.3), (2.5), (2.9) následující pásy (viz obrázek 2.1). Interval spolehlivosti pro pevné x

$$\left(2{,}41 + 0{,}35x \pm s\ t_{10}(1-\alpha/2)\sqrt{\frac{1}{12} + \frac{(x-6{,}5)^2}{143}}\right).$$

Predikční interval

$$\left(2{,}41 + 0{,}35x \pm s\ t_{10}(1-\alpha/2)\sqrt{1+\frac{1}{12}+\frac{(x-6{,}5)^2}{143}}\right).$$

Pás spolehlivosti

$$\left(2{,}41 + 0{,}35x \pm s\ \sqrt{2F_{2,10}(1-\alpha)\left(\frac{1}{12}+\frac{(x-6{,}5)^2}{143}\right)}\right).$$

Příklad 6. Nyní sestrojíme pásy spolehlivosti pro zajímavější polynomickou regresi. Uvažujme regresní polynom z příkladu 1

$$Y_{x_i} = 1{,}6x_i - 0{,}2x_i^2 + 0{,}01x_i^3 + e_i,$$

Z 1.1 již víme, že

$$\hat{\boldsymbol{\theta}} = (0{,}72;\ 1{,}25;\ -0{,}11;\ 0{,}004)^\top$$

a tedy

$$\boldsymbol{x}^\top(x)\hat{\boldsymbol{\theta}} = 0{,}72 + 1{,}25x - 0{,}11x^2 + 0{,}004x^3.$$

Pro polynom třetího stupně platí

$$d^2(x) = \begin{pmatrix}1 & x & x^2 & x^3\end{pmatrix}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\begin{pmatrix}1 & x & x^2 & x^3\end{pmatrix}^\top.$$

Požadujme pravděpodobnost pokrytí 95%, tedy $\alpha = 0{,}05$.

Dosazením do vzorce (2.2) dostaneme interval spolehlivosti pro pevné x

$$\begin{aligned}(&0{,}72+1{,}25x - 0{,}11x^2 + 0{,}004x^3 - s\ t_8(1-\alpha/2)d(x),\\ &0{,}72 + 1{,}25x - 0{,}11x^2 + 0{,}004x^3 + s\ t_8(1-\alpha/2)d(x)).\end{aligned}$$

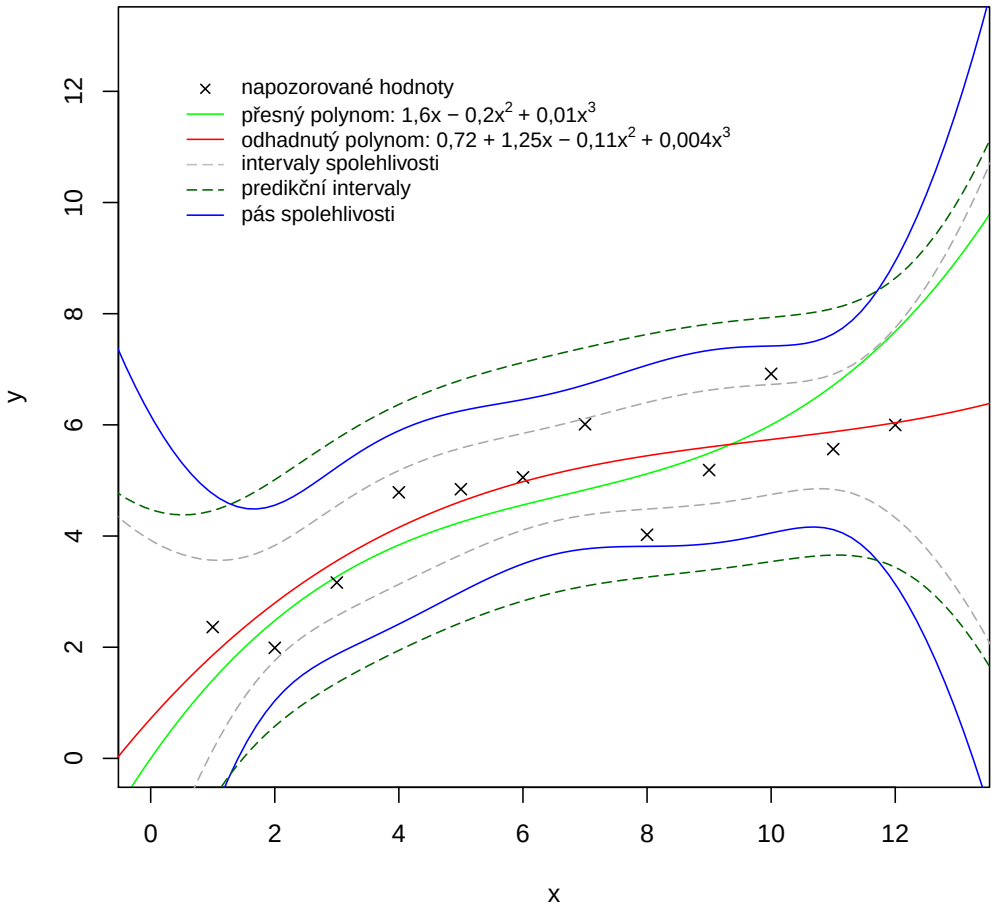
Podle vzorce (2.4) dostaneme predikční interval pro pevné x

$$\begin{aligned}(&0{,}72+1{,}25x - 0{,}11x^2 + 0{,}004x^3 - s\ t_8(1-\alpha/2)\sqrt{1+d^2(x)},\\ &0{,}72 + 1{,}25x - 0{,}11x^2 + 0{,}004x^3 + s\ t_8(1-\alpha/2)\sqrt{1+d^2(x)}).\end{aligned}$$

Meze pásu spolehlivosti jsou podle (2.7)

$$\begin{aligned}L(x) &= 0{,}72 + 1{,}25x - 0{,}11x^2 + 0{,}004x^3 - s\ \sqrt{4d^2(x)F_{4,8}(1-\alpha)},\\ U(x) &= 0{,}72 + 1{,}25x - 0{,}11x^2 + 0{,}004x^3 + s\ \sqrt{4d^2(x)F_{4,8}(1-\alpha)}.\end{aligned}$$

Na obrázku 2.2 vidíme, že pás spolehlivosti je širší než pás sestrojený z intervalů spolehlivosti pro pevné x, neboť pás spolehlivosti dodržuje pravděpodobnost pokrytí pro všechna $x \in \mathbb{R}$ současně. Také pás sestrojený z predikčních intervalů je širší než pás sestrojený z intervalů spolehlivosti pro pevné x, neboť bodový odhad má větší rozptyl.



Obrázek 2.2: Pás spolehlivosti, interval spolehlivosti a predikční interval pro kubickou závislost.

3. Pás spolehlivosti pro regresní polynomy

Pásy z předchozí kapitoly jsou konzervativní (skutečná pravděpodobnost pokrytí je větší než požadovaná) a tedy moc široké, proto se podíváme na konstrukci přesných pásů spolehlivosti, kterou popsal Wynn (1984). Budeme sestrojovat přesný pás spolehlivosti pro jednodimensionální polynom stupně $p-1$ na intervalu $T \subseteq (-\infty, \infty)$. Pás je konstruován umísťováním intervalů kolem předpokládané hodnoty regresního modelu v nulových bodech p-tého ortogonálního polynomu vzhledem k daným bodům. Z výpočtů vyplývá, že pás je po částech polynomický stupně $p-1$.

Uvažujme regresní polynom stupně $p-1$:

$$\begin{aligned} Y_{x_i} &= \theta_0 + \theta_1 x_i + \cdots + \theta_{p-1} x_i^{p-1} + e_i, \\ \mathsf{E}\, Y_{x_i} &= \theta_0 + \theta_1 x_i + \cdots + \theta_{p-1} x_i^{p-1}, \end{aligned}$$

kde Y_{x_i} jsou pozorované hodnoty v daných bodech x_i $(i = 1, \ldots, n)$. Předpokládejme, že $n > p$ a Y_i jsou nezávislé s rozptylem σ^2. Najdeme odhad $\hat{\theta}$ parametru $\boldsymbol{\theta}$ pomocí metody nejmenších čtverců. Tedy $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$, kde $\boldsymbol{Y} = (Y_{x_1}, \ldots, Y_{x_n})^\top$ a

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{p-1} \end{pmatrix}$$

je matice typu $n \times p$. Označme

$$\hat{Y}_x = \hat{\theta}_0 + \hat{\theta}_1 x + \cdots + \hat{\theta}_{p-1} x^{p-1}.$$

3.1 Kvadraturní body

Sestrojíme ortogonální polynomy $\phi_0(x), \ldots, \phi_p(x)$ takové, že

$$\sum_{k=1}^{n} \phi_i(x_k)\phi_j(x_k) = \begin{cases} n, & i = j, \\ 0, & i \neq j. \end{cases}$$

Označme $z_1, \ldots, z_p$ kořeny polynomu $\phi_p(x)$. Z podmínky regularity modelu jsou tyto kořeny různé.

Věta 12 (Szegö, 1967, str. 47)**.** *Jsou-li $z_1 < \cdots < z_n$ kořeny polynomu $p_n(z)$, pak existují $\lambda_1, \ldots \lambda_n \in \mathbb{R}$ takové, že*

$$\int_a^b \rho(z) d\alpha(z) = \lambda_1 \rho(z_1) + \cdots + \lambda_n \rho(z_n),$$

pro libovolný polynom $\rho(z)$ stupně $\leq 2n-1$. Míra $d\alpha(z)$ a číslo n udávají $\lambda_1, \ldots, \lambda_n$ jednoznačně.

Je-li $\pi(x)$ polynom stupně $\leq 2p-1$ potom podle Věty 12 existují $\lambda_1, \dots \lambda_p > 0$ takové, že

$$n^{-1} \sum_{i=1}^{p} \pi(x_i) = \lambda_1 \pi(z_1) + \cdots + \lambda_p \pi(z_p), \tag{3.1}$$

kde $z_1, \dots, z_p$ jsou kořeny polynomu $\phi_p(x)$

Lemma 13. *Existují body z_i a váhy $\lambda_i > 0$ $(i = 1, \dots, n)$ takové, že*

$$M = n^{-1} \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{Z}^\top \boldsymbol{\Lambda} \boldsymbol{Z},$$

kde $\boldsymbol{\Lambda} = diag(\lambda_1, \dots, \lambda_p)$ a

$$\boldsymbol{Z} = \begin{pmatrix} 1 & z_1 & z_1^2 & \dots & z_1^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & z_p & z_p^2 & \dots & z_p^{p-1} \end{pmatrix}$$

je matice typu $p \times p$.

Důkaz. Plyne přímo z (3.1).

□

Lemma 14. *Je-li $\hat{Y}_{z_i}$ odhad $\mathsf{E}\, Y_{z_i}$ v kvadraturních bodech z_i $(i = 1, \dots, p)$ potom pro i,$j \in \{1, \dots, p\}$ platí*

$$cov(\hat{Y}_{z_i}, \hat{Y}_{z_j}) = \begin{cases} 0, & i \neq j, \\ \sigma^2 n^{-1} \lambda_i^{-1}, & i = j. \end{cases}$$

Důkaz. Nechť $\hat{\boldsymbol{Y}}_{\boldsymbol{Z}} = (\hat{Y}_{z_1}, \dots, \hat{Y}_{z_p})^\top = \boldsymbol{Z}\hat{\boldsymbol{\theta}}$. Potom varianční matice vektoru $\hat{\boldsymbol{Y}}_Z$ je

$$\mathrm{Var}(\hat{\boldsymbol{Y}}_{\boldsymbol{Z}}) = \mathrm{Var}(\boldsymbol{Z}\hat{\boldsymbol{\theta}}) = \boldsymbol{Z}\mathrm{Var}(\hat{\boldsymbol{\theta}})\boldsymbol{Z}^\top = \sigma^2 \boldsymbol{Z}(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{Z}^\top = \sigma^2 n^{-1} \boldsymbol{\Lambda}^{-1},$$

kde poslední rovnost plyne z Lemma 13.

□

Dále podle Szegö (1967) pro všechna λ_j, $j = 1, \dots, p$ platí

$$\lambda_j^{-1} = \sum_{i=0}^{p-1} \phi_i^2(z_j) \qquad (j = 1, \dots, n).$$

3.2 Pás

Pás je konstruovám tak, aby pokrýval odpovídající hodnotu $\mathsf{E}\, Y_{z_i}$ současně pro všechna $i = 1, \dots, p$, tedy

$$y_{*i} \leq \mathsf{E}\, Y_{z_i} \leq y_i^*, \tag{3.2}$$

kde

$$y_{*i} = \hat{Y}_{z_i} - c\sqrt{\mathrm{var}\, \hat{Y}_{z_i}}, \qquad y_i^* = \hat{Y}_{z_i} + c\sqrt{\mathrm{var}\, \hat{Y}_{z_i}}.$$

Výpočtu konstanty c se budeme věnovat později. Jelikož $\mathsf{E}\, Y_x$ je polynom stupně $p-1$, pak lze nerovnost (3.2) zapsat ve tvaru $g_*(x) \le \mathsf{E}\, Y_x \le g^*(x)$, $\forall x \in T$, kde

$$\begin{aligned} g_*(x) &= \inf\{q(x) \mid y_{*i} \le q(z_i) \le y_i^*;\ i = 1, \dots, p;\ deg\{q(x)\} = p-1\} \\ g^*(x) &= \sup\{q(x) \mid y_{*i} \le q(z_i) \le y_i^*;\ i = 1, \dots, p;\ deg\{q(x)\} = p-1\} \end{aligned} \tag{3.3}$$

Zde inf a sup uvažujeme nad všemi polynomy a T je jakákoli množina reálných čísel obsahující z_i $(i = 1, \dots, p)$. Problémem zůstává určit $g_*(x)$ a $g^*(x)$. Uvažme Lagrangeův interpolační polynom $q(x)$ (viz Věta o interpolaci Stanovský (2010), str. 54) stupně $p-1$ splňující $q(z_i) = y_i$, $i = 1, \dots, p$, kde y_i jsou libovolné. Tedy

$$q(x) = \sum_{i=1}^{p} y_i l_i(x), \tag{3.4}$$

kde

$$l_i(x) = \prod_{j \ne i} \frac{x - z_j}{z_i - z_j}.$$

Berme y_i jako proměnnou, určíme

$$\operatorname{sgn}\left\{\frac{\partial q(x)}{\partial y_i}\right\} = \operatorname{sgn}\{l_i(x)\} = \begin{cases} \operatorname{sgn}\{(-1)^{i+1}\} & x < z_1, \\ \operatorname{sgn}\{(-1)^{i+j+1}\} & z_j < x < z_{j+1}, j = 1, \dots, p-1, \\ \operatorname{sgn}\{(-1)^{i+p+1}\} & z_p < x. \end{cases} \tag{3.5}$$

Vhodnou náhradou y_i za y_{*i} popř. y_i^*, podle znaménka derivace v (3.5) dosáhneme sup nebo inf z (3.3) v každém intervalu s krajními body z_i. Konečné řešení pro $g^*(x)$ je dáno vztahem

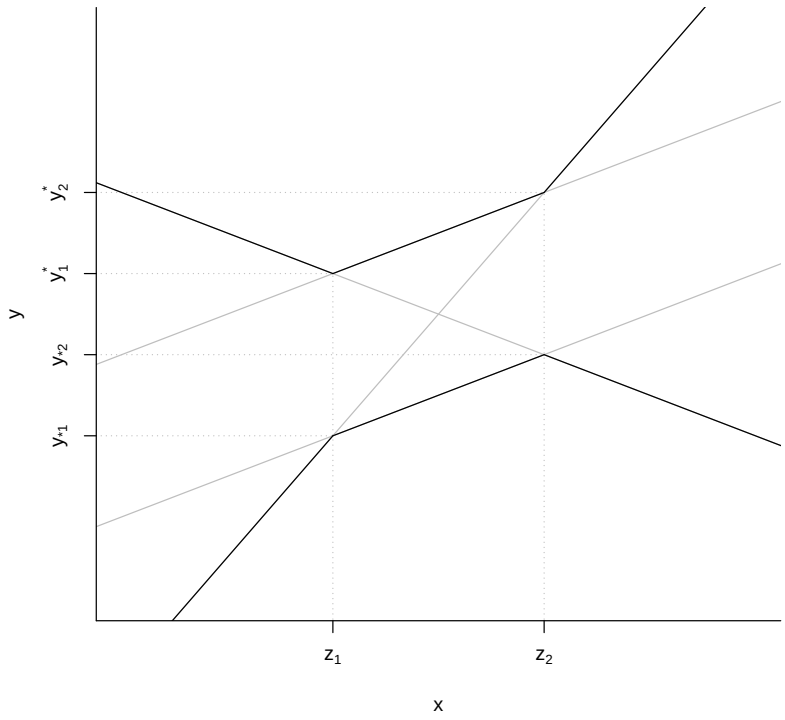
$$\begin{aligned} &\sum_{k=1}^{[\frac{1}{2}p+\frac{1}{2}]} y^*_{2k-1} l_{2k-1}(x) + \sum_{k=1}^{[\frac{1}{2}p]} y_{*2k} l_{2k}(x) && x \le z_1 \\ &\sum_{k=1}^{[\frac{1}{2}j+\frac{1}{2}]} y^*_{j-2k+2} l_{j-2k+2}(x) + \sum_{k=1}^{[\frac{1}{2}j]} y_{*j-2k+1} l_{j-2k+1}(x) \\ &\qquad + \sum_{k=1}^{[\frac{1}{2}p-\frac{1}{2}j+\frac{1}{2}]} y^*_{j+2k-1} l_{j+2k-1}(x) + \sum_{k=1}^{[\frac{1}{2}p-\frac{1}{2}j]} y_{*j+2k} l_{j+2k}(x) && z_j \le x \le z_{j+1}, \\ &\sum_{k=1}^{[\frac{1}{2}p+\frac{1}{2}]} y^*_{p-2k+2} l_{p-2k+2}(x) + \sum_{k=1}^{[\frac{1}{2}p]} y_{*p-2k+1} l_{p-2k+1}(x) && z_p \le x, \end{aligned} \tag{3.6}$$

kde [.] značí dolní celou část. Zřejmě platí $g^*(z_j) = y_j^*$. Vzorec pro $g_*(x)$ se získá záměnou y_* a y^* a obráceně.

Příklad 7. Speciálně pro regresní přímku $Y_{x_i} = \theta_0 + \theta_1 x_i + e_i$ platí

$$g^*(x) = \begin{cases} y_1^* l_1(x) + y_{*2} l_2(x) & x \le z_1, \\ y_1^* l_1(x) + y_2^* l_2(x) & z_1 \le x \le z_2, \\ y_{*1} l_1(x) + y_2^* l_2(x) & z_2 \le x. \end{cases}$$

$$g_*(x) = \begin{cases} y_{1*} l_1(x) + y_2^* l_2(x) & x \le z_1, \\ y_{*1} l_1(x) + y_{*2} l_2(x) & z_1 \le x \le z_2, \\ y_1^* l_1(x) + y_{*2} l_2(x) & z_2 \le x. \end{cases} \tag{3.7}$$



Obrázek 3.1: Pás spolehlivosti pro přímku.

Na obrázku 3.1 je ukázka horní meze $g^*(x)$ a dolní meze $g_*(x)$ pásu spolehlivosti pro regresní přímku. Pás spolehlivosti pro skutečná data je na obrázku 3.2.

3.3 Konstanta c

Zbývá dopočítat konstantu c. Předpokládejme nyní, že $\boldsymbol{Y}$ má p-rozměrné normální rozdělení. Jestliže σ^2 je známé, využijeme nezávislosti $\hat{Y}_{z_i}$, která vyplývá z Lemma 14, potom

$$\mathsf{P}\{y_{*i} \leq \mathsf{E}\ Y_{z_i} \leq y_i^*;\ i = 1, \ldots, p\} = \prod \mathsf{P}\{y_{*i} \leq \mathsf{E}\ Y_z \leq y_i^*\} = (2\Phi(c) - 1)^p,$$

kde Φ je distribuční funkce normálního rozdělení. Konstantu c dopočteme z rovnice $(2\Phi(c) - 1)^p = 1 - \alpha$, kde $1 - \alpha$ je požadovaná pravděpodobnost pokrytí. Je-li σ^2 neznámé, budeme uvažovat nestranný odhad s^2, pro něž platí $(n-p)s^2/\sigma^2 \sim \chi^2_{n-p}$. Potom

$$Z_i = \frac{\frac{\hat{Y}_{z_i} - \mathsf{E}\ Y_{z_i}}{\sqrt{\mathrm{var}(\hat{Y}_{z_i})}}}{\sqrt{\frac{(n-p)s^2/\sigma^2}{n-p}}} = \frac{\sigma(\hat{Y}_{z_i} - \mathsf{E}\ Y_{z_i})}{s\sqrt{\mathrm{var}(\hat{Y}_{z_i})}},\ \ (i = 1, \ldots, p)$$

má p-rozměrné t rozdělení s $(n-p)$ stupni volnosti. Požadujeme

$$\mathsf{P}(\mid Z_i \mid \leq c\sigma s^{-1};\ i = 1, \ldots, p) = 1 - \alpha.$$

Odtud můžeme určit $c\sigma s^{-1}$ z tabulek k-rozměrného t rozdělení (viz např. Hahn a Hendrickson, 1971) nebo počítačovou simulací (metoda Monte Carlo).

Příklad 8. Popsanou konstrukci pásu spolehlivosti předvedeme na datech z příkladu 1 za předpokladu lineární závislosti. Máme tedy $n = 12$, $p = 2$. Sestrojíme ortogonální polynomy ϕ_0, ϕ_1, ϕ_2 tak, aby

$$\sum_{k=1}^{12} \phi_i(x_k)\phi_j(x_k) = \begin{cases} 12, & i = j, \\ 0, & i \neq j \end{cases} \tag{3.8}$$

a najdeme kořeny z_1, z_2 polynomu $\phi_2(x)$. Můžeme využít knihovnu *polynom* v programu R a funkci *poly.orth*(x, *degree* = p), která sestrojí ortonormální polynomy do stupně p takové, že v bodech $x = (x_1, \ldots, x_n)$

$$\sum_{k=1}^{n} \phi_i(x_k)\phi_j(x_k) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

proto koeficienty všech polynomů vynásobíme $\sqrt{12}$, aby splňovaly rovnost (3.8) a použijeme funkci *polyroot* k nalezení kořenů $\phi_2(x)$.

```
> library(polynom)
> x <- 1:12
> phi <- poly.orth(x, degree = 2)
> for(i in 1:3){phi[[i]] = phi[[i]]*sqrt(12)}
> z = polyroot(fi[[3]])
```

$$\begin{aligned} \phi_0(x) &= 1; \\ \phi_1(x) &= -1{,}88 + 0{,}29x; \\ \phi_2(x) &= 2{,}88 - 1{,}23x + 0{,}09x^2; \\ z_1 &= 3{,}05; \\ z_2 &= 9{,}95. \end{aligned}$$

Spočítáme $\lambda_1^{-1}, \lambda_2^{-1}$, kde $\lambda_j^{-1} = \sum_{i=0}^{2} \phi_i^2(z_j)$

$$\begin{aligned} \lambda_1^{-1} &= 2; \\ \lambda_2^{-1} &= 2. \end{aligned}$$

Pro neznámý rozptyl lze pro 2-rozměrné t rozdělení s 10 stupni volnosti v tabulce Hahn a Hendrickson (1971) vyhledat

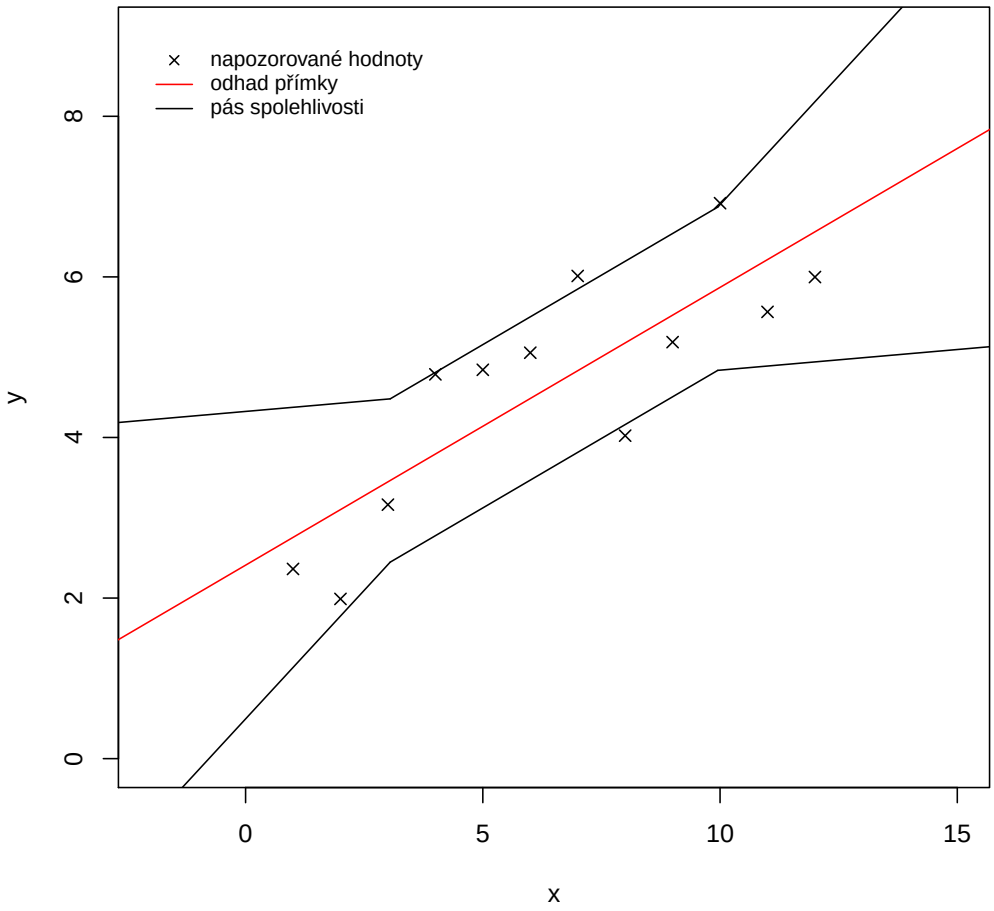
$$c\sigma s^{-1} = 2{,}609$$

a tedy

$$c\sigma = 2{,}609s = 2{,}22,$$

kde $s = \sqrt{s^2} = 0{,}89$. Pak krajní hodnoty pásu v bodech z_1, z_2 jsou

$$\begin{aligned} y_{*i} &= \hat{Y}_{z_i} - c\sigma\sqrt{n^{-1}\lambda_i^{-1}}, & y_i^* &= \hat{Y}_{z_i} + c\sigma^2\sqrt{n^{-1}\lambda_i^{-1}}, \\ y_{*1} &= 2{,}51, & y_1^* &= 4{,}41; \\ y_{*2} &= 4{,}90, & y_2^* &= 6{,}80, \end{aligned}$$



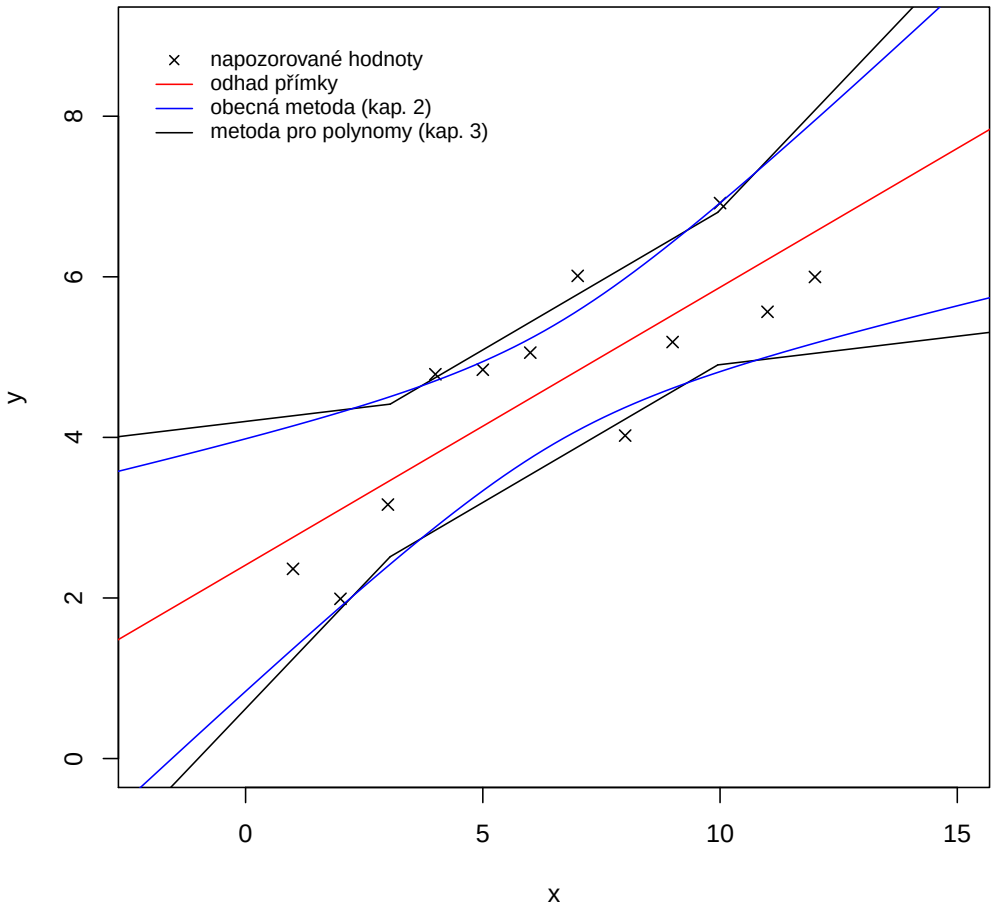
Obrázek 3.2: Pásy spolehlivosti pro přímku sestrojené motodou pro polynomy.

kde $\hat{Y}_{z_i} = \hat{\theta}_0 + \hat{\theta}_1 z_i = 2{,}41 + 0{,}35 z_i$. Lagrangeův polynom (3.4) lze v R sestrojit pomocí funkce *poly.calc*(z,y) z knihovny *polynom*, kde $z = (z_1, z_2)$, $y = (y_1, y_2)$ a $y_i = y_i^*$ popř. $y_i = y_{*i}$ podle (3.7). Výsledné meze pásu spolehlivosti (viz obrázek 3.2) jsou

$$g^* = \begin{cases} 4{,}20 + 0{,}07x & x \leq 3{,}05 \\ 3{,}36 + 0{,}35x & 3{,}05 \leq x \leq 9{,}95 \\ 0{,}62 + 0{,}62x & 9{,}95 \leq x; \end{cases}$$

$$g_* = \begin{cases} 0{,}62 + 0{,}62x & x \leq 3{,}05 \\ 1{,}45 + 0{,}35x & 3{,}05 \leq x \leq 9{,}95 \\ 4{,}20 + 0{,}07x & 9{,}95 \leq x. \end{cases}$$

Pás sestrojený tímto postupem je po částech lineární a na okolí kvadraturních bodů z_1, z_2 užší než pás spolehlivosti sestrojený postupem pro obecné funkce (z části 2.3), mimo okolí těchto bodů je však širší (viz obrázek 3.3).



Obrázek 3.3: Pásy spolehlivosti pro přímku.

Příklad 9. Nyní předvedeme konstrukci pásu spolehlivosti pro polynomiální regresi. Opět budeme uvažovat regresní polynom z příkladu 1

$$Y_{x_i} = 1{,}6x_i - 0{,}2x_i^2 + 0{,}01x_i^3 + e_i.$$

Sestrojíme ortogonální polynomy $\phi_0, \ldots, \phi_4$ tak, aby

$$\sum_{k=1}^{12} \phi_i(x_k)\phi_j(x_k) = \begin{cases} 12, & i = j, \\ 0, & i \neq j \end{cases} \tag{3.9}$$

a najdeme kořeny z_1, z_2, z_3, z_4 polynomu $\phi_4(x)$. Můžeme využít knihovnu *polynom* v programu *R* a funkci $poly.orth(x, degree = 4)$, koeficienty všech polynomů vynásobíme $\sqrt{12}$, aby splňovaly rovnost (3.9) a použijeme funkci *polyroot* k nalezení kořenů $\phi_4(x)$.

$$\begin{aligned}
\phi_0(x) &= 1; \\
\phi_1(x) &= -1{,}88 + 0{,}29x; \\
\phi_2(x) &= 2{,}88 - 1{,}23x + 0{,}09x^2; \\
\phi_3(x) &= -4{,}39 + 3{,}39x - 0{,}63x^2 + 0{,}03x^3; \\
\phi_4(x) &= 7{,}05 - 8{,}01x + 2{,}52x^2 - 0{,}29x^3 + 0{,}01x^4; \\
z_1 &= 1{,}41; \\
z_2 &= 4{,}50; \\
z_3 &= 8{,}50; \\
z_4 &= 11{,}59.
\end{aligned}$$

Spočítáme $\lambda_1^{-1}, \lambda_2^{-1}, \lambda_3^{-1}, \lambda_4^{-1}$, kde $\lambda_j^{-1} = \sum_{i=0}^{4} \phi_i^2(z_j)$

$$\begin{aligned}
\lambda_1^{-1} &= 5{,}54; \\
\lambda_2^{-1} &= 3{,}13; \\
\lambda_3^{-1} &= 3{,}13; \\
\lambda_4^{-1} &= 5{,}54.
\end{aligned}$$

Pro známý rozptyl $\sigma^2 = 1$ vyjádříme konstantu $c = u((\sqrt[12]{1-\alpha}+1)/2)$, kde $u(\alpha)$ značí $\alpha-$kvantil normálního rozdělení a $\alpha = 0{,}05$ je požadovaná pravděpodobnost pokrytí

$$c = 2{,}49.$$

Spočítáme krajní hodnoty pásu v bodech z_1, z_2, z_3, z_4

$$\begin{aligned}
y_{*i} &= \hat{Y}_{z_i} - c\sqrt{\sigma^2 n^{-1}\lambda_i^{-1}}, & y_i^* &= \hat{Y}_{z_i} + c\sqrt{\sigma^2 n^{-1}\lambda_i^{-1}}, \\
y_{*1} &= 0{,}57, & y_1^* &= 3{,}95; \\
y_{*2} &= 3{,}13, & y_2^* &= 5{,}68; \\
y_{*3} &= 4{,}25, & y_3^* &= 6{,}80; \\
y_{*4} &= 4{,}28, & y_5^* &= 7{,}66,
\end{aligned}$$

kde

$$\hat{Y}_{z_i} = \hat{\theta}_0 + \hat{\theta}_1 z_i + \cdots + \hat{\theta}_{p-1} z_i^{p-1} = 0{,}72 + 1{,}25z_i - 0{,}11z_i^2 + 0{,}004z_i^3.$$

Pro polynom stupně 3 je horní mez pásu spolehlivosti (ze vzorce 3.6)

$$g^*(x) = \begin{cases} y_1^* l_1(x) + y_{*2} l_2(x) + y_3^* l_3(x) + y_{*4} l_4(x) & x \le z_1, \\ y_1^* l_1(x) + y_2^* l_2(x) + y_{*3} l_3(x) + y_4^* l_4(x) & z_1 \le x \le z_2, \\ y_{*1} l_1(x) + y_2^* l_2(x) + y_3^* l_3(x) + y_{*4} l_4(x) & z_2 \le x \le z_3, \\ y_1^* l_1(x) + y_{*2} l_2(x) + y_3^* l_3(x) + y_4^* l_4(x) & z_3 \le x \le z_4, \\ y_{*1} l_1(x) + y_2^* l_2(x) + y_{*3} l_3(x) + y_4^* l_4(x) & z_4 \le x. \end{cases} \tag{3.10}$$

Dolní mez $g_*(x)$ získáme z (3.10) záměnnou y_i^* za y_{*i}, $i = 1, \ldots, 4$, popř. obráceně. Lagrangeův polynom (3.4) lze v R sestrojit pomocí funkce *poly.calc*(z,y) z knihovny *polynom*, kde $z = (z_1,\ z_2,\ z_3,\ z_4)$ a $y = (y_1,\ y_2,\ y_2,\ y_4)$, kde $y_i = y_i^*$ popř. $y_i = y_{*i}$ podle (3.10), $i = 1, \ldots, 4$. Výsledné meze pásu spolehlivosti jsou

$$g^* = \begin{cases} 7{,}56 - 3{,}53x + 0{,}75x^2 - 0{,}040x^3 & x \le 1{,}41 \\ 0{,}59 + 3{,}17x - 0{,}60x^2 + 0{,}033x^3 & 1{,}41 \le x \le 4{,}50 \\ -3{,}17 + 3{,}00x - 0{,}25x^2 + 0{,}004x^3 & 4{,}50 \le x \le 8{,}50 \\ 6{,}74 - 2{,}68x + 0{,}53x^2 - 0{,}025x^3 & 8{,}50 \le x \le 11{,}59 \\ -6{,}12 + 6{,}03x - 0{,}97x^2 + 0{,}048x^3 & 11{,}59 \le x; \end{cases}$$

$$g_* = \begin{cases} -6{,}12 + 6{,}03x - 0{,}97x^2 + 0{,}048x^3 & x \le 1{,}41 \\ 0{,}84 - 0{,}67x + 0{,}38x^2 - 0{,}025x^3 & 1{,}41 \le x \le 4{,}50 \\ 4{,}61 - 0{,}51x + 0{,}02x^2 + 0{,}004x^3 & 4{,}50 \le x \le 8{,}50 \\ -5{,}31 + 5{,}18x - 0{,}76x^2 + 0{,}033x^3 & 8{,}50 \le x \le 11{,}59 \\ 7{,}56 - 3{,}53x + 0{,}75x^2 - 0{,}040x^3 & 11{,}59 \le x. \end{cases}$$

Je-li rozptyl neznámý lze pro 4-rozměrné t rozdělení s 8 stupni volnosti v tabulce Hahn a Hendrickson (1971) vyhledat

$$c\sigma s^{-1} = 3{,}128$$

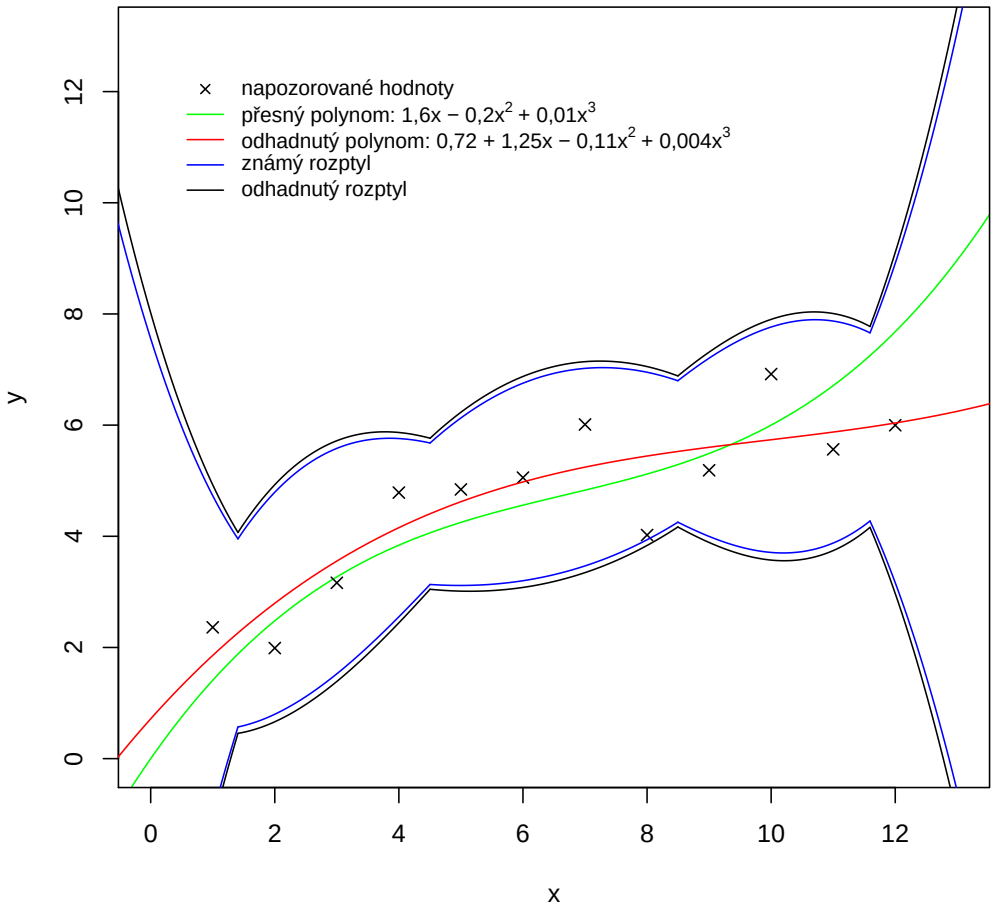
a tedy

$$c\sigma = 3{,}128 s = 2{,}66,$$

kde $s = \sqrt{s^2} = 0{,}85$.

Pak krajní hodnoty pásu v bodech z_1, z_2, z_3, z_4 jsou

$$\begin{aligned} y_{*i} &= \hat{Y}_{z_i} - c\sigma\sqrt{n^{-1}\lambda_i^{-1}}, & y_i^* &= \hat{Y}_{z_i} + c\sigma^2\sqrt{n^{-1}\lambda_i^{-1}}, \\ y_{*1} &= 0{,}45, & y_1^* &= 4{,}07; \\ y_{*2} &= 3{,}05, & y_2^* &= 5{,}76; \\ y_{*3} &= 4{,}17, & y_3^* &= 6{,}88; \\ y_{*4} &= 4{,}16, & y_5^* &= 7{,}77. \end{aligned}$$



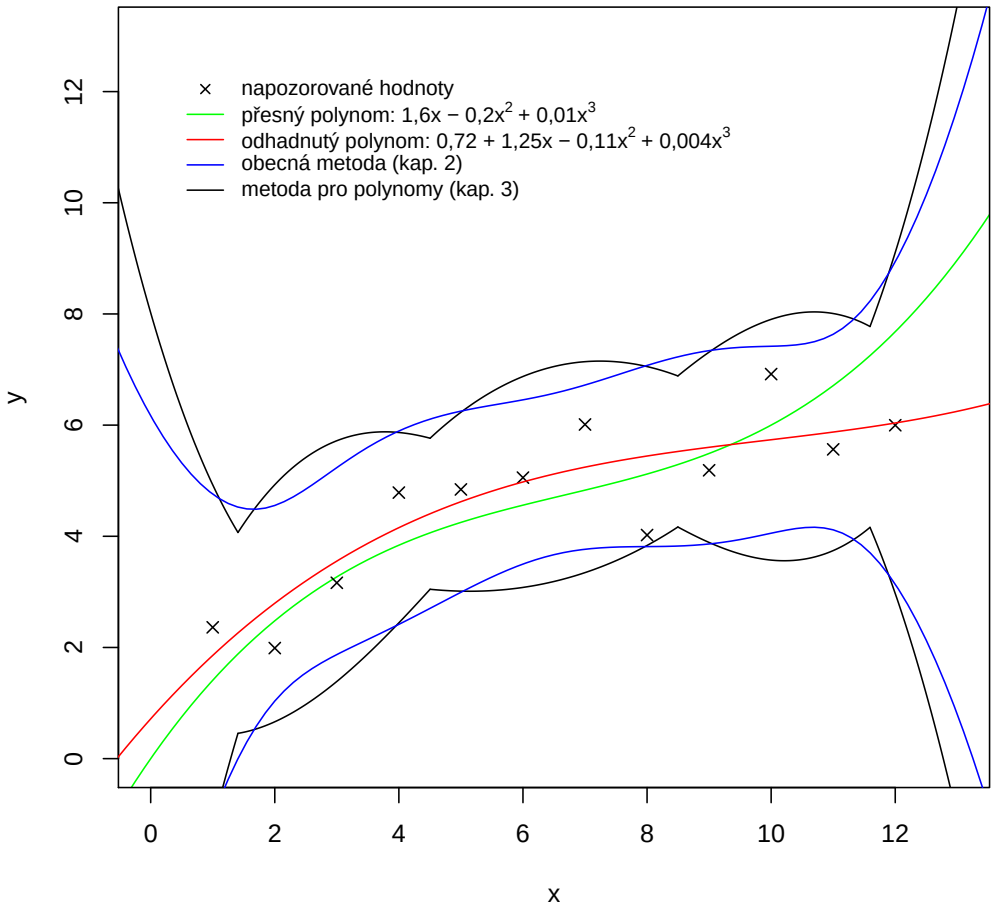
Obrázek 3.4: Porovnání pásů spolehlivosti pro polynom se známým a odhadnutým rozptylem.

Výsledné meze pásu spolehlivosti s odhadnutým rozptylem jsou

$$g^* = \begin{cases} 8{,}02 - 3{,}86x + 0{,}81x^2 - 0{,}043x^3 & x \leq 1{,}41 \\ 0{,}58 + 3{,}30x - 0{,}63x^2 + 0{,}035x^3 & 1{,}41 \leq x \leq 4{,}50 \\ -3{,}44 + 3{,}12x - 0{,}26x^2 + 0{,}004x^3 & 4{,}50 \leq x \leq 8{,}50 \\ 7{,}15 - 2{,}94x + 0{,}57x^2 - 0{,}027x^3 & 8{,}50 \leq x \leq 11{,}59 \\ -6{,}59 + 6{,}36x - 1{,}03x^2 + 0{,}051x^3 & 11{,}59 \leq x; \end{cases}$$

$$g_* = \begin{cases} -6{,}59 + 6{,}36x - 1{,}03x^2 + 0{,}051x^3 & x \leq 1{,}41 \\ 0{,}85 - 0{,}80x + 0{,}41x^2 - 0{,}027x^3 & 1{,}41 \leq x \leq 4{,}50 \\ 4{,}87 - 0{,}62x + 0{,}03x^2 + 0{,}004x^3 & 4{,}50 \leq x \leq 8{,}50 \\ -5{,}72 + 5{,}44x - 0{,}80x^2 + 0{,}035x^3 & 8{,}50 \leq x \leq 11{,}59 \\ 8{,}02 - 3{,}86x + 0{,}81x^2 - 0{,}043x^3 & 11{,}59 \leq x. \end{cases}$$

Pás spolehlivosti s odhadnutým rozptylem je širší než pás se známým rozptylem (viz obrázek 3.4).



Obrázek 3.5: Porovnání pásů spolehlivosti pro kubickou závislost.

Pás sestrojený postupem pouze pro polynomy (kapitola 3) je na okolí kvadraturních bodů $z_1, \ldots, z_4$ užší než pás spolehlivosti sestrojený postupem pro obecné funkce (část 2.3), mimo okolí těchto bodů je však širší (viz obrázek 3.5).

Závěr

V této práci jsme se seznámili se základními vlastnostmi lineárního modelu a zaměřili jsme se na sestrojení pásů spolehlivosti. Nejdříve jsme sestrojili interval spolehlivosti a predikční interval pro jediný pevný bod. Dále jsme sestrojili pás spolehlivosti dodržující pravděpodobnost pokrytí střední hodnoty lineárního modelu pro všechny body $\mathbb{R}$ současně. Tento pás vycházel z konfidenční množiny pro neznámé parametry regresního modelu. Výsledné meze vyšly stejné jako meze pásu sestrojeného Scheffého metodou, kterou popsal např. Naiman (1986).

V poslední části jsme popsali postup sestrojení přesného pásu spolehlivosti pro polynomický regresní model stupně p podle Wynn (1984). Tato metoda je založena na sestrojení přesných intervalů spolehlivosti pro střední hodnotu v tzv. kvadraturních bodech. Kvadraturní body nalezneme jako kořeny ortogonálního polynomu, stupně $p+1$, vzhledem k zadaným bodům. Výsledný pás je po částech polynomiální stupně p a prochází krajními body intervalů spolehlivosti střední hodnoty v kvadraturních bodech.

Všechny uvedené postupy jsou předvedeny na příkladu pro lineární závislost a pro polynom stupně 3. Pásy sestrojené Wynnovou metodou jsou na okolí kvadraturních bodů užší než pásy sestrojené Scheffého metodou, mimo tato okolí jsou však širší. Pro lineární závislost dává Wynnova metoda jednoduchý po částech lineární pás, avšak pro polynomy vyššího stupně může být tento pás nepřehledný.

Seznam použité literatury

Anděl, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Hahn, G. J. a Hendrickson, R. W. (1971). A table of percentage points of the distribution of the largest absolute value of k student t variates and its applications. *Biometrika*, **58**(2), 323–332.

Lin, S. (2016). Comparison of simultaneous confidence bands for univariete polyomial regression over an interval. *Communications in statistics: Theory and methods*, **45**(3), 589–598.

Naiman, D. Q. (1986). Conservative confidence bands in curvilinear regression. *The Annals of Statistics*, **14**(3), 896–906.

Spurier, J. D. (1993). Comparison of simultaneous confindence bands for quadratic regression over a finite interval. *Technometrics*, **35**(3), 315–320.

Stanovský, D. (2010). *Základy algebry*. První vydání. Matfyzpress, Praha. ISBN 978-80-7378-105-7.

Szegö, G. (1967). *Orthogonal Polynomials*, volume 23 of *American Math. Soc: Colloquium publ.* American Mathematical Soc. ISBN 978-0821810231.

Wynn, H. P. (1984). An exact confidence band for one-dimensional polynomial regression. *Biometrika*, **71**(2), 375–379.

Zvára, K. (2008). *Regrese*. První vydání. Matfyzpress, Praha. ISBN 978-80-7378-041-8.