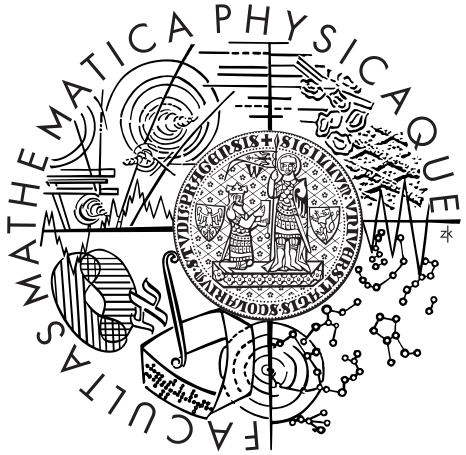
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Milan Benko

Regresní analýza a spliny

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	Mgr. Milan Bašta, Ph.D.
Studijní program:	Matematika
Studijní obor:	Finanční matematika

Praha 2015

Na tomto mieste by som rád poďakoval pánovi Mgr. Milanovi Baštovi, Ph.D. za jeho čas a cenné rady pri písaní tejto práce. Taktiež chcem poďakovať svojím priateľom a rodine za ich podporu.

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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne .....

Title: Regression analysis and splines

Author: Milan Benko

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Milan Bašta, Ph.D., University of Economics, Prague, Faculty of Informatics and Statistics, Department of Statistics and Probability

Abstract: The aim of this Bachelor's thesis is to introduce the basic concepts of regression analysis and subsequently regression splines as parametric models for regression function. I have looked upon the main characteristics of regression splines (coherence, coherence of derivations, the choice of placement and a number of knots). Further on in the thesis I have studied two bases as the examples of regression splines (truncated power basis and B-spline basis). I have also presented a model of natural cubic splines and a suitable basis for its representation has been derived. In the other part of my thesis I have looked upon the use of natural splines in order to increases the appraisal precision of regression function, mean square error formula has been derived and I have been trying to find out and illustrate under what conditions the use of natural splines is applicable. The thesis is complemented with a Monte Carlo Simulation, contextualized into models of splines. The results show that the criteria commonly used for the choice of a model ($\mathbb{R}^2_{adj}$, $PRESS$ statistic, hypothesis testing) do not always enable us to choose the right model in order to achieve the greatest precision of the estimation of regression function. All the calculations are done in R software and are in the electronic attachment.

Keywords: regression model, least squares method, regression splines, natural splines, mean square error

Názov práce: Regresná analýza a spliny

Autor: Milan Benko

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: Mgr. Milan Bašta, Ph.D., VŠE Praha, Fakulta informatiky a statistiky, Katedra statistiky a pravděpodobnosti

Abstrakt:

Cieľom bakalárskej práce je predstaviť základné pojmy regresnej analýzy a následne regresné spliny ako parametrické modely pre regresnú funkciu. Sú diskutované základné vlastnosti regresných splinov (spojitosť, spojitosť derivácií, voľba polohy a počtu uzlových bodov). Ďalej sú predstavené dve bázy vhodné pre reprezentáciu regresných splinov (useknutá mocninová báza a B-spliny). Taktiež je predstavený model prirodzených (kubických) splinov a je odvodená vhodná báza pre jeho reprezentáciu. Následne je diskutované použitie prirodzených splinov za účelom zvýšenia presnosti odhadu regresnej funkcie, sú odvodené vzorce pre strednú štvorcovú chybu odhadu a je kvalitatívne diskutované a ilustrované, za akých okolností je použitie prirodzených splinov vhodné. Práca je doplnená Monte Carlo simuláciou, zasadenou do kontextu modelov splinov, ktorej výsledky naznačujú, že v praxi bežne používané kritéria pre výber modelu ($\mathbb{R}^2_{adj}$, $PRESS$ štatistika, test hypotézy) neposkytujú vždy správne rozhodnutie, aký model je skutočne optimálny z hľadiska presnosti odhadu regresnej funkcie. Všetky výpočty sú prevedené v softvéri R a sú k dispozícii v elektronickej prílohe.

Kľúčové slová: regresný lineárny model, metóda najmenších štvorcov, regresné spliny, prirodzené spliny, stredná štvorcová chyba

Obsah

Kapitola 1

Úvod

Mnoho problémov, či už v technických alebo ekonomických odvetviach, zahŕňa skúmanie vzťahu, teda závislosti alebo nezávislosti medzi dvoma alebo viacerými premennými. Riešením práve týchto problémov sa zaoberá regresná analýza.

Regresná analýza je dôležitým štatistickým nástrojom zahrňujúci množstvo metód, pomocou ktorých odhadujeme mimo iné aj strednú hodnotu nejakej náhodnej veličiny Y. Túto náhodnú veličinu nazývame *odozva* alebo *závislá premenná*. Jej strednú hodnotu odhadujeme na základe jednej alebo viacerých spojitých či diskrétnych veličín nazývaných *regresory*, *vysvetľujúce* alebo *nezávislé premenné*. Pokiaľ skúmame závislosť strednej hodnoty závislej premennej na hodnotách jednej nezávislej premennej, hovoríme o tzv. *jednoduchej lineárnej regresii*, kdežto pri skúmaní závislosti na hodnotách viacerých nezávislých premenných ide o *viacnásobnú regresnú analýzu*.

Závislosť medzi strednou hodnotou závislej a danými nezávislými premennými bude popisovať funkcia, ktorú nazývame *regresná funkcia*. Krivka určujúca parametrický predpis regresnej funkcie môže nadobúdať veľa rôznych podôb, z ktorých sa primárne budeme venovať regresnej funkcii splinov[1].

V úvode kapitoly 2 si predstavíme jeden z najjednoduchších regresných modelov - regresný model priamky. Z názvu je zrejmé, že vzťah popisujúci závislosť strednej hodnoty závislej a nezávislej premennej bude lineárny, teda regresnou funkciou bude v tomto prípade (regresná) priamka. Práve na tomto modeli si ukážeme ako z dát, ktoré tvorí súbor n pozorovaní spomínaných premenných, vypočítame odhady *regresných koeficientov* určujúcich tvar odhadnutej regresnej priamky. Odhady vypočítame *metódou najmenších štvorcov*, ktorú si bližšie popíšeme v sekcii 2.3. Ďalej si v kapitole 2 zadefinujeme lineárny model viacnásobnej regresie. Následne si v sekcii 2.4 zadefinujeme podmodel a ukážeme si dva prístupy ako ho môžeme z modelu získať. Ďalej si v danej sekcii zadefinujeme *F-test*, ktorým skúmame prípadnú platnosť podmodelu. Záver kapitoly 2 je venovaný rôznym kritériám pre výber regresného modelu.

V kapitole 3 si predstavíme regresné spliny. Popíšeme si ich základné vlastnosti a následne si predstavíme dva rôzne systémy bázových funkcií, ktorých lineárnou kombináciou získame regresnú funkciu regresného splinu. Následne budeme v sekcii 3.5 rozoberať rôzne prístupy voľby polohy a počtu uzlových bodov. Ďalej si predstavíme model prirodzených splinov a odvodíme si vhodnú bázu pre jeho reprezentáciu. V sekcii 3.7 budeme diskutovať, za akých okolností je vhodné použiť

[1] čítaj splajnov

prirodzené spliny za účelom zvýšenia presnosti odhadu regresnej funkcie. V závere práce budeme Monte Carlo simuláciou kvalitatívne skúmať, či sa na základe v praxi bežne používaných kritérií pre výber regresného modelu môžeme správne rozhodnúť, aký model je z hľadiska presnosti odhadu regresnej funkcie skutočne optimálny.

Kapitola 2

Model lineárnej regresie

Model lineárnej regresie môžeme chápať ako predpis, ktorý nám bližšie popisuje vzťah medzi závislou a nezávislou premennou. Definíciu modelu viacnásobnej lineárnej regresie môžeme nájsť v sekcii 2.2, no pre názornosť a lepšie pochopenie si postupy a metódy vyšetrovania vzťahu medzi strednou hodnotou náhodnej veličiny Y a hodnotami jednej nenáhodnej veličiny, ukážeme na príklade regresného modelu priamky. V tomto kontexte budeme uvažovať, že naša nezávislá premenná je nejaká *spojitá* nenáhodná veličina, ktorú budeme značiť x.

2.1 Príklad regresného modelu priamky

Ako sme už naznačili v úvode kapitoly, predpokladáme, že regresná funkcia popisujúca vzťah medzi premennými bude priamka. Naše dáta budú pozostávať z n pozorovaní $(Y_i, x_i), i = 1,...,n$, kde Y_i predstavuje i-tú hodnotu závislej premennej a x_i bude zase predstavovať i-tú hodnotu nezávislej premennej. Regresný model priamky môžeme zapísať v tvare

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{2.1}$$

ktorý môžeme interpretovať tak, že pre dané konštanty x_i pozostávajú odpovedajúce hodnoty Y_i z hodnoty $\beta_0 + \beta_1 x_i$ a nejakého náhodného vychýlenia označeného ϵ_i. Aby tento model spĺňal predpoklady lineárneho regresného modelu (Def. 1), musí mimo iné platiť $\mathbb{E}\,\epsilon_i = 0$ pre každé $i = 1,...,n$. Potom z rovnice (2.1) vyplýva

$$\mathbb{E}\,Y_i = \beta_0 + \beta_1 x_i. \tag{2.2}$$

Rovnica (2.2) nám udáva hodnotu regresnej funkcie v i-tom pozorovaní. Neznáme parametre β_0 a β_1, nazývané *regresné koeficienty*, predstavujú *prienik* priamky s osou y a *smernicu* našej regresnej priamky. Tieto parametre sa budeme snažiť z našich dát čo najlepšie odhadnúť.

Odhady, či už koeficientov β_0 a β_1, alebo strednej hodnoty odozvy Y_i, budeme značiť symbolom striešky, tj. $\hat{\beta}_0$, $\hat{\beta}_1$ alebo $\hat{Y}_i$. Odhad $\hat{Y}_i$, ktorý nazývame i-tou *vyrovnanou hodnotou*, môžeme zapísať v tvare

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{2.3}$$

Odhady $\hat{\beta}_0$ a $\hat{\beta}_1$ budeme hľadať pomocou *metódy najmenších štvorcov*, bližšie popísanú v sekcii 2.3. Touto metódou budeme minimalizovať sumu štvorcov chyby,

ktorú si označíme SS_ϵ (z angl. sum of squares), určenú predpisom

$$SS_\epsilon = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2, \tag{2.4}$$

pričom výraz (2.4) je funkciou v neznámych parametroch β_0 a β_1. Ako sme už naznačili, chceme nájsť také hodnoty parametrov β_0 a β_1, v ktorých je (2.4) minimálna, teda pôjde o problém hľadania minima funkcie dvoch premenných. Hodnoty, v ktorých je (2.4) minimálna, sú práve odhady metódou najmenších štvorcov.

Výraz (2.4) si najprv zderivujeme podľa β_0 a β_1

$$\begin{aligned} \frac{\partial SS_\epsilon}{\partial \beta_0} &= 2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)(-1), \\ \frac{\partial SS_\epsilon}{\partial \beta_1} &= 2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)(-x_i). \end{aligned} \tag{2.5}$$

Následne si derivácie (2.5) položíme rovné nule, čím získame systém rovníc, ktorý nazývame *systémom normálnych rovníc.* Po vydelení týchto rovníc číslom (-2) a rozvedení súm dostane naše hľadané odhady $\hat{\beta}_0$ a $\hat{\beta}_1$ ako

$$\begin{aligned} \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i &= 0, \\ \sum_{i=1}^{n} Y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= 0. \end{aligned} \tag{2.6}$$

Systém rovníc (2.6) má práve jedno riešenie, ak aspoň dve z hodnôt $x_i, i =$ 1,...,n sú navzájom rôzne. Je zrejmé, že minimalizovaná funkcia (2.4) je konvexná (viď sekciu 2.3), riešenie je teda skutočne globálnym minimom.

Riešením rovníc (2.6) dostaneme odhady $\hat{\beta}_0$ a $\hat{\beta}_1$ v tvare

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{Y} - \hat{\beta}_1 \bar{x} \tag{2.7}$$

a

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i Y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \ \frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2}, \tag{2.8}$$

kde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ sú aritmetické priemery.

V praxi sa však tento postup výpočtu odhadov regresných koeficientov, tj. dosadením do (2.7) a (2.8), nepoužíva. Vhodnejšou alternatívou je použitie vbudovanej funkcie $lm()$ v softvéri R. Ďalej v práci budeme dopočítavať odhady práve využitím tejto funkcie.

Regresný model priamky nemusí byť v niektorých prípadoch dosť flexibilný na to, aby dostatočne dobre popísal vzťah závislosti strednej hodnoty Y na x. V praxi je preto bežné tento vzťah popisovať flexibilnejšími funkciami, ako napríklad kvadratickým alebo kubickým polynómom, či regresnými splinami, ktoré sú špeciálnymi prípadmi lineárneho modelu viacnásobnej regresie. Ten si zadefinujeme v nasledujúcej sekcii 2.2. Spomenutými špeciálnymi prípadmi regresných funkcií (a mnohými ďalšími) sa potom budeme zaoberať v kapitole 3.

2.2 Model viacnásobnej lineárnej regresie

Model viacnásobnej lineárnej regresie sa od modelu regresnej priamky líši predpokladom, že pre každú hodnotu náhodnej veličiny Y_i, $i = 1, \ldots, n$ je pozorovaných obecne k rôznych hodnôt známych nenáhodných konštánt $x_{ij}, j = 1, \ldots, k$, tj. nezávislých premenných, ktoré nejakým spôsobom ovplyvňujú strednú hodnotu veličín $Y_1, \ldots, Y_n$.

Predpokladajme, že vzťah medzi strednou hodnotou $Y_i, i = 1, \ldots, n$ a danými hodnotami nenáhodných konštánt môžeme popísať nasledovne

$$\mathbb{E}\, Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik},\ i = 1, \ldots, n, \tag{2.9}$$

pričom $\beta_j, j = 0, \ldots, k$, sú neznáme parametre, tj. regresné koeficienty. Známe konštanty vyskytujúce sa v rovnici (2.9) je vhodné pre ich ďalšie použitie usporiadať do matice, ktorú si označíme $\mathbf{X}$ a nazveme *regresnou maticou* alebo *maticou modelu,*

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}. \tag{2.10}$$

Poznámka. V práci budeme uvažovať, že stĺpce regresnej matice $\mathbf{X}$ sú lineárne nezávislé, teda matica $\mathbf{X}$ bude mať plnú hodnosť $h(\mathbf{X}) = k + 1$.

Prvý stĺpec regresnej matice $\mathbf{X}$, odpovedajúci jednotkovému vektoru, prislúcha regresnému parametru β_0, ktorý nazývame *absolútny člen.* Tento člen nám udáva očakávanú hodnotu veličiny Y v prípade, ak by všetky vysvetľujúce premenné boli rovné nule. Naďalej v práci budeme uvažovať, že pojmom *lineárny regresný model* (viď Def. 1) myslíme práve *lineárny regresný model s absolútnym členom.* Pojem *lineárny* zdôrazňuje, že vzťah popisujúci závislosť medzi závislou a nezávislými premennými je lineárny v regresných koeficientoch.

Poznámka. Konštanty x_{ij} nemusia obsahovať iba hodnoty pozorovaní, ale môžu nadobúdať tvar stransformovaných pôvodných porovaní, napr. transformáciou jednej nezávislej premennej x_i v tvare $x_{i1} = x_i$ a $x_{i2} = x_i^2$ sa v Kapitole 3 dostaneme k predpisu modelu kvadratickej regresie. Stransformované konštanty x_{ij} zvykneme nazývať *regresory.*

Nasledujúca definícia lineárneho regresného modelu je sformulovaná podľa [3].

Definícia 1. *Nech* $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ *je náhodný vektor* $n \times 1$ *a nech* $\mathbf{X}_{n\times(k+1)}$ *je matica známych reálnych konštánt určená predpisom* (2.10)*. Nech* $n > k + 1$ *a* $h(\mathbf{X}) = k + 1$*. Povieme, že* $\mathbf{Y}$ *spolu s maticou modelu* $\mathbf{X}$ *spĺňa lineárny regresný model, ak pre vektor neznámych regresných parametrov* $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ *platí*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \tag{2.11}$$

kde $\epsilon_i, i = 1, \ldots, n$*, ktoré nazveme chybové členy, sú zložky náhodného vektora* $\boldsymbol{\epsilon}$*, pre ktorý platí* $\mathbb{E}\, \boldsymbol{\epsilon} = \boldsymbol{0}$ *a* $\mathsf{var}\, \boldsymbol{\epsilon} = \sigma^2 \mathbf{I}_n$*, pričom* $\sigma^2 > 0$ *je ďalší neznámy parameter.*

Z Definície 1 môžeme vyvodzovať, že náhodný vektor $\mathbf{Y}$ má strednú hodnotu $\mathbf{X}\boldsymbol{\beta}$ a variačnú maticu $\sigma^2 \mathbf{I}_n$, ktorou predpokladáme rovnaký rozptyl a nekorelovanosť jednotlivých zložiek náhodného vektora $\mathbf{Y}$. Lineárny (regresný) model

budeme zapisovať v tvare $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$ alebo ho prípadne môžeme zapísať vektorovo ako $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, kde $\boldsymbol{\epsilon} \sim (\mathbf{0},\, \sigma^2\mathbf{I}_n)$.

Ak by sme špeciálne uvažovali, že vektor $\mathbf{Y}$ má mnohorozmerné normálne rozdelenie, potom model $\mathbf{Y} \sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ budeme nazývať *normálny lineárny model.* V tomto modeli platí $\boldsymbol{\epsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Uvažovaním daného predpokladu si v sekcii 2.3 ukážeme základne vlastnosti a rozdelenie štatistík týkajúcich sa (nielen) odhadu vektora regresných koeficientov $\boldsymbol{\beta}$, ktorý získame metódou najmenších štvorcov.

2.3 Metóda najmenších štvorcov

Metóda najmenších štvorcov (MNŠ) patrí medzi najpoužívanejšie metódy na výpočet odhadu regresných koeficientov. Označme $\hat{\boldsymbol{\beta}}$ vektor odhadnutých regresných koeficientov $\hat{\beta}_j$, $j = 0, \ldots, k$ a rovnako ako v sekcii 2.1, nám hodnoty vektora $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, ktorý nazývame vektorom vyrovnaných hodnôt, budú predstavovať odhadnuté stredné hodnoty závislej premennej. Jednotlivé zložky vektora $\hat{\mathbf{Y}}$ sú dané nasledovne

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}, i = 1, \ldots, n. \tag{2.12}$$

Vzťah (2.12) môžeme vektorovo zapísať ako $\hat{Y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, i = 1, \ldots, n$, kde vektor $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})^\top, i = 1, \ldots, n$ je i-tým riadkom matice $\mathbf{X}$. Predtým, ako si ukážeme výpočet vektora $\hat{\boldsymbol{\beta}}$, uvedieme si definíciu sformulovanú podľa [1], ktorá s metódou najmenších štvorcov úzko súvisí.

Definícia 2. *Náhodný vektor $\mathbf{u} = \mathbf{Y} - \hat{\mathbf{Y}}$ budeme nazývať vektorom rezíduí a jeho jednotlivé zložky určujúce vzdialenosť medzi Y_i a $\hat{Y}_i$, tj. $u_i = Y_i - \hat{Y}_i = Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ budeme nazývať rezíduum. Ďalej, náhodnú veličinu*

$$RSS = \mathbf{u}^\top \mathbf{u} = \|\mathbf{u}\|^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

budeme nazývať reziduálny súčet štvorcov.

Vektor $\hat{\boldsymbol{\beta}}$ budeme hľadať tak, aby *RSS*, teda vzdialenosť medzi vektormi $\mathbf{Y}$ a $\hat{\mathbf{Y}}$, bola v euklidovskom priestore čo najmenšia. Podľa [4] je

$$\hat{\boldsymbol{\beta}} = arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.13}$$

Maticové derivácie (2.13) podľa vektoru $\boldsymbol{\beta}$

$$\frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = -\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{X} = -2(\mathbf{X}^\top \mathbf{Y} - \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}),$$

položíme rovné nule, čím získame systém $k+1$ lineárnych rovníc. Vektor $\hat{\boldsymbol{\beta}}$ získame riešením systému (normálnych) rovníc (2.14)

$$(\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y}. \tag{2.14}$$

Minimalizovaná funkcia (2.13) je konvexná, keďže jej druhá derivácia podľa vektoru $\boldsymbol{\beta}$ je rovná $2\mathbf{X}^\top \mathbf{X}$, čo je pozitívne semidefinitná matica (viď Tvrd. 1). Riešenie sústavy (2.14) bude preto globálne minimum.

Definícia 3. *Symetrická matica* $\mathbf{A}$ *je pozitívne definitná, ak pre každý vektor* $\mathbf{y} \neq \boldsymbol{0}$ *je* $\mathbf{y}^\top \mathbf{A}\mathbf{y} > 0$, *([6, str. 209]). Štvorcová matica* $\boldsymbol{A}$ *je symetrická ak* $\boldsymbol{A} = \boldsymbol{A}^\top$,*([7, str. 45]).*

Tvrdenie 1. *Matica* $\mathbf{X}^\top\mathbf{X}$ *je pozitívne definitná.*

Dôkaz. Pre $\mathbf{X}^\top\mathbf{X}$ platí

$$(\mathbf{X}^\top\mathbf{X})^\top = \mathbf{X}^\top\mathbf{X},$$

teda matica je symetrická. Uvažujme nejaký vektor $\mathbf{y} \neq \mathbf{0}$. Výraz

$$\mathbf{y}^\top\mathbf{X}^\top\mathbf{X}\mathbf{y} = (\mathbf{X}\mathbf{y})^\top(\mathbf{X}\mathbf{y}), \tag{2.15}$$

je suma štvorcov, teda je nezáporný. Z lineárnej nezávislosti stĺpcov matice $\mathbf{X}$ vyplýva, že $\mathbf{X}\mathbf{y} \neq \mathbf{0}$, (pre $\mathbf{y} \neq \mathbf{0}$), viď [7, str. 25], teda výraz 2.15 je kladný. ❑

Ak predpokladáme lineárnu nezávislosť stĺpcov matice $\mathbf{X}$, štvorcová matica $\mathbf{X}^\top\mathbf{X}$ je regulárna, a preto existuje práve jedno riešenie danej sústavy (2.14), a to

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

Vektor $\hat{\mathbf{Y}}$ môžeme písať v tvare $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, pričom maticu $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ si označíme $\mathbf{H}$, takže platí $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$.

Poznámka. Označenie matice $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ písmenom $\mathbf{H}$ je zaužívané v regresnej analýze podľa anglického slova *hat*, tj. strieška.

Tvrdenie 2. *Matica* $\boldsymbol{H}$ *je idempotentná (platí* $\boldsymbol{HH} = \boldsymbol{H}$*) a platí* $\boldsymbol{H}\mathbf{X} = \mathbf{X}$.

Dôkaz. Je zrejmý dosadením $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

❑

Vlastnosti odhadov získaných MNŠ

V nasledujúcich vetách si ukážeme základné vlastnosti odhadu vektora regresných koeficientov $\hat{\boldsymbol{\beta}}$ získaného metódou najmenších štvorcov, ako aj odhadu vektora strednej hodnoty odozvy $\hat{\mathbf{Y}}$.

Veta 3. *Nech* $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$. *Pre* $\hat{\boldsymbol{\beta}}$ *platí* $\mathbb{E}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\mathsf{var}\,\hat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

Dôkaz. Viď [3, str. 82]. ❑

Poznámka. Pre naše ďalšie potreby si zavedieme maticu $\mathbf{V} = (\mathbf{X}^\top\mathbf{X})^{-1}$ a jej jednotlivé zložky si označíme $v_{ij}, 0 \leq i,j \leq k$.

Veta 4. *Nech* $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$. *Pre* $\hat{\mathbf{Y}}$ *platí* $\mathbb{E}\,\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$, $\mathsf{var}\,\hat{\mathbf{Y}} = \sigma^2\mathbf{H}$.

Dôkaz. S využitím vety 3 môžeme dôkaz zjednodušiť nasledovne

$$\mathbb{E}\,\hat{\mathbf{Y}} = \mathbb{E}\,\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbb{E}\,\hat{\boldsymbol{\beta}} = \mathbf{X}\boldsymbol{\beta},$$

$$\mathsf{var}\,\hat{\mathbf{Y}} = \mathsf{var}\,(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}\,\mathsf{var}\,\hat{\boldsymbol{\beta}}\,\mathbf{X}^\top = \mathbf{X}\,\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\,\mathbf{X}^\top = \sigma^2\mathbf{H}.$$

❑

Dôsledok 1. Metódou najmenších štvorcov sme získali nestranné odhady vektoru $\boldsymbol{\beta}$ a $\mathbb{E}\mathbf{Y}$.

Uvažovaním normálneho lineárneho modelu $\mathbf{Y} \sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ môžeme podľa [1] sformulovať ďalšie vety, prostredníctvom ktorých si určíme rozdelenia jednotlivých štatistík.

Veta 5. *Nech* $\mathbf{Y} \sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. *Potom*

$$(i)\quad \hat{\mathbf{Y}} \sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n), \tag{2.16}$$

$$(ii)\ \mathbf{u} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})), \tag{2.17}$$

$$(iii)\frac{RSS}{\sigma^2} \sim \chi^2_{n-k-1}, \tag{2.18}$$

kde χ^2_{n-k-1} *je chí-kvadrát rozdelenie s* $n-k-1$ *stupňami voľnosti.*

Dôkaz. Viď [2, str. 62].

❑

Teraz si v nasledujúcej sekcii 2.4 zadefinujeme ďalší dôležitý pojem lineárnej regresie, a to podmodel. Následne si ukážeme dva spôsoby získania podmodelu, a to vypustenením stĺpcov regresnej matice $\mathbf{X}$ a zavedením lineárnych obmedzení na regresné koeficientu lineárneho regresného modelu.

2.4 Podmodel

V praxi sa často vyžaduje, aby model popisujúci vzťah medzi závislou premennou a regresormi bol čo najjednoduchší. Hľadáme teda taký model, ktorý zmenšuje priestor $\mathcal{M}(\mathbf{X})$ (viď *Poznámku* nižšie) všetkých možných stredných hodnôt náhodného vektoru $\mathbf{Y}$. Tento model nazývame *podmodel*. Uveďme si jeho definíciu sformulovanú podľa [1].

Poznámka. Symbolom $\mathcal{M}(\mathbf{X})$ budeme značiť lineárny obal stĺpcov matice $\mathbf{X}$. Tento lineárny priestor, ktorý je tvorený všetkými lineárnymi kombináciami stĺpcov matice $\mathbf{X}$, nazývame *regresným priestorom.* Je zrejmé, že $\mathbb{E}\mathbf{Y} \in \mathcal{M}(\mathbf{X})$, keďže $\mathbb{E}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$.

Definícia 4. *Povieme, že platí podmodel modelu* $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, *ak pre nejaký vektor* $\boldsymbol{\beta}_0$ *a maticu (nenáhodných) konštánt* $\mathbf{X}_0$, *ktorá spľňa* $\mathcal{M}(\mathbf{X}_0) \subset \mathcal{M}(\mathbf{X}), 0 < h(\mathbf{X}_0) = r_0 < k+1$, *platí* $\mathbb{E}\,\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0$.

Porovnanie modelu a podmodelu patrí medzi jednu z najdôležitejších metód v regresnej analýze. Ako môžeme vidieť vo vete 6, porovnávanie modelov je založené na rozdiely medzi reziduálnym súčtom štvorcov v modeli, ktorý budeme značiť štandardne (podľa Def. 2) RSS a reziduálnym súčtom štvorcov v podmodeli, ktorý budeme značiť RSS_0. Pre odvodenie tvaru testovej štatistiky o prechode k podmodelu uvažujme opäť normálny lineárny model.

Veta 6 (F-test)**.** *Ak platí v normálnom lineárnom modeli* $\mathbf{Y} \sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *podmodel, potom*

$$F_0 = \frac{n-k-1}{k+1-r_0}\frac{RSS_0 - RSS}{RSS} \sim F_{k+1-r_0,n-k-1}, \tag{2.19}$$

kde $F_{k+1-r_0,n-k-1}$ *predstavuje Fisherovo rozdelenie s* $k+1-r_0$ *a* $n-k-1$ *stupňami voľnosti.*

Dôkaz. Viď [1, str. 31].

❑

Štatistika F_0 z vety 6 je testovou štatistikou tzv. F-testu, ktorého nulová hypotéza je $H_0 : \mathbb{E}\mathbf{Y} \in \mathcal{M}(\mathbf{X}_0)$ a alternatíva $H_1 : \mathbb{E}\mathbf{Y} \notin \mathcal{M}(\mathbf{X}_0)$. Ak nulovú hypotézu H_0 nezamietame v prospech alternatívy, znamená to, že platí podmodel $\mathbf{Y} \sim \mathsf{N}_n(\mathbf{X}_0\boldsymbol{\beta}_0, \sigma^2\mathbf{I}_n)$. H_0 zamietame, ak $F_0 \geq F_{k+1-r_0,n-k-1}(1-\alpha)$, kde $F_{k+1-r_0,n-k-1}(1-\alpha)$ značí $(1-\alpha)$-kvantil rozdelenia $F_{k+1-r_0,n-k-1}$.

V tejto práci si ukážeme dva prístupy získania podmodelu, a to vypustením stĺpcov matice $\mathbf{X}$ a zavedením lineárnych obmedzení na vektor $\boldsymbol{\beta}$. Druhý zo spomenutých spôsobov získania podmodelu budeme následne využívať v kapitole 3 napr. pri získavaní podmodelu regresných splinov, a to model *prirodzených splinov*, (viď sekciu 3.6).

2.4.1 Vypustenie stĺpcov matice X

Prvý, jednoduchší postup ako od modelu prejsť k jeho podmodelu, je vypustením stĺpcov regresnej matice $\mathbf{X}$. Predpokladajme rozklad matice $\mathbf{X} = (\mathbf{X}_0|\mathbf{X}_1)$ modelu $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ a k nemu odpovedajúci rozklad vektoru $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top)^\top, \boldsymbol{\beta}_0 \in \mathbb{R}^{r_0}$ a $\boldsymbol{\beta}_1 \in \mathbb{R}^{k+1-r_0}$. Náš model $\mathbf{Y}$ môžeme písať v tvare

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}. \tag{2.20}$$

Ak si označíme RSS_0 reziduálny súčet štvorcov v podmodeli $\mathbf{Y} \sim \mathsf{N}_n(\mathbf{X}_0\boldsymbol{\beta}_0, \sigma^2\mathbf{I}_n)$, môžeme (uvažovaním $\boldsymbol{\epsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$) štatistikou (2.19) a hypotézou $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ testovať prechod od modelu (2.20) k spomínanému podmodelu.

2.4.2 Lineárne obmedzenia na regresné koeficienty

Druhý postup, ako môžeme získať podmodel, je zavedením lineárnych obmedzení na vektor regresných koeficientov $\boldsymbol{\beta}$. Lineárne obmedzenia môžeme vyjadriť v tvare $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$, pričom matica $\mathbf{T}$ má rozmery $d \times (k+1)$ a vektor konštánt $\mathbf{c}$ má rozmery $d \times 1$.

Poznámka. Hodnota d nám určuje počet obmedzení aplikovaných na regresné koeficienty

Teraz si uveďme vetu, sformulovanú podľa [1] o tom, že dané lineárne obmedzenia nám skutočne určujú podmodel.

Veta 7. *Nech matica $\boldsymbol{T}_{d\times(k+1)}$ má lineárne nezávislé riadky. Ďalej nech platí $0 < d < k+1 = h(\mathbf{X})$ a súčasne $\mathcal{M}(\boldsymbol{T}^\top) \subset \mathcal{M}(\mathbf{X}^\top)$. Potom konzistentná sústava lineárnych rovníc*

$$\boldsymbol{T}\boldsymbol{\beta} = \mathbf{c}$$

určuje podmodel modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ dimenzie $r_0 = k + 1 - d$.

Dôkaz. Viď [1, str. 33].

❑

Odhad strednej hodnoty $\mathbf{Y}$ budeme opäť hľadať metódou najmenších štvorcov, teda minimalizovaním $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$ s tým, že zároveň platí $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$. Vektor $\hat{\mathbf{Y}}_0 = \mathbf{X}\hat{\boldsymbol{\beta}}_0$, ktorý predstavuje odhad $\mathbb{E}\mathbf{Y}$ v podmodeli, budeme hľadať tak, aby vzdialenosť medzi vektormi $\hat{\mathbf{Y}}_0$ a $\mathbf{Y}$ bola v priestore $\mathcal{M}(\mathbf{X})$ čo najmenšia. Vektor $\hat{\boldsymbol{\beta}}_0$, teda vektor odhadnutých regresných koeficientov v podmodeli, ktorý spĺňa $\mathbf{T}\hat{\boldsymbol{\beta}}_0 = \mathbf{c}$, získame pomocou metódy Langrageových multiplikátorov, pričom minimalizovanú funkciu uvažujeme v tvare

$$\Phi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{\lambda}^\top(\mathbf{T}\boldsymbol{\beta} - \mathbf{c}),$$

kde zložky vektora $\boldsymbol{\lambda}_{d\times 1}$ sú Langrangeové multiplikátory. Postup riešenia daného minimalizujúceho problému môže čitateľ nájsť v [1, str. 33], odkiaľ je prevzatý nasledujúci tvar vektoru $\hat{\boldsymbol{\beta}}_0$ a $\mathsf{var}\hat{\boldsymbol{\beta}}_0$ nasledovne

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{T}^T(\mathbf{T}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{T}^T)^{-1}(\mathbf{T}\hat{\boldsymbol{\beta}} - \mathbf{c}), \tag{2.21}$$

$$\mathsf{var}\hat{\boldsymbol{\beta}}_0 = \sigma^2\left((\mathbf{X}^\top\mathbf{X})^{-1} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{T}^\top\left(\mathbf{T}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{T}^\top\right)^{-1}\mathbf{T}(\mathbf{X}^\top\mathbf{X})^{-1}\right). \tag{2.22}$$

Na záver tejto kapitoly si povieme niečo o predpovediach. Táto téma tvorí základ regresnej analýzy, keďže často chceme zistiť predpoveď individuálnej alebo odhad strednej hodnoty závislej premennej v nejakom novom, nami doteraz nepozorovanom bode. Tejto téme ako aj

V sekcii 2.5 si zadefinujeme tvar predpovede individuálnej a odhadu strednej hodnoty závislej premennej ako aj ukazovatele toho, ako dobre prekladá dáta odhadnutá regresná funkcia lineárneho modelu.

2.5 Ďalšie pojmy regresnej analýzy

2.5.1 Odhad strednej hodnoty závislej premennej v novom bode

Pri pojme predpoveď sa nám často vybaví odhadovanie budúcej hodnoty závislej premennej na základe nami už pozorovaných hodnotách regresorov. V

kontexte lineárnej regresie sa predpoveďou myslí odhad hodnoty náhodnej veličiny Y v nejakom novom bode $\mathbf{x}_* = (1, x_{*1}, \ldots, x_{*k})^\top$, pričom uvažujeme $\mathbf{x}_* \in \mathcal{M}(\mathbf{X}^\top)$ a regresory $\mathbf{x}_{*j}, j = 1, \ldots, k$ sú opäť nenáhodné konštanty. Nasledujúca teória sa opiera o výklad knihy [5, Sek. 3.6.5].

Uvažujme model $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. Predpoveď individuálnej hodnoty závislej premennej v bode $\mathbf{x}_*$ uvažujeme v tvare

$$Y_* = \beta_0 + \beta_1 x_{*1} + \ldots + \beta_k x_{*k} + \epsilon_*, \tag{2.23}$$

kde chyba ϵ_* je nezávislá so zložkami vektora $\boldsymbol{\epsilon}$ lineárneho regresného modelu. Predpokladajme, že $\mathbb{E}\epsilon_* = 0$ a $\mathsf{var}\epsilon_* = \sigma^2$. Pre (teoretickú) strednú hodnotu Y_* (v bode $\mathbf{x}_*$) potom platí

$$\mathbb{E}Y_* = \mathbf{x}_*^\top \boldsymbol{\beta}. \tag{2.24}$$

Odhad výrazu (2.24) uvažujeme podľa [5] v tvare

$$\hat{Y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_{*1} + \ldots + \hat{\beta}_k x_{*k} = \mathbf{x}_*^\top \hat{\boldsymbol{\beta}}. \tag{2.25}$$

Tvrdenie 8. *Pre $\hat{Y}_*$ platí $\mathbb{E}\hat{Y}_* = \mathbf{x}_*^\top\boldsymbol{\beta}$ a $\mathsf{var}\hat{Y}_* = \sigma^2\mathbf{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_*$.*

Dôkaz. Platnosť $\mathbb{E}\hat{Y}_* = \mathbf{x}_*^\top\boldsymbol{\beta}$ je využitím Vety 3 zrejmá.

Opäť využitím Vety 3 dostaneme $\mathsf{var}\hat{Y}_* = \mathbf{x}_*^\top\mathsf{var}(\hat{\boldsymbol{\beta}})\mathbf{x}_* = \sigma^2\mathbf{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_*$. ❑

Dôsledok 2. $\hat{Y}_*$ je nevychýleným odhadom (viď Def. 5) predpovede Y_*.

Definícia 5. *Povieme, že odhad $\hat{\theta}$ je nevychýleným odhadom (skalárneho) parametru θ, ak platí*

$$\mathbb{E}(\hat{\theta} - \theta) = \mathsf{bias}(\hat{\theta}, \theta) = 0.$$

Dôkaz. (Dôsledku 2). Plynie priamo z tvrdenia 8.

❑

K predpovediam sa vrátime v podsekcii 2.5.3, kde vyššie popísanú teóriu použijeme napr. pre definíciu tzv. *PRESS* štatistiky, a potom neskôr v sekcii 3.7. Predtým, ako sa k spomínanej podsekcii dostaneme, zadefinujme si niekoľko ďalších pojmov potrebných pre pochopenie a odvodenie výrazov v danej podsekcii.

2.5.2 Sumy štvorcov v lineárnom modeli

V tejto podsekcii si zadefinujeme ďalšie sumy štvorcov často používané v regresnej analýze a uvedieme si vetu, ktorá popisuje vzťah medzi nimi a RSS. Pripomeňme si, že v práci uvažujeme lineárny (regresný) model, ktorý (spravidla) obsahuje absolútny člen. Tento predpoklad odpovedá tomu, že niektorý zo stĺpcov regresnej matice $\mathbf{X}$ je identicky rovný $\mathbf{1}_n$. Podľa [2, str. 66] model $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ potom splňuje slabšiu požiadavku

$$\mathbf{1}_n \in \mathcal{M}(X). \tag{2.26}$$

Poznámka. Pre modely nespĺňajúce požiadavku (2.26) nasledujúca definícia a veta neplatí.

Definícia 6. *Nech* $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. *Potom výrazy definované nasledujúcimi predpismi nazveme*

1. *Celkový súčet štvorcov (angl. Total sum of squares), ozn.* SS_T,

$$SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2. \tag{2.27}$$

2. *Regresný súčet štvorcov (angl. Regression sum of squares), ozn.* SS_R,

$$SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2. \tag{2.28}$$

Výraz SS_T definovaný predpisom (2.27) môžeme chápať ako celkovú variabilitu hodnôt závislej premennej. Výraz SS_R určený (2.28) predstavuje časť variability, ktorú môžeme vysvetliť pôsobením regresorov, tzv. vysvetlená variabilita, zatiaľ čo hodnota RSS udáva tzv. nevysvetlenú variabilitu.

Veta 9. *Nech* $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, *potom pre sumy štvorcov platí rovnosť*

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SS_T} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SS_R} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{RSS}. \tag{2.29}$$

Dôkaz. Viď [2, str. 66].

❑

Pomocou Definície 6 a Vety 9, ako aj teóriou popísanou v podsekcii 2.5.1, môžeme prejsť k definícii a určení výrazov, ktoré nám vyjadrujú ako dobre model popisuje vzťah medzi závislou premennou a regresormi. V nasledujúcej podsekcii si ukážeme predpis *koeficientu determinácie* spolu s jeho upravenou formou a následne si pomocou metódy n-násobnej krížovej validácie zadefinujeme tzv. *PRESS* štatistiku. Tieto ukazovatele budeme v nasledujúcej kapitole 3 často používať.

2.5.3 Kritéria pre výber regresného modelu

Medzi najbežnejšie ukazovatele toho, ako dobre model popisuje závislosť medzi závislou premennou a regresormi nachádzajucími sa v modeli, patria *koeficient determinácie* a *upravený koeficient determinácie*. Ich hodnoty môžeme vyčítať z výstupu funkcie *summary()*(aplikovanej na výstup funkcie *lm*() v softvéri R), ako hodnoty `R-squared` a `Adjusted R-squared`.

Definícia 7. *Nech* $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *a nech* $\mathbf{1}_n \in \mathcal{M}(\mathbf{X})$. *Hodnotu*

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{RSS}{SS_T}, \tag{2.30}$$

nazveme koeficientom determinácie (angl. coefficient of determination) lineárneho regresného modelu (2.11) *a hodnotu*

$$R^2_{adj} = 1 - \frac{\frac{1}{n-k-1}RSS}{\frac{1}{n-1}SS_T}, \tag{2.31}$$

nazveme upraveným koeficientom determinácie (angl. adjusted coefficient of determination).

Hodnota R^2 nám vypovedá o percentuálnom vysvetlení variability hodnôt Y_i naším modelom a poskytuje náznak toho, ako dobre môžeme predpovedať hodnoty Y na základe regresorov, obsiahnutých v modeli ([8]). Táto veličina nadobúda hodnoty z intervalu [0,1], pričom hodnotu 1 nadobúda pri dokonalom preložení dát odhadom regresnej funkcie, tj. ak pre každé $i = 1, \ldots, n$ platí $Y_i = \hat{Y}_i$. Za platnosti tohto predpokladu je $RSS = 0$ a teda $R^2 = 1$. Naopak, koeficient R^2 nadobúda hodnotu 0 ak model obsahuje iba absolútny člen (viď Lemma 10 nižšie), teda ani jeden z regresorov nám nedáva žiadnu informáciu o správaní sa strednej hodnoty závislej premennej.

Lemma 10. *Pre model* $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, *ktorý obsahuje iba absolútny člen, tj.* $Y_i = \beta_0 + \epsilon_i$, *platí* $SS_R = 0$.

Dôkaz. Regresná matica modelu obsahujúci iba absolútny člen je $\mathbf{X} = \mathbf{1}_n$. Z platnosti $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ a predpisu matice $\mathbf{H}$ je zrejmé, že

$$\hat{\mathbf{Y}} = \mathbf{1}_n \underbrace{(\mathbf{1}_n^\top \mathbf{1}_n)^{-1}\mathbf{1}_n^\top \mathbf{Y}}_{\frac{1}{n}\sum_{i=1}^n Y_i} = \mathbf{1}_n \bar{Y},$$

z čoho vyplýva $\hat{Y}_i = \bar{Y}$ pre každé $i = 1, \ldots, n$. Platnosť $SS_R = 0$ je z predpisu (2.28) zrejmá.

❑

Nie je náročné si rozmyslieť, že R^2 je neklesajúca funkcia počtu regresorov. Vyplýva to z toho, že SS_T je nezávislá od počtu regresorov, zatiaľ čo RSS sa pridaním regresoru (resp. regresorov) do modelu zníži (alebo ostane nemenné). Z tejto vlastnosti vyplýva, že pri porovnávaní modelu a jeho podmodelu s rovnakou závislou premennou, nie je vhodné tento koeficient používať ako ukazovateľ lepšej aproximácie dát odhadnutou regresnou funkciou modelu. Na porovnávanie takýchto modelov je vhodnejšie používať napr. práve upravenú verziu koeficientu determinácie R^2_{adj}, ktorý berie v úvahu počet regresorov (ten je rovný $h(\mathbf{X})$) vyskytujúcich sa v modeli a je definovaný predpisom (2.31).

Ďalšou veličinou popisujúcou presnosť regresného modelu je hodnota *PRESS* (angl. *predicted residual sum of squares*). Tú získame metódou n-násobnej krížovej validácie tak, že z naších dát postupne vynecháme i-té pozorovanie a zvyšné $n-1$ pozorovania použijeme na výpočet odhadu vektora regresných koeficientov v modeli. Odhad vektora regresných koeficientov v modeli s vynechaným i-tým pozorovaním označíme $\hat{\boldsymbol{\beta}}_{(-i)}$ a strednú hodnotu závislej premennej vo vynechanom pozorovaní $\mathbf{x}_i$ odhadneme ako

$$\hat{Y}_{(-i)} = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{(-i)}. \tag{2.32}$$

Ďalej si i-té vynechané rezíduum označme $u_{(-i)}$, pričom

$$u_{(-i)} = Y_i - \hat{Y}_{(-i)}. \tag{2.33}$$

Poznámka. Odhad $\hat{\boldsymbol{\beta}}_{(-i)}$ počítame vždy pri jednom vynechanom pozorovaní, a preto sa daný postup získania hodnoty rovnice (2.35) občas nazýva $LOOCV$ (angl. *Leave-one-out cross validation*).

Aby sme sa vyhli n-násobnému odhadovaniu regresných koeficientov v modeli, využijeme platnosť rovnosti (2.34) (viď [13, str. 45])

$$u_{(-i)} = Y_i - \hat{Y}_{(-i)} = \frac{u_i}{1 - h_{ii}}, \tag{2.34}$$

kde h_{ii} je i-tý diagonálny prvok matice $\mathbf{H}$ a PRESS štatistiku si podľa [13, str. 45] určíme v tvare

$$PRESS = \sum_{i=1}^{n} (u_{(-i)})^2. \tag{2.35}$$

Pri rozhodovaní sa, ktorý z (rôznych) modelov popisuje vzťah medzi premennými lepšie si vyberieme ten, ktorého hodnota (2.35) bude nižšia.

V tejto kapitole sme si predstavili základné pojmy lineárnej regresie. Zadefinovali sme si lineárny model viacnásobnej regresie a predviedli sme si, ako odhady regresných koeficientov dopočítame metódou najmenších štvorcov. Následne sme si ukázali základne vlastnosti týchto odhadov a ich vlastnosti ak by sme uvažovali mnohorozmerné normálne rozdelenie náhodného vektora $\mathbf{Y}$. Ďalej sme si zadefinovali pojem podmodelu a vetu o prechode k podmodelu, na ktorú sa budeme v nasledujúcej kapitole 3 často odkazovať. V závere kapitoly (2) sme si zadefinovali niekoľko ukazovateľov určujúcich ako dobre model popisuje závislosť medzi závislou a nezávislými premennými.

Teraz sa v nasledujúcej kapitole pozrieme na niektoré konkrétne predpisy rôznych lineárnych modelov, ktorých postupnou kombináciou sa dostaneme k predpisu modelu *regresných splinov* a ako sa uvažovaním lineárnych obmedzení rozoberaných v sekcii 2.4 dostaneme k ich podmodelu - modelu *prirodzených splinov.*

Kapitola 3

Spliny

Pôvod slova *spline* môžeme nájsť v lodnom inžinierstve, kde sa týmto pojmom označoval nástroj na kreslenie kriviek. Dnes sa so slovom spline stretneme mimo iné aj v regresnej analýze, kde splinom nazývame po častiach spojitú polynomickú funkciu. Body, v ktorých sa jednotlivé polynómy napájajú nazývame *uzly* a ich význam si vysvetlíme v sekcii 3.2. Predtým, ako sa dostaneme k predpisu modelu regresného splinu, prevedieme čitateľa modelom regresných polynómov (sekcia 3.1), ako aj modelom skokovej regresnej funkcie (sekcia 3.2), ktoré, ako uvidíme, s modelom regresného splinu úzko súvisia.

V kapitole 2 sme zadefinovali lineárny model viacnásobnej regresie, v ktorom sme uvažovali, že spolu s hodnotami závislej premennej je obecne pozorovaných k rôznych regresorov. Ďalej v práci sa však budeme venovať iba modelom, v ktorých spolu s náhodou veličinou Y je pozorovaná jedna spojitá (nenáhodná) veličina x. Rovnako ako v prechádzajúcej kapitole bude uvažovať, že naše dáta pozostávajú z n pozorovaní daných veličín, $(Y_i, x_i), i = 1, \ldots, n$. Regresnú funkciu takýchto modelov budeme uvažovať v tvare

$$\mathbb{E}Y_i = \beta_0 f_0(x_i) + \beta_1 f_1(x_i) + ... + \beta_k f_k(x_i) = \sum_{j=0}^{k} \beta_j f_j(x_i) = f(x_i), \tag{3.1}$$

kde $f_j : \mathbb{R} \to \mathbb{R}$, $j = 0$,...,k sú známe, lineárne nezávislé funkcie. Tieto funkcie budeme nazývať *bázové funkcie*.

Definícia 8. *Množinu lineárne nezávislých známych funkcií $f_j(x), j = 0, \ldots, k$ budeme nazývať systémom bázových funkcií.*

Poznámka. Bázovú funkciu odpovedajúcu absolútnemu členu β_0 budeme automaticky (až na niektoré výnimky) uvažovať ako $f_0(x_i) = 1$.

Jednotlivé príklady modelov predstavené v tejto kapitole získame uvažovaním rôznych predpisov bázových funkcií $f_j(x), j = 0, \ldots, k$. Je zrejmé, že náš regresný model priamky (2.1) je vyjadrený lineárnou kombináciou bázových funkcií $f_0(x_i) = 1$ a $f_1(x_i) = x_i$, teda $\{1, x\}$ je báza vektorového priestoru všetkých priamok. Predpisom bázových funkcií $f_j(x_i) = x_i^j$ môžeme pre $j \in \mathbb{N}_0$ generovať bázu *regresného modelu polynómu* ľubovoľného stupňa v tvare $\{1, x, x^2, ..., x^j, ...\}$, ktorý si rovno predstavíme v nasledujúcej sekcii 3.1.

3.1 Regresný model polynómu

Ako sme už spomenuli v kapitole 2, model regresnej priamky nemusí, a v praxi často ani nie je kvôli svojej jednoduchosti veľmi vhodným modelom pre popis vzťahu závislosti medzi závislou a nezávislou premennou. Zvlášť, ak je vzťah popisujúci závislosť medzi premennými nelineárny, je model priamky priam nepoužiteľný. Ako nám napovedá názov sekcie, uvažovaním komplexnejších funkcií, akými sú polynómy, môžeme preloženie dát odhadom regresnej funkcie modelu polynómu výrazne zlepšiť.

Uvažujme, že regresory v lineárnom modeli viacnásobnej regresie (2.11) nahradíme transformáciou našej jednej nezávislej premennej, a to konkrétne mocninami vyššieho rádu. Náš model nadobúda pre $d \in \mathbb{N}_0$ tvar

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_d x_i^d + \epsilon_i, i = 1,...,n, \tag{3.2}$$

ktorý nám zachováva vlastnosti obecného modelu (2.11) a nazveme ho *regresný model polynómu (jednej nezávislej premennej).*

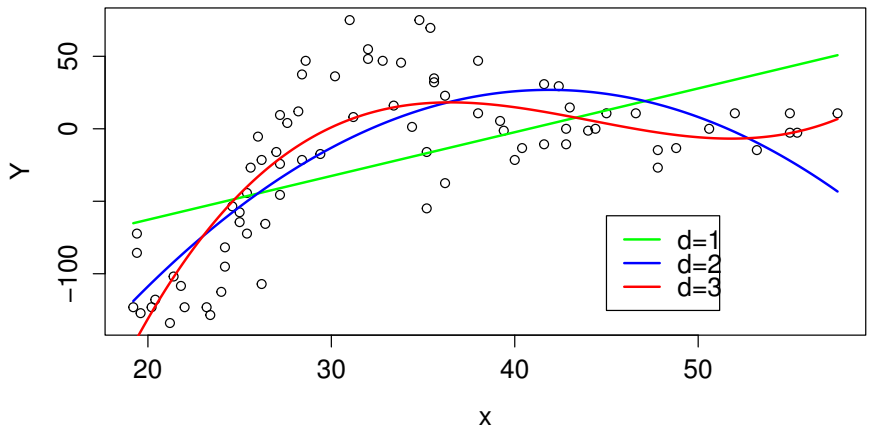
V tomto modeli môžeme pomocou štatistiky z Vety 6 a hypotézou $H_0 : \beta_r = \ldots = \beta_d = 0$ pre $r \leq d$ testovať, či by nebolo vhodné prejsť k podmodelu modelu regresného polynómu stupňa d, a to k modelu regresného polynómu stupňa $r-1$.

Model (3.2) je stále lineárny v parametroch, keďže jeho regresná funkcia je vyjadrená lineárnou kombináciou bázových funkcií, hoci funkcia popisujúca vzťah strednej hodnoty závislej premennej a nezávislej premennej lineárna nie je (pre voľbou $d > 1$). V praxi sa stupeň polynómu d volí maximálne 3 alebo 4, keďže krivka vyšších polynómov môže byť príliš flexibilná a nadobúdať príliš oscilujúci tvar. Je zrejmé, že pri voľbe $d = 1$ dostaneme tvar nášho regresného modelu priamky (vzťah (2.1)) a voľbou $d = 2$ (resp. $d = 3$) dostaneme kvadratickú (resp. kubickú) závislosť strednej hodnoty Y na vysvetľujúcej premennej. Ako vidíme nižšie na obrázku 3.1, zvyšovaním stupňa d polynómu dostávame lepšie odhady strednej hodnoty odozvy, v zmysle zväčšovania hodnoty R^2_{adj} pre voľbu $d = 1, 2, 3$.

Poznámka. Dáta na obrázku 3.1 pochádzajú z balíčka MASS v softvéri R. Hodnotami závislej premennej Y_i budú pozorovania *accel* (zrýchlenie) a hodnoty nezávislej premennej x_i pozorovania *times* (čas) datasetu *mcycle*. Uvažujme iba pozorovania, pre ktoré je hodnota premennej *times* väčšia ako 19. Veľkosť tohto dátového súboru je $n = 78$. S týmito dátami budeme pracovať v sekciách 3.2 - 3.4.

3.2 Model skokovej regresnej funkcie

Hlavnou myšlienkou modelu skokovej funkcie je rozdelenie si intervalu nameraných hodnôt nezávislej premennej $[min(x_i), max(x_i)]$ na niekoľko rôznych neprekrývajúcich sa úsekov. Body, ktoré nám rozdelia interval hodnôt nezávislej premennej na menšie podintervaly, označíme $min(x_i) = \xi_0 < \xi_1 < \ldots < \xi_{K+1} = max(x_i)$. Body $\xi_k, k = 1, \ldots, K$, nazývame *vnútorné uzly* (resp. iba *uzly* či *uzlové body*) a body ξ_0 a ξ_{K+1} nazývame *hraničné uzly* - určujú nám hranice intervalu prekladania dát odhadom regresnej funkcie. Predpokladajme, že tieto body sú vopred dané konštanty.



Obr. 3.1: Dáta (viď Pozn. 3.1) preložené odhadmi polynómov rôznych stupňov d. Upravené koeficienty determinácie jednotlivých (zaokrúhlene na 2 desatinné miesta) volieb d sú $R^2_{adj} = 0.30$ pre $d = 1$, $R^2_{adj} = 0.60$ pre $d = 2$ a $R^2_{adj} = 0.67$ pre $d = 3$.

Uzlové body nám rozdelia škálu nameraných hodnôt nezávislej premennej na $K + 1$ podintervalov

$$\mathcal{I}_k = [\xi_k, \xi_{k+1}), k = 0, \ldots, K, \tag{3.3}$$

na ktorých budeme dáta zvlášť aproximovať najprv odhadom konštanty a potom lineárnou funkciou. Aby sme ukázali prvý typ aproximácie, tj. odhadom konštanty, zadefinujeme si indikátorovú funkciu $I_A(x)$ nasledovne

$$I_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \text{inak.} \end{cases}$$

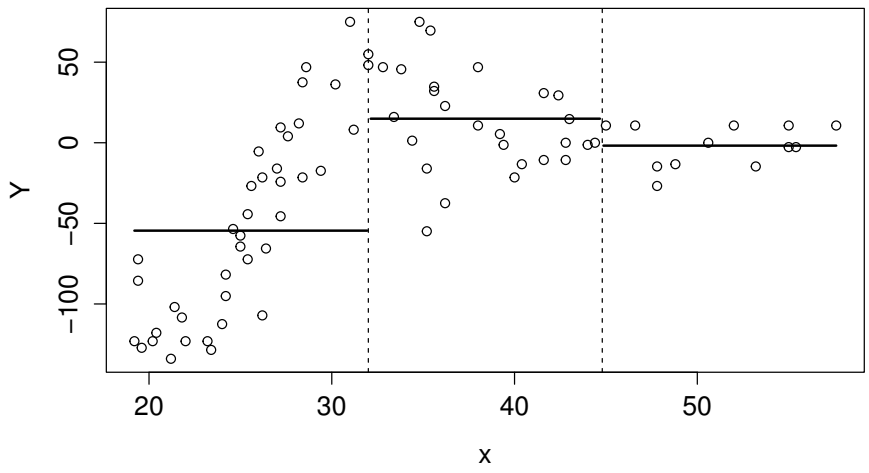
Model skokovej funkcie s K uzlovými bodmi nadobúda pre i =1,...,n tvar

$$Y_i = \beta_0 + \beta_1 I_{\mathcal{I}_1}(x_i) + ... + \beta_K I_{\mathcal{I}_k}(x_i) + \epsilon_i. \tag{3.4}$$

Pre $x_i \in [\xi_0, \xi_1)$ je v modeli (3.4) $\mathbb{E}Y_i = \beta_0$ a pre $x_i \in [\xi_k, \xi_{k+1}), k$ =1,...,K je $\mathbb{E}Y_i = \beta_0 + \beta_k$. Jednotlivé regresné koeficienty β_k, k =1,...,K, predstavujú rozdiel medzi strednou hodnotou závislej premennej Y_i na podintervale $\mathcal{I}_0$ a podintervale $\mathcal{I}_k$. Z predpisu $\mathbb{E}Y_i$ (na jednotlivých intervaloch) vidíme, že $\mathbb{E}Y_i$ je po častiach konštantná funkcia (premennej x_i) so skokmi v uzloch ξ_k, k =1,...,K. Na obrázku 3.2 môžeme vidieť naše dáta preložené odhadom regresnej funkcie modelu (3.4) s dvoma uzlovými bodmi ξ_1 a ξ_2 volenými rovnomerne na intervale $[min(x_i), max(x_i)]$.

3.3 Model po častiach lineárnej funkcie

Teraz prejdeme k trochu komplexnejšiemu modelu, a to k aproximácií dát odhadom po častiach lineárnej funkcie. Jednotlivé podintervaly (3.3) budeme



Obr. 3.2: Dáta preložené odhadom regresnej funkcie modelu skokovej funkcie, teda konštantou na jednotlivých podintervaloch oddelených zvislou prerušovanou čiarou znázorňujúcou polohu uzlových bodov ξ_1 a ξ_2.

prekladať odhadmi priamok. Je zrejmé, že daný model bude obsahovať dvakrát viac regresných koeficientov ako model (3.4).

Model po častiach lineárnych funkcií s K uzlovými bodmi definujeme predpisom

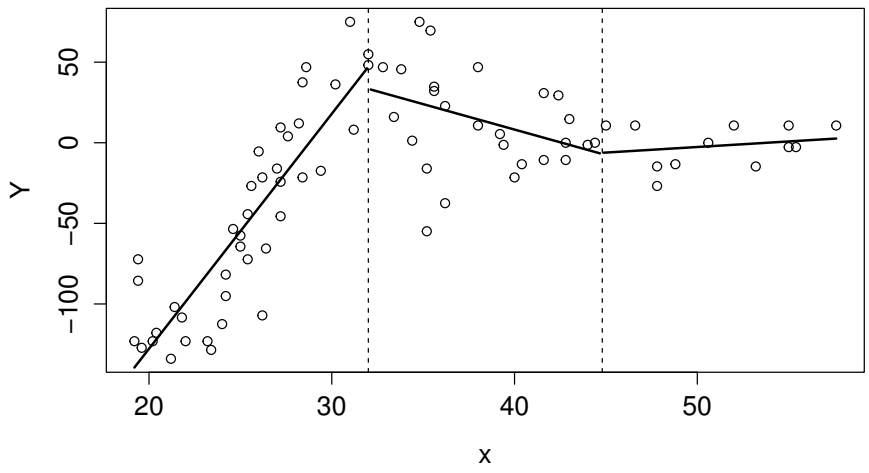
$$\begin{aligned} Y_i = &\beta_0 I_{\mathcal{I}_0}(x_i) + \beta_1\, x_i\, I_{\mathcal{I}_0}\,(x_i) + \beta_2\, I_{\mathcal{I}_1}\,(x_i) + \beta_3\, x_i\, I_{\mathcal{I}_1}\,(x_i) + ... \\ &+ \beta_{2K}\, I_{\mathcal{I}_K}\,(x_i) + \beta_{2K+1}\, x_i\, I_{\mathcal{I}_K}\,(x_i) + \epsilon_i. \end{aligned} \tag{3.5}$$

Pre $x_i \in [\xi_r, \xi_{r+1})\,, r = 0, \ldots, K$ platí $\mathbb{E}Y_i = \beta_{2r} + \beta_{2r+1} x_i$, teda na jednotlivých intervaloch je regresná funkcia priamkou. Hypotézou $H_0 : \beta_1 = \beta_3 = \ldots = \beta_{2K+1}$ môžeme testovať prechod k podmodelu modelu (3.5) - modelu skokovej funkcie (3.4). V grafe na obrázku 3.3 vidíme preklad dát odhadom priamky na jednolivých podintervaloch s dvoma uzlovými bodmi ξ_1 a ξ_2.

Pridávaním ďalších mocninových transformácií nezávislej premennej by sme mohli model (3.5) rozšíriť tak, aby sme jednotlivé úseky prekladali odhadom polynómu stupňa $d > 1$. Predtým, ako tak urobíme, si náš model (3.5) upravíme, aby sa jednotlivé priamky v uzlových bodoch na seba napájali, teda aby bola naša regresná funkcia v uzloch spojitá.

3.3.1 Model po častiach spojitej lineárnej regresnej funkcie

Teraz si podľa [14] ukážeme intuitívny spôsob získania modelu po častiach spojitej lineárnej regresnej funkcie. Bez ujmy na obecnosti uvažujme nejaký ľubovoľný (vnútorný) uzol ξ. Ďalej predpokladajme, že tento bod nám hodnoty nezávislej premennej rozdeľuje na dve podintervaly, na ktorých budeme dáta prekladať odhadmi priamky. Na intervale $x < \xi$ uvažujme priamku určenú predpisom $\beta_0 + \beta_1 x$



Obr. 3.3: Dáta preložené odhadmi po častiach lineárnych funkcií oddelené zvislými prerušovanými čiarami znázorňujúce polohu uzlových bodov ξ_1 a ξ_2.

a na intervale $x \geq \xi$ priamku určenú predpisom $\beta_0' + \beta_1' x$. Požiadavkou spojitosti v bode $x = \xi$ vyžadujeme

$$\beta_0 + \beta_1 \xi = \beta_0' + \beta_1' \xi. \tag{3.6}$$

Parameter β_1' si môžeme vyjadriť v tvare $\beta_1' = \beta_1 + \alpha$, kde α predstavuje zmenu smernice priamky pri prechode z intervalu $x < \xi$ do intervalu $x \geq \xi$. Rovnicu (3.6) si teda môžeme prepísať do tvaru

$$\beta_0 + \beta_1 \xi = \beta_0' + (\beta_1 + \alpha)\xi,$$

odkiaľ dostaneme $\beta_0' = \beta_0 - \alpha\xi$. Hodnotu regresnej funkcie v i-tom pozorovaní nášho uvažovaného modelu, teda modelu s jedným uzlovým bodom, v ktorom je regresná funkcia spojitá, môžeme vyjadriť v tvare

$$\begin{aligned} \mathbb{E}\, Y_i &= (\beta_0 + \beta_1 x_i)\, I_{(-\infty,\xi)}(x_i) + (\beta_0' + (\beta_1 + \alpha)x_i)\, I_{[\xi,\infty)}(x_i) \\ &= (\beta_0 + \beta_1 x_i)\, I_{(-\infty,\xi)}(x_i) + (\beta_0 + \beta_1 x_i + \alpha(x_i - \xi))\, I_{[\xi,\infty)}(x_i) \\ &= \beta_0 + \beta_1 x_i + \alpha(x_i - \xi)\, I_{[\xi,\infty)}(x_i). \end{aligned} \tag{3.7}$$

Výraz $(x_i - \xi)\, I_{[\xi,\infty)}(x_i)$ budeme značiť $(x_i - \xi)_+$ a definujeme predpisom

$$(x_i - \xi)_+ = \begin{cases} x_i - \xi, & x_i \geq \xi, \\ 0, & \text{inak.} \end{cases}$$

Definícia 9. *Funkciu $h(x)$ stupňa $d \in \mathbb{N}_0$ určenú predpisom*

$$h(x) = (x - \xi)_+^d = ((x - \xi)_+)^d \tag{3.8}$$

budeme nazývať useknutou mocninovou funkciou (stupňa d) (angl. truncated power function).

Pre $d > 0$ je funkcia (3.8) v bode ξ pre spojitá a s predpokladom $0^0 = 0$ má táto funkcia pre $d = 0$ v bode ξ skok o veľkosti 1 ([9, str. 82]).

Dôsledok 3. Derivácia $h(x)$ je rovná

$$\frac{\mathsf{d}h(x)}{\mathsf{d}x} = d(x-\xi)_+^{d-1},$$

teda funkcia má $d-1$ spojitých derivácií, pričom d-tá derivácia má v bode ξ skok o veľkosti $d!$.

Model po častiach spojitej lineárnej regresnej funkcie s K (vnútornými) uzlami $\xi_1, \ldots, \xi_K$, si zadefinuje zobecnením (3.7) ako

$$Y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \beta_{1+k}(x_i - \xi_k)_+ + \epsilon_i, \ i = 1,...,n. \tag{3.9}$$

Regresnú funkciu modelu (3.9) môžeme písať v tvare

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \beta_{1+k}(x - \xi_k)_+,$$

a jej bázu ako

$$\{1, x, (x-\xi_1)_+, \ldots, (x-\xi_K)_+\}.$$

V grafe na Obrázku 3.3.1 môžeme vidieť preložené dáta odhadom regresnej funkcie modelu 3.9. Ako môžeme vidieť z obrázku 3.3.1, odhad regresnej funkcie modelu (3.9) (s dvoma uzlovými bodmi) nadobúda v uzloch ostré hrany, tzn. že v daných uzloch nie je funkcia diferencovateľná. Tento problém môžeme vyriešiť pomocou bázy generovanej vyššími mocninami useknutej mocninovej funkcie. Báza funkcie spojitého polynómu v K uzloch stupňa d pozostáva z funkcií $1, x, ..., x^d, (x-\xi_1)_+^d, ..., (x-\xi_K)_+^d$. Lineárnou kombináciou daných bazových funkcií sa dostávame k prepisu a definícii *regresného splinu.*

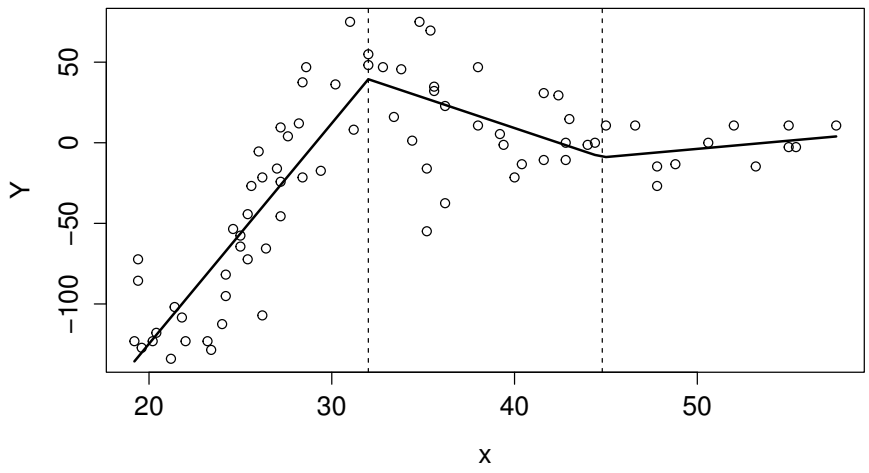
3.4 Obecný model splinu

Ako sme si ukázali postupnými nadväznosťami regresného modelu polynómu a modelu skokovej regresnej funkcie, *spliny* si môžeme predstaviť ako po častiach spojité polynómické funkcie.

Definícia 10. *Funkciu* $f : [\xi_0, \xi_{K+1}] \to \mathbb{R}$ *nazveme splinom stupňa* $d \geq 0$ *s uzlami* $\xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1}$ *ak spĺňa nasledujúce podmienky:*

1. $f(x)$ *je* $(d-1)-$*krát spojito diferencovateľná. Výnimka nastáva pre* $d = 1$*, keď funkcia* $f(x)$ *je spojitá, ale nie diferencovateľná. (Pre* $d = 0$ *funkcia nie je spojitá).*

2. $f(x)$ *je polynóm stupňa* d *na jednotlivých intervaloch* $[\xi_k, \xi_{k+1})$.



Obr. 3.4: Dáta preložené odhadom regresnej funkcie modelu (3.9) s 2 uzlami, ktorých poloha je znázornená zvislou prerušovanou čiarou.

Spliny s vopred pevne danými uzlovými bodmi budeme nazývať *regresné spliny*. Model regresného splinu stupňa d s K vnútornými uzlovými bodmi môžeme vyjadriť v tvare

$$Y_i = \beta_0 + \beta_1 x_i + ... + \beta_d x_i^d + \sum_{k=1}^{K} \beta_{d+k}(x_i - \xi_k)_+^d + \epsilon_i, \;\; i = 1,...,n, \tag{3.10}$$

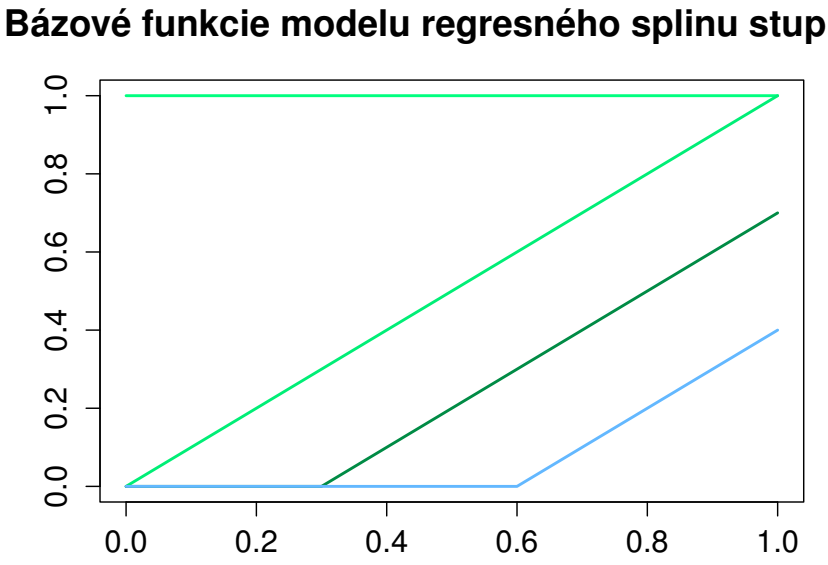
kde $\beta_k, k = 0, \ldots, K + d + 1$, sú regresné koeficienty a systém bázových funkcií daného modelu

$$\{1, x,..., x^d, (x - \xi_1)_+^d,..., (x - \xi_K)_+^d\} \tag{3.11}$$

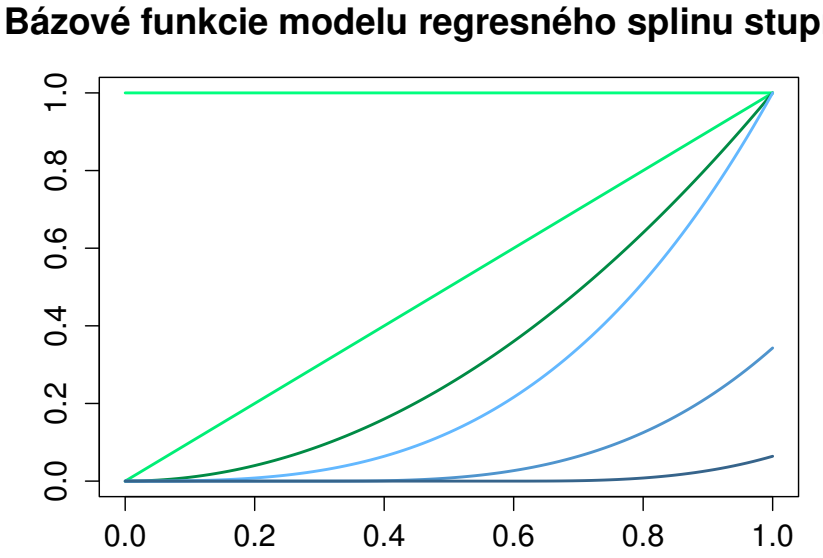
nazývame *useknutá mocninová báza*. Voľbou $d = 1, 2, 3$ v (3.10) dostaneme (v odpovedajúcom poradí) model lineárneho(viď (3.9)), kvadratického a kubického splinu. Práve model kubického splinu je jedným z najpoužívanejších modelov regresných splinov. Keďže má spojitú prvú a druhú deriváciu, krivka určujúca kubický spline je dostatočne hladká pre väčšinu praktických problémov. K modelu kubického splinu sa vrátime v sekcii 3.6, kde uvažovaním reštrikcií na regresnú funkciu tohto modelu získame predpis jeho podmodelu, a to model *prirodzených splinov*.

Na obrázku 3.5 môžeme vidieť systém bázových funkcií modelu regresného splinu (3.10) s voľbou $d = 1$ a voľbou $d = 3$ na obrázku 3.6 s dvoma uzlovými bodmi $\xi_1 = 0.3$ a $\xi_2 = 0.6$.

Z predpisu (3.10) je zrejmé, že uvažovaním $\beta_{d+k} = 0, k = 1,...,K$ by sme dostali predpis (3.2) regresného modelu polynómu stupňa d, teda model polynómu je podmodelom modelu regresných splinov. Uvažovaním $\boldsymbol{\epsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ v modeli (3.10) by sme mohli testovou štatistikou zavedenou vo vete 6 a hypotézou $H_0 : \beta_{d+1} = \ldots = \beta_{d+K} = 0$ skúmať platnosť podmodelu, teda či by vzťah medzi premennými Y a x nebolo vhodnejšie popisovať práve modelom regresného polynómu.



Obr. 3.5: Systém bázových funkcií lineárneho regresného splinu (d=1) s dvoma uzlovými bodmi $\xi_1 = 0.3$ a $\xi_2 = 0.6$.

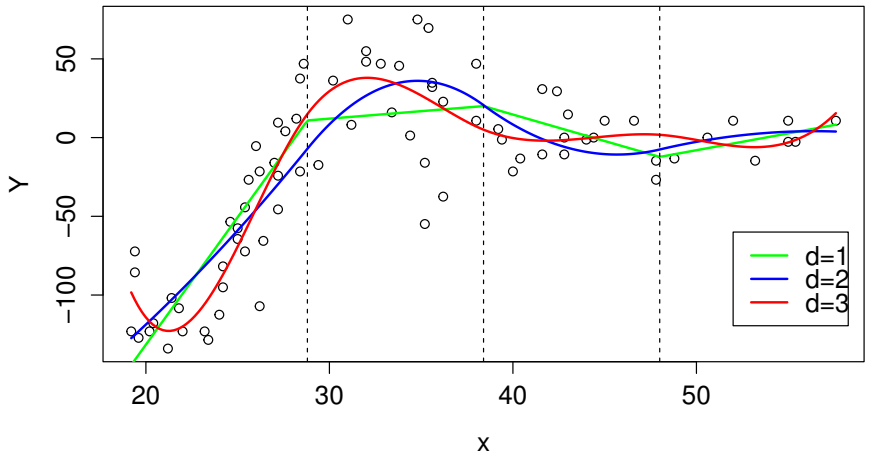


Obr. 3.6: Systém bázových funkcií kubického regresného splinu (d=3) s dvoma uzlovými bodmi $\xi_1 = 0.3$ a $\xi_2 = 0.6$.

Poznámka. V elektronickej prílohe (súbor `funkcie.R`) môžeme nájsť nami naprogramovanú funkciu *rs()*, ktorú sme používali pri konštrukcii Obrázku 3.7. Táto funkcia nám generuje regresnú maticu modelu (3.10). Jej argumenty sú `x`, do

ktorého zadávame hodnoty nezávislej premennej, potom argument `uzly`, ktorým udávame postupnosť (vnútorných) uzlov a argument `stup`, ktorým zadávame stupeň splinu.

Z obrázku 3.7 nižšie môžeme vidieť, ako postupným zvyšovaním stupňa splinu d dostávame (v rámci nášho dátového súboru) stále lepší odhad regresnej funkcie, v zmysle zvyšovania hodnoty R^2_{adj} so zvyšujúcim sa stupňom d. Prerušovanou čiarou sú znázornené uzly volené v rovnomerne na intervale $[min(x_i), max(x_i)]$. Upravené koeficienty determinácie (zaokrúhlene na 2 desatinné miesta) sú pre $d = 1$ $R^2_{adj} = 0.71$, pre $d = 2$ je $R^2_{adj} = 0.71$ a pre $d = 3$ je R^2_{adj} najvyšší, a to 0.78.



Obr. 3.7: Dáta preložené odhadmi regresnej funkcie splinu s odpovedajúcim stupňom d a troma uzlovými bodmi, ktorých poloha je vyznačená zvislou prerušovanou čiarou.

3.4.1 B-spline báza

Pre odhad regresných koeficientov modelu splinov je možné v softvéri R použiť vbudovanú funkciu *bs*() z balíčka *splines*. Táto funkcia nám generuje regresnú maticu s inými bázovými funkciami, ako sú nami vyššie uvedené bázové funkcie (3.11) modelu regresných splinov. Bázové funkcie tohto modelu totiž nemusia byť vhodné, keďže pre uzly nachádzajúce sa bezprostredne blízko seba sú tieto funkcie takmer lineárne závislé, či ako sa dočítame v ([13, str. 70]), pri vysokom počte uzlov môže viesť výpočet odhadov regresných koeficientov k numerickej nestabilite. Preto sa v praxi používajú ekvivalentné bázy (bázy generujúce rovnaký priestor), pomocou ktorých je výpočet odhadov regresných koeficientov numericky stabilnejší. Medzi najpoužívanejšie patrí *B-spline* báza. Ako sa môžeme dočítať v [13, str. 70], vyrovnané hodnoty $\hat{Y}_i, i = 1,...,n$ modelu regresných splinov s bázou (3.11) a bázou tvorenou B-spline bázovými funkciami, sú pre obe bázy totožné.

Predtým, ako si zadefinujeme predpis B-spline bázových funkcií, musíme si našu postupnosť vnútorných a hraničných uzlov rozšíriť. Uvažujme ďalších $2(m-$

1) *vonkajších uzlov*, ktoré sú potrebné pre konštrukciu bázových funkcií B-splinu rádu $m = d + 1$ označených $B_{j,m}(x), j = -(m-1),...,K$.

Rastúcu postupnosť uzlov si označme $\boldsymbol{\xi} = (\xi_{-(m-1)},...,\xi_{K+m})$. Potom $B_{j,m}(x)$ je j-tá B-spline bázová funkcia rádu m (stupňa $d+1$) s postupnosťou uzlov $\boldsymbol{\xi}$, daná rekurzívnym predpisom podľa [9, str. 90] nasledovne

$$B_{j,1}(x) = \begin{cases} 1, & \xi_j \leq x < \xi_{j+1}, \\ 0, & \text{inak}, \end{cases} \tag{3.12}$$

a

$$B_{j,m}(x) = \frac{x - \xi_j}{\xi_{j+m-1} - \xi_j} B_{j,m-1}(x) - \frac{x - \xi_{j+m}}{\xi_{j+m} - \xi_{j+1}} B_{j+1,m-1}(x). \tag{3.13}$$

Týmto predpisom rekurzie si môžeme napočítať B-spline bázu ľubovoľného rádu. Teraz si podľa výkladu kníh [9] a [10] sformulujeme základné vlastnosti B-spline bázových rádu m.

1. *Pozitivizmus*: Platí
$$B_{j,m}(x) \geq 0, \qquad x \in \mathbb{R}.$$

2. *Nosič funkcie*: Platí
$$B_{j,m}(x) = 0, \qquad x \notin [\xi_j, \xi_{j+m}].$$

3. Pre $x \in [\xi_0, \xi_{K+1}]$ platí:
$$\sum_{j=-(m-1)}^{K} B_{j,m}(x) = 1.$$

4. *Derivácia* ($m > 1$):
$$\frac{\partial B_{j,m}(x)}{\partial x} = (m-1)\left(\frac{-B_{j+1,m-1}(x)}{\xi_{j+m} - \xi_{j+1}} + \frac{B_{j,m-1}(x)}{\xi_{j+m-1} - \xi_j}\right),$$
a pre $m = 1$ je derivácia bázovej funkcie rovná nule ([9, str. 117]). Z daného predpisu vyplýva, že derivácia B-spline bázových funkcií rádu m sa dá vyjadriť ako lineárna kombinácia B-spline bázových funkcií rádu $m-1$, teda spline generovaný B-spline bázou je diferencovateľný do rádu 1.

Poznámka. O ďalších vlastnostiach B-spline báze a B-spline bázových funkcií sa môžeme dočítať v [9], prípadne v [10].

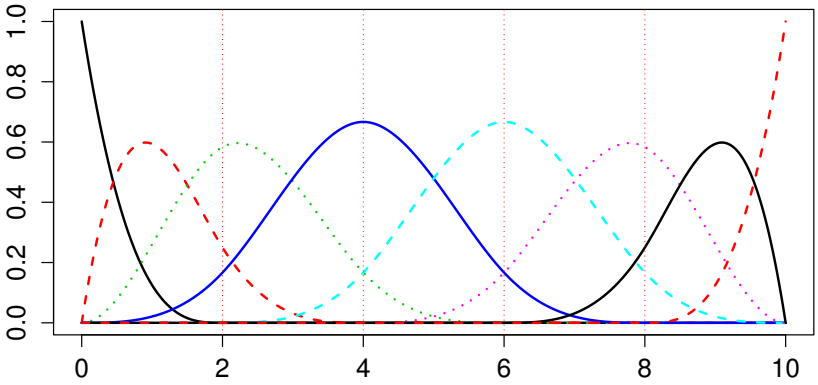
Regresnú funkciu splinu získame lineárnou kombináciou jeho bázových funkcií [10, str. 11], teda spline s B-spline bázou rádu m s uzlovými bodmi $\boldsymbol{\xi}$ nadobúda tvar

$$f(x) = B_{\boldsymbol{\beta}}^{\boldsymbol{\xi},m}(x) = \sum_{j=-(m-1)}^{K} \beta_j B_{j,m}(x), \tag{3.14}$$

kde $\beta_j, j = -(m-1),..., K$ sú regresné koeficienty.

Na obrázku 3.8 môžeme vidieť B-spline bázové funkcie rádu $m = 4$ so štyrmi rovnomerne rozmiestnenými (vnútornými) uzlami na intervale $[0, 10]$.

Teraz si v nasledujúcej sekcii 3.5 zhrnieme niekoľko najbežnejšie používaných prístupov voľby polohy a počtu uzlových bodov.



Obr. 3.8: B-spline bázové funkcie rádu $m = 4$ so štyrmi rovnomerne rozmiestnenými (vnútornými) uzlami, ktorých poloha je znázornená zvislou prerušovanou čiarou.

3.5 Voľba polohy a počtu uzlov

V tejto sekcii si podľa výkladu kníh [11] a [13] popíšeme niektoré základne prístupy voľby polohy a počtu uzlových bodov. Správnou voľbou polohy a počtu uzlov môžeme totiž dosiahnuť výrazné zlepšenie preloženia dát odhadom regresnej funkcie.

Jedným z intuitívnych prístupov voľby polohy uzlov je voľba väčšieho počtu uzlov v miestach, kde je regresná funkcia výraznejšie variabilná, zatiaľ čo v miestach, kde je táto funkcia stabilnejšia, je vhodné voliť menej uzlov. V praxi sa však skôr stretávame s voľbou rovnomerného rozmiestnenia uzlov. V prípade, ak sú hodnoty nezávislej premennej rozmiestnené na intervale $[min(x_i), max(x_i)]$ s (približne) rovnakým odstupom, vhodným prístupom voľby polohy uzlov je rovnomerné rozmiestnenie na danom intervale. Pre voľbu K (vnútorných) uzlov môžeme použiť nasledujúcu formulu

$$\xi_k = min(x_i) + (max(x_i) - min(x_i))k/(K+1),\ k = 1, \ldots, K.$$

Naopak, ak hodnoty nezávislej premennej nie sú (približne) rovnako rozmiestnené na intervale prekladania, vhodným prístupom voľby uzlov je rovnomerne v kvantiloch hodnôt nezávislej premennej. Tento prístup nám zaručí väčší počet uzlov v miestach intervalu $[min(x_i), max(x_i)]$, kde máme väčšie množstvo pozorovaní, zatiaľ čo v častiach intervalu s menším počtom pozorovaní bude počet uzlov malý.

Ako sa môžeme dočítať v [12, str. 23], na rozdiel od polohy uzlov je pre zlepšenie odhadu regresnej funkcie práve dôležitejším kritériom správna voľba počtu uzlov K. Medzi najpoužívanejšie prístupy posúdenia optimálneho počtu uzlov (pri predom danej voľbe ich polohy) patrí napr. rozhodovanie sa na základe R^2_{adj}, či $PRESS$ štatistiky (sekcia 2.5.3) alebo napr. v modeli regresného alebo prostredníctvom F-testu, ktorým môžeme skúmať prípadnú platnosť podmodelu s odstráneným jedným alebo viacero uzlov.

V nasledujúcej sekcii si zadefinujeme podmodel regresného kubického splinu, a to model prirodzeného splinu. Následne si odvodíme pred bázových funkcií vhodných na jeho reprezentáciu.

3.6 Prirodzený kubický spline

Ako sme si ukázali v sekcii 3.4, regresné spliny môžeme chápať ako po častiach spojité polynomické funkcie. Ako sa môžeme dočítať v [16], odhady regresných funkcií modelu regresného splinu (3.10) majú v hraničných častiach intervalu hodnôt nezávislej premennej veľký rozptyl, čo môže viesť k veľmi nepresnej extrapolácií, teda odhadu stredných hodnôt závislej premennej mimo interval prekladania $[min(x_i), max(x_i)]$. Tento problém môžeme potenciálne vyriešiť zavedením reštrikcií na regresné koeficienty modelu regresného splinu s K (vnútornými) uzlami $\xi_k, k = 1, \ldots, K$, tak, aby regresná funkcia daného modelu bola vo chvostoch, teda naľavo od najmenšieho uzla ξ_1 a napravo od najväčšieho uzla ξ_K, lineárna. Uvažovaním takýchto reštrikcií sa dostávame k definícií modelu *prirodzeného (kubického) splinu.*

Poznámka. V tejto sekcii budeme uvažovať iba model kubického regresného splinu, a preto slovo kubický budeme vynechávať.

Regresnú funkciu regresného splinu s K uzlami si označme $f(x)$, a kvôli potrebe ďalšej práce s danou funkciou si preznačme regresné koeficienty. Regresnú funkciu uvažujme v tvare

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3, \tag{3.15}$$

kde $\beta_j, j = 0, \ldots, 3$ a $\theta_k, k = 1, \ldots, K$ sú regresné koeficienty. Linearita regresnej funkcie (3.15) je v jej chvostoch zaručená položením $f''(x) = f'''(x) = 0$ pre $x \leq \xi_1$ a $x \geq \xi_K$. Uvažovaním daného obmedzenia sa dostaneme k reštrikciám na regresné koeficienty, ktoré nám zaručia linearitu $f(x)$ vo chvostoch a ich aplikovaním si určíme predpis bázových funkcií prirodzeného splinu.

Pre hodnoty $x \leq \xi_1$ je zrejmé, že výraz $(x - \xi_k)_+^3$ bude pre k =1,...,K nulový, teda linearitu $f(x)$ (pre $x \leq \xi_1$) dosiahneme položením

$$\beta_2 = \beta_3 = 0. \tag{3.16}$$

Pre $x \geq \xi_K$ je výraz $(x - \xi_k)_+^3$ pre k =1,...,K nenulový, teda aplikovaním (3.16) na (3.15) dostávame druhú a tretiu deriváciu $f(x)$ v tvare

$$f''(x) = 6 \sum_{k=1}^{K} \theta_k (x - \xi_k), \ x \geq \xi_K, \tag{3.17}$$

$$f'''(x) = 6 \sum_{k=1}^{K} \theta_k, \ x \geq \xi_K. \tag{3.18}$$

Položením (3.17) a (3.18) nule dostaneme ďalšie reštrikcie na regresné koeficienty v tvare

$$\sum_{k=1}^{K} \theta_k = 0, \tag{3.19}$$

$$\sum_{k=1}^{K} \xi_k \theta_k = 0. \tag{3.20}$$

Zavedením reštrikcií (3.19) a (3.20) si môžeme dva parametre, θ_K a θ_{K-1}, vyjadriť pomocou zvyšných $K-2$ voľných parametrov $\theta_k, k = 1, \ldots, K-2$.

Riešením sústavy o dvoch neznámych θ_K a θ_{K-1},

$$\begin{aligned} \theta_K &= -\theta_{K-1} - \sum_{k=1}^{K-2} \theta_k, \\ \xi_K \theta_K &= -\xi_{K-1}\theta_{K-1} - \sum_{k=1}^{K-2} \xi_k \theta_k, \end{aligned} \tag{3.21}$$

dostaneme parameter θ_{K-1} v tvare

$$\theta_{K-1} = \frac{\xi_K \sum_{k=1}^{K-2} \theta_k - \sum_{k=1}^{K-2} \xi_k \theta_k}{(\xi_{K-1} - \xi_K)}. \tag{3.22}$$

Po dosadení θ_{K-1} do prvej rovnice (3.21) vyjadríme θ_K v tvare

$$\theta_K = \frac{\sum_{k=1}^{K-2} \xi_k \theta_k - \xi_{K-1} \sum_{k=1}^{K-2} \theta_k}{(\xi_{K-1} - \xi_K)}. \tag{3.23}$$

Aplikovaním reštrikcií (3.16) na regresnú funkciu (3.15) dostávame predpis v tvare

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3, \tag{3.24}$$

kde výraz $\sum_{k=1}^{K} \theta_k (x-\xi_k)_+^3$ si upravíme dosadením výrazov (3.23) a (3.22) do tvaru

$$\begin{aligned} \sum_{k=1}^{K} \theta_k (x-\xi_k)_+^3 = \sum_{k=1}^{K-2} \theta_k (x-\xi_k)_+^3 &+ \frac{\xi_K \sum_{k=1}^{K-2} \theta_k - \sum_{k=1}^{K-2} \xi_k \theta_k}{(\xi_{K-1} - \xi_K)} (x - \xi_{K-1})_+^3 \\ &+ \frac{\sum_{k=1}^{K-2} \xi_k \theta_k - \xi_{K-1} \sum_{k=1}^{K-2} \theta_k}{(\xi_{K-1} - \xi_K)} (x - \xi_K)_+^3 . \end{aligned}$$

Označme si z_k regresor odpovedajúci koeficientu $\theta_k, k = 1, \ldots, K-2$. Potom z_k nadobúda pre $k = 1, \ldots, K-2$ tvar

$$z_k = (x-\xi_k)_+^3 + \frac{\xi_K - \xi_k}{\xi_{K-1} - \xi_K} (x - \xi_{K-1})_+^3 + \frac{\xi_k - \xi_{K-1}}{(\xi_{K-1} - \xi_K)} (x - \xi_K)_+^3. \tag{3.25}$$

Poznámka. Jednoduchou úpravou (3.25) sa dostaneme ku rovnakému tvaru regresorov, aký je uvedený v [12, str. 20].

Zhrnutím vyššie prevedených výpočtov sa dostávame k predpisu regresného modelu prirodzeného splinu s K uzlovými bodmi $\xi_1, \ldots, \xi_K$, a K bázovými funkciami v tvare

$$Y_i = \sum_{k=0}^{K-1} \beta_k f_k(x_i) + \epsilon_i, \ i = 1, \ldots, n, \tag{3.26}$$

kde

$$f_0(x_i) = 1, \qquad f_1(x_i) = x_i,$$

a pre $k = 1, \ldots, K-2$ je

$$f_{k+1}(x_i) = (x_i - \xi_k)_+^3 + \frac{\xi_K - \xi_k}{\xi_{K-1} - \xi_K}(x_i - \xi_{K-1})_+^3 + \frac{\xi_k - \xi_{K-1}}{(\xi_{K-1} - \xi_K)}(x_i - \xi_K)_+^3. \quad (3.27)$$

Opäť, podobne ako pri modeli regresného splinu (3.10), môžeme uvažovaním $\boldsymbol{\epsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ v modeli (3.26) prostredníctvom F-testu (veta 6) a hypotézou $H_0 : \beta_2 = \ldots = \beta_{K-1} = 0$ testovať platnosť podmodelu modelu prirodzeného splinu (3.26), a to model regresnej priamky.

Odhady regresných koeficientov modelu prirodzeného splinu (s K uzlovými bodmi) môžeme vypočítať v softvéri R (prostredníctvom funkcie $lm()$) pomocou funkcie $ns()$, ktorá generuje (regresnú) maticu s B-spline bázovými funkciami modelu prirodzeného splinu.

Poznámka. Je potrebné si uvedomiť, že na rozdiel od funkcie $bs()$ argumentom `Boundary.knots` funkcie $ns()$ udávame uzly, od ktorých je regresná funkcia prirodzeného splinu lineárna, tj. uzly ξ_1 a ξ_K.

Ďalší možný spôsob získania odhadu regresnej funkcie prirodzeného splinu je zavedením lineárnych obmedzení (viď sekciu 2.4) na vektor regresných koeficientov $\boldsymbol{\beta}$ v modeli regresných splinov s regresnou funkciou v tvare (3.15). Uvažujme lineárne obmedzenia $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$ na základe reštrikcií (3.16), (3.19) a (3.20) v tvare

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 1 \\ 0 & 0 & 0 & 0 & \xi_1 & \cdots & \xi_K \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \theta_1 \\ \vdots \\ \theta_K \end{pmatrix} \text{ a } \mathbf{c} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \quad (3.28)$$

Vektor odhadov regresných koeficientov $\hat{\boldsymbol{\beta}}_0$, ktorý spĺňa podmienku $\mathbf{T}\hat{\boldsymbol{\beta}}_0 = \mathbf{c}$ vypočítame podľa vzorca (2.21) v sekcii 2.4. Odhad regresnej funkcie prirodzeného splinu je daný jednoznačne vektorom $\hat{\boldsymbol{\beta}}_0$.

Poznámka. V priloženom súbore môžeme nájsť zdrojový kód, kde sme ilustrovali zhodnosť odhadov regresných funkcií (s pevne zvolenými troma vnútornými uzlami), vypočítaných troma rôznymi prístupmi, a to:

1. Výpočtom odhadov regresných koeficientov modelu (3.26) prostredníctvom nami naprogramovanej funkcie $ps()$, ktorá nám generuje regresnú maticu s danými bázovými funkciami.

2. Výpočtom odhadov regresných koeficientov prostredníctvom funkcie $ns()$.

3. Výpočtom odhadov regresných koeficientov za platnosti lineárneho obmedzenia v tvare (3.28) (pre K=3).

V tejto kapitole sme si predstavili model regresného splinu a jeho podmodel - model prirodzeného splinu. Teraz budeme skúmať rozdiel medzi odhadmi regresných funkcií daných modelov za platnosti väčšieho modelu, teda modelu regresného splinu. Podobný problém je riešený v článku [17], kde autor skúma rozdiel medzi odhadmi regresných funkcií regresného modelu viacnásobnej regresie a jeho podmodelu za platnosti väčšieho modelu.

3.7 Porovnanie presnosti odhadu regresnej funkcie v modeli regresného a prirodzeného splinu

Ako sme už spomenuli na začiatku predošlej sekcie 3.6, odhady regresných funkcií v modeli regresného splinu (3.10) sú vo chvostoch variabilné, čo vedie k nepresnej predpovedi individuálnej, či odhadu strednej hodnoty závislej premennej na krajoch a mimo interval nameraných hodnôt nezávislej premennej. Tento problém môžeme potenciálne vyriešiť uvažovaním modelu prirodzeného splinu (3.26), keďže ako si v tejto sekcii ukážeme, zavedením reštrikcií zaručujúcich linearitu regresnej funkcie v modeli regresného splinu v hraničných častiach intervalu $[min(x_i), max(x_i)]$, získame v obecnosti odhady strednej hodnoty závislej premennej v novom bode $\mathbf{x}_*$ s potenciálne menším rozptylom.

Poznámka. Slovo *kubický* budeme v nasledujúcom texte vypúšťať, keďže zo sekcie 3.6 je zrejmé, že model prirodzených splinoch získame práve z modelu kubického regresného splinu.

Uvažujme platnosť modelu regresného splinu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ s K (vnútornými) uzlovými bodmi $\xi_k, k = 1, \ldots, K$ a regresnou maticou $\mathbf{X}$ generovanou bázovými funkciami (3.11) (d=3). My napriek tomu predpokladajme platnosť nesprávneho modelu, a to modelu prirodzeného splinu. Ten získame aplikovaním lineárnych obmedzení $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$ tak, ako sme si predviedli v sekcii 3.6. Označme $\hat{\boldsymbol{\beta}}_0$ vektor odhadnutých regresných koeficientov modelu prirodzených splinov. Podľa vzorca (2.22) platí

$$\mathsf{var}\hat{\boldsymbol{\beta}}_0 = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} - \sigma^2 \underbrace{\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{T}^\top\left(\mathbf{T}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{T}^\top\right)^{-1}\mathbf{T}(\mathbf{X}^\top\mathbf{X})^{-1}\right)}_{\boldsymbol{\Omega}}. \quad (3.29)$$

Ďalej vieme, že $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \mathsf{var}\hat{\boldsymbol{\beta}}$ (viď Veta 3), kde $\hat{\boldsymbol{\beta}}$ je vektor odhadnutých regresných koeficientov modelu regresného splinu.

Pre odhad strednej hodnoty závislej premennej v bode $\mathbf{x}_* \in \mathcal{M}(\mathbf{X}^\top)$, v modeli prirodzeného splinu, ktorý si označíme $\hat{\eta}_*^{(0)}$, a jeho rozptyl platí

$$\hat{\eta}_*^{(0)} = \mathbf{x}_*^\top\hat{\boldsymbol{\beta}}_0, \quad (3.30)$$

$$\mathsf{var}(\hat{\eta}_*^{(0)}) = \mathbf{x}_*^\top\mathsf{var}(\hat{\boldsymbol{\beta}}_0)\mathbf{x}_* = \mathsf{var}(\hat{\eta}_*) - \sigma^2\mathbf{x}_*^T\boldsymbol{\Omega}\mathbf{x}_* \leq \mathsf{var}(\hat{\eta}_*), \quad (3.31)$$

kde $\mathsf{var}(\hat{\eta}_*)$ je rozptyl odhadu strednej hodnoty v bode $\mathbf{x}_*$ v modeli regresného splinu, $\hat{\eta}_* = \mathbf{x}_*^\top\hat{\boldsymbol{\beta}}$. Posledná nerovnosť platí, keďže matica $\boldsymbol{\Omega}$ je pozitívne semidefinitná.

Nerovnosť (3.31) implikuje, že rozptyl odhadu strednej hodnoty závislej premennej, teda odhadu regresnej funkcie v bode $\mathbf{x}_*$, nebude v modeli prirodzeného splinu (oproti rozptylu odhadu v modeli regresného splinu) nikdy väčší.

Naopak, ako si ukážeme nižšie, vychýlenie odhadu $\hat{\eta}_*^{(0)}$ je potenciálne nenulové, kdežto nenulovosť odhadu strednej hodnoty závislej premennej v modeli regresného splinu bez lineárneho obmedzenia na regresné koeficienty priamo plynie z definície 5.

Pre vychýlenie odhadu $\hat{\eta}_*^{(0)}$ od skutočnej hodnoty regresnej funkcie $\eta_* = \mathbf{x}_*^\top \boldsymbol{\beta}$ platí

$$\mathsf{bias}(\hat{\eta}_*^{(0)}, \eta_*) = \mathbb{E}\,\hat{\eta}_*^{(0)} - \eta_*, \tag{3.32}$$

kde pre $\mathbb{E}\,\hat{\eta}_*^{(0)}$ s využitím vzorca (2.21) platí

$$\begin{aligned}
\mathbb{E}\,\hat{\eta}_*^{(0)} &= \mathbb{E}\mathbf{x}_*^\top \hat{\boldsymbol{\beta}}_0 \\
&= \mathbb{E}(\mathbf{x}_*^T(\hat{\boldsymbol{\beta}} - \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{T}^T(\mathbf{T}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{T}^T)^{-1}}_{\mathbf{A}}(\mathbf{T}\hat{\boldsymbol{\beta}} - \mathbf{c})) \\
&= \mathbf{x}_*^T\mathbb{E}\hat{\boldsymbol{\beta}} - \mathbf{x}_*^T\mathbf{A}\mathbf{T}\mathbb{E}\hat{\boldsymbol{\beta}} + \mathbf{x}_*^T\mathbf{A}\mathbf{c} \\
&= \mathbf{x}_*^T\boldsymbol{\beta} - \mathbf{x}_*^T\mathbf{A}\mathbf{T}\boldsymbol{\beta} + \mathbf{x}_*^T\mathbf{A}\mathbf{c},
\end{aligned} \tag{3.33}$$

z čoho vyplýva

$$\mathsf{bias}(\hat{\eta}_*^{(0)}, \eta_*) = \mathbb{E}\,\hat{\eta}_*^{(0)} - \eta_* = \mathbf{x}_*^\top\boldsymbol{\beta} - \mathbf{x}_*^\top\mathbf{A}\mathbf{T}\boldsymbol{\beta} + \mathbf{x}_*^\top\mathbf{A}\mathbf{c} - \mathbf{x}_*^T\boldsymbol{\beta} = \mathbf{x}_*^\top\mathbf{A}(\mathbf{c} - \mathbf{T}\boldsymbol{\beta}). \tag{3.34}$$

Ako sme si ukázali, uvažovaním "nesprávneho"modelu prirodzených splinov sme za platnosti väčšieho modelu regresných splinov získali odhad strednej hodnoty závislej premennej, ktorý je potenciálne vychýlený, no v žiadnom bode $\mathbf{x}_*$ nemá oproti nevychýlenému odhadu strednej hodnoty v modeli regresných splinov väčší rozptyl.

Charakteristika, na základe ktorej budeme porovnávať presnosť týchto odhadov sa nazýva *stredná štvorcová chyba* a je určená nasledujúcou definíciou.

Definícia 11. *Pre odhad $\hat{\theta}$ skalárneho parametru θ definujeme strednú štvorcovú chybu odhadu ako*

$$\mathsf{MSE}(\hat{\theta}) = \mathbb{E}\left(\hat{\theta} - \theta\right)^2. \tag{3.35}$$

Tvrdenie 11. *Pre strednú štvorcovú chybu odhadu platí*

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{var}(\hat{\theta}) + \left[\mathsf{bias}(\hat{\theta}, \theta)\right]^2. \tag{3.36}$$

Dôkaz.

$$\begin{aligned}
\mathbb{E}(\hat{\theta} - \theta)^2 &= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta}) + (\mathbb{E}\hat{\theta} - \theta)\right]^2 \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2 + (\mathbb{E}\hat{\theta} - \theta)^2 + 2(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta)\right] \\
&= \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 + \mathbb{E}(\mathbb{E}\hat{\theta} - \theta)^2 + 2\mathbb{E}\left[\hat{\theta}\mathbb{E}\hat{\theta} - (\mathbb{E}\hat{\theta})^2 - \theta\hat{\theta} + \theta\mathbb{E}\hat{\theta}\right] \\
&= \mathsf{var}(\hat{\theta}) + \left[\mathsf{bias}(\hat{\theta},\theta)\right]^2 + 2\left[(\mathbb{E}\hat{\theta})^2 - (\mathbb{E}\hat{\theta})^2 - \theta\mathbb{E}\hat{\theta} + \theta\mathbb{E}\hat{\theta}\right] \\
&= \mathsf{var}(\hat{\theta}) + \left[\mathsf{bias}(\hat{\theta}, \theta)\right]^2.
\end{aligned}$$

❑

Porovnajme odhady strednej hodnoty závislej premennej našich modelov pomocou vyššie definovanej strednej štvorcovej chyby. Platí

$$\begin{aligned} \mathsf{MSE}(\hat{\eta}_*) &= \mathsf{var}(\hat{\eta}_*), \\ \mathsf{MSE}(\hat{\eta}_*^{(0)}) &= \mathsf{var}(\hat{\eta}_*) - \sigma^2 \mathbf{x}_*^\top \mathbf{\Omega} \mathbf{x}_* + \left[\mathbf{x}_*^\top \mathbf{A}(\mathbf{c} - \mathbf{T}\boldsymbol{\beta})\right]^2. \end{aligned} \tag{3.37}$$

Poznámka. Odhad s menšou *strednou štvorcovou chybou* je lepší.

Skúmajme, za akých podmienok je odhad strednej hodnoty závislej premennej na základe modelu prirodzeného splinu lepším ako na základe modelu regresných splinov.

$$\mathsf{MSE}(\hat{\eta}_*) > \mathsf{MSE}(\hat{\eta}_*^{(0)}) \iff \sigma^2 \mathbf{x}_*^T \mathbf{\Omega} \mathbf{x}_* > \left[\mathbf{x}_*^\top \mathbf{A}(\mathbf{c} - \mathbf{T}\boldsymbol{\beta})\right]^2. \tag{3.38}$$

Z nerovnosti (3.38) môžeme pozorovať, že odhad strednej hodnoty závislej premennej modelu prirodzeného splinu bude lepší za väčšieho poklesu v rozptyle odhadu ako je vo štvorci jeho vychýlenia.

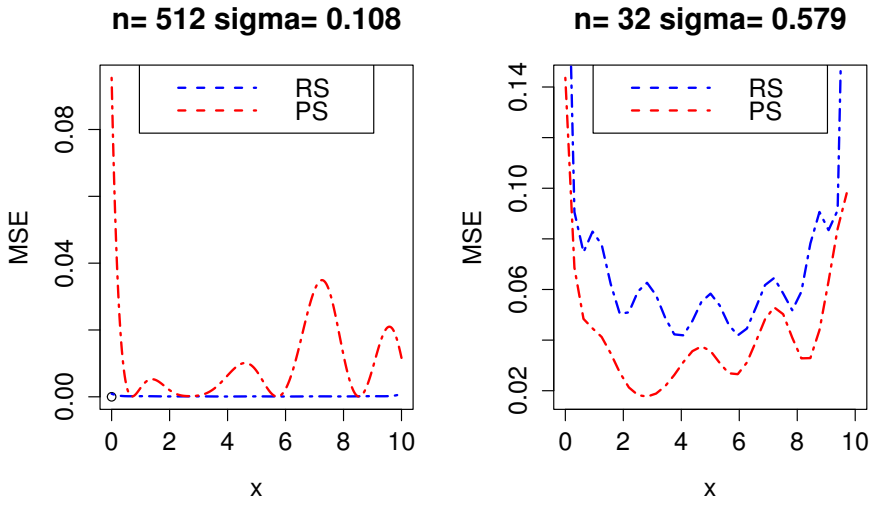
Dalo by sa skúmať, aký typ dát danú nerovnosť (určite) spĺňa, no tejto problematike sa v práci venovať nebudeme. Namiesto toho preberieme výsledky z článku [17], ktoré uvádzajú, že pre zvyšujúcu sa smerodajnú odchýlku chybovej zložky σ a zmenšujúci sa počet pozorovaní n bude mať MSE odhadu strednej hodnoty závislej premennej v podmodeli tendenciu byť nižšia ako MSE odhadu strednej hodnoty vo väčšom modeli. Túto teóriu si budeme teraz ilustrovať na príklade.

Uvažujme platnosť modelu kubického splinu s troma pevne zvolenými uzlami rovnomerne rozmiestnenými na intervale hodnôt nezávislej premennej. Pomocou funkcie *generateSpline*() si náhodne[1] vygenerujeme vektor regresných koeficientov, ktoré nám určujú skutočný tvar regresnej funkcie, a tú si následne vyhodnotíme v nami uvažonom maximálnom počte hodnôt nezávislých premenných. Uvažujeme, že hodnoty nezávislej premennej sú rovnomerne rozložené na intervale $[0, 10]$.

Vygenerujeme si dve sady pozorovaní. Prvú s počtom pozorovaní $n_1 = 512$ smerodajnou odchýlkou chybovej zložky $\sigma_1 = 0.108$ a druhú s počtom pozorovaní $n_2 = 32$ a smerodajnou odchýlkou chybovej zložky $\sigma_2 = 0.579$. Vďaka znalosti presného predpisu skutočnej regresnej funkcie si môžeme ľahko dopočítať hodnoty MSE (použitím vzorcov (3.37)) odhadov strednej hodnoty zavislej premennej v modeli prirodzeného a regresného splinu.

Na obrázku 3.7 vidíme na ľavom grafe MSE odhadu strednej hodnoty závislej premennej v modeli regresného splinu (modrá čiara) a odhadu strednej hodnoty v modeli prirodzeného splinu (červená čiara) prvého datasetu, teda pri počte pozorovaní $n_1 = 512$ a $\sigma_1 = 0.108$ a na pravom grafom eventuálne MSE odhadov druhého datasetu s počtom pozorovaní $n_2 = 32$ a $\sigma_2 = 0.579$. Obdobné výsledky dostávame zmenou hodnotou *seed*, teda generovaním iných regresných funkcií, ako je nami použitá v tejto ukážke. Z obrázku 3.7 môžeme vidieť, že dané výsledky sú v súlade s nami vyššie popísanou teóriu.

[1] reprodukovateľnosť výsledkov je zaručená funkciou *set.seed*()



Obr. 3.9: Graf priebehu MSE odhadu regresnej funkcie modelu regresného splinu (RS - modrá farba) a MSE odhadu regresnej funkcie modelu prirodzeného splinu (PS - červená čiara) za platnosti modelu regresného (kubického) splinu s troma pevne zvolenými uzlami rovnomerne rozmiestnenými na intervale $[0, 10]$.

3.7.1 Simulácie Monte Carlo

Na posúdenie toho, pri akom type dát[2] je model prirodzených splinov lepší ako model regresných splinov s tým, že sme uvažovali platnosť modelu regresného splinu s pevne zvolenými troma uzlovými bodmi, sme v prechádzajúcej sekcii použili strednú štvorcovú chybu odhadu (regresnej funkcie). Na jej výpočet je však potrebná znalosť skutočnej regresnej funkcie, ktorú v praxi nepoznáme. V praxi teda na zhodnotenie toho, ktorý z modelov[3] popisuje závislosť medzi strednou hodnotou závislej a hodnotami nezávislej premennej lepšie, používame mimo iné napr. $R^2_{adj}, PRESS$ štatistiku, či F-test (viď sekcia 2.4 a 2.5.3).

V tejto časti práce budeme chcieť pomocou jednoduchých Monte Carlo simulácií posúdiť, do akej miery sme schopní na základe týchto charakteristík korektne vybrať skutočne lepší model. Model, ktorý je skutočne lepší, budeme posudzovať pomocou MSE odhadu regresnej funkcie daných modelov.

Budeme uvažovať rôzne počty pozorovaní $n_j = 2^{10-j}, j = 1, \ldots, 6$ a rôzne hodnoty smerodajnej odchýlky chybovej zložky $\sigma_l = 0.02 * 1.4^{2(l-1)+1}, l = 1, \ldots, 6$. Označme ${}_{n_j}\mathbf{x}, j = 1, \ldots, 6$ vektor hodnôt nezávislej premennej dĺžky n_j a ${}_{n_j}\boldsymbol{\eta}$ vektor hodnôt skutočnej regresnej funkcie v bodoch ${}_{n_j}\mathbf{x}$, rovnomerne rozmiestnených na intervale $[0, 10]$. Pre každú hodnotu $\sigma_l, l = 1, \ldots, 6$ si vygeneruje vektor hodnôt závislej premennej ${}_{n_j}\mathbf{Y}$ dĺžky $n_j, j = 1, \ldots, 6$ ako

$$ {}_{n_j}\mathbf{Y} = {}_{n_j}\boldsymbol{\eta} + {}_{n_j}\boldsymbol{\epsilon}, $$

kde ${}_{n_j}\boldsymbol{\epsilon} \sim \mathsf{N}_{n_j}(\mathbf{0}, \sigma_l^2 \mathbf{I}_{n_j}), j = 1, \ldots, 6, l = 1, \ldots, 6$, čím získame celkovo 36 rôznych kombinácií počtu pozorovaní a hodnoty smerodajnej odchýlka chybovej zložky.

[2] tj. pre aký počet pozorovaní n a smerodajnú odchýlku chybovej zložky σ

[3] nielen modelu prirodzeného a regresného splinu, ale rôznych modelov v obecnosti

Regresná funkcia, ktorú sme vygenerovali pomocou *generateSpline*() je pevná a pre rôzne kombinácie n_j a σ_l sa nemení. Jednotlivé simulácie sa líšia iba v realizovaných hodnotách chybovej zložky[4].

Za skutočne lepší model budeme považovať ten, ktorého hodnoty MSE vypočítané v primárnych hodnotách jednotlivých vektorov ${}_{n_j}\mathbf{x}, j = 1,\ldots,6,$, sú vo všetkých hodnotách daného vektoru menšie ako u druhého modelu. Ak by žiadny z modelov nemal nižšiu MSE vo všetkých hodnotách vektoru, za skutočne lepší model budeme považovať ten, ktorého priemerná MSE je nižšia.

Upravený koeficient determinácie

Uvažujme počet simulácií S. Pre každú simuláciu dát $s = 1,...,S$ rôznej kombinácie $n_j, j = 1,\ldots,6$ a $\sigma_l, l = 1,\ldots,6$, si odhadneme model regresného a prirodzeného splinu. Pre oba odhady si zistíme hodnotu upraveného koeficientu determinácie (viď sekcie 2.5.3). Položme $Z_s = g({}^s_{PS}R^2_{adj}, {}^s_{RS}R^2_{adj}), s = 1,...,S$, kde ${}^s_{RS}R^2_{adj}$ predstavuje hodnotu R^2_{adj} odhadu modelu regresného splinu s-tej simulácie. ${}^s_{PS}R^2_{adj}$ si definujeme obdobne pre odhad modelu prirodzeného splinu. Funkciu $g(\cdot,\cdot)$ si definujeme predpisom

$$g(x,y) = \begin{cases} 1, & x > y, \\ 0, & \text{inak.} \end{cases}$$

Potom platí, že $Z_1,...,Z_S$ je náhodný výber z alternatívneho rozdelenia Alt(p), $p \in (0,1)$, kde p označuje pravdepodobnosť, že odhad modelu regresného splinu je na základe hodnoty R^2_{adj} horší ako odhad modelu prirodzeného splinu. Nestranným odhadom parametru p je relatívna početnosť

$$\hat{p} = \frac{\sum_{s=1}^{S} Z_s}{S} = \bar{Z}.$$

V tabuľke 3.1 nižšie môžeme vyčítať hodnoty relatívnej početnosti $\hat{p}$ pre rôzne kombinácie n_j a σ_l. Modrou farbou sú vyznačené hodnoty, v ktorých je priemerná MSE odhadu regresnej funkcie regresného splinu menšia ako priemerná MSE odhadu regresnej funkcie prirodzeného splinu. Hrubou modrou kurzívou je vyznačená hodnota, kde je hodnota MSE menšia v každom bode odhadnutej regresnej funkcie regresného splinu, v ktorom bola MSE vyhodnotená. Obdobný vzťah platí pre červenú farbu a odhad regresnej funkcie prirodzeného splinu.

n_j a σ_l	**0.028**	**0.055**	**0.108**	**0.211**	**0.413**	**0.810**
512	0.00	0.00	0.00	0.00	0.00	0.08
256	0.00	0.00	0.00	0.00	0.01	0.21
128	0.00	0.00	0.00	0.00	0.07	0.36
64	**0.00**	0.00	0.00	0.01	0.21	**0.45**
32	0.00	0.00	0.00	0.08	0.35	**0.50**
16	0.00	0.00	0.01	0.19	**0.43**	**0.51**

Tabuľka 3.1: Tabuľka relatívnych početností $\hat{p}$ rôznych kombinácií n_j a σ_l.

[4]reprodukovateľnosť výsledkov je opäť zaručená pevnou hodnotou *seed*

Podľa [3, Veta 2.10] platí

$$\mathsf{var}\hat{p} = \mathsf{var}(\bar{Z}) = \frac{p(1-p)}{S}. \tag{3.39}$$

Smerodajnú chybu odhadu $\hat{p}$, teda chybu spôsobenú konečným počtom simulácií, môžeme pre $S = 2500$ obmedziť zhora hodnotou 0.01, keďže

$$\mathsf{S.E.}(\hat{p}) = \sqrt{\frac{p(1-p)}{S}} \leq \frac{0.5}{\sqrt{2500}} = 0.01. \tag{3.40}$$

PRESS štatistika

Inou charakteristikou modelu, na základe ktorej môžeme rozhodnúť, ktorý z modelov popisuje vzťah medzi strednou hodnotou závislej premennej a hodnotami nezávislej premennej je $PRESS$ štatistika, bližšie popísaná v sekcii 2.5.3.

Tabuľku relatívnych početností získame analogickým spôsobom ako je popísaný v predošlej časti s tým, že tabuľka bude prezentovať relatívne početnosti, v ktorých bol zvolený ako lepší model na základe $PRESS$ štatistiky opäť model prirodzeného splinu. Kvôli následnej interpretácií výsledkov si pravdepodobnosť, že na základe $PRESS$ štatistiky vyberieme model prirodzeného splinu označme q. Farebné odlíšenie hodnôt nachádzajúcich sa v tabuľke je totožné s tabuľkou 3.1 .

n a $\boldsymbol{\sigma}$	**0.028**	**0.055**	**0.108**	**0.211**	**0.413**	**0.810**
512	0.00	0.00	0.00	0.00	0.00	0.31
256	0.00	0.00	0.00	0.00	0.06	0.56
128	0.00	0.00	0.00	0.00	0.33	0.75
64	**0.00**	0.00	0.00	0.10	0.60	**0.83**
32	0.00	0.00	0.03	0.45	0.78	**0.87**
16	0.00	0.11	0.56	0.83	**0.92**	**0.94**

Tabuľka 3.2: Tabuľka relatívnych početností výberu modelu prirodzeného splinu na základe $PRESS$ štatistiky pre rôzne kombinácie n_j a σ_l.

Smerodajnú chybu jednotlivých relatívnych početností pre rôznu kombináciu n_j a σ_l môžeme podobne ako v predchádzajúcej časti obmedziť zhora hodnotou 0.01.

F-test

Posledným prístupom rozhodovania sa, ktorý z odhadov regresnej funkcie uvažovaného modelu a podmodelu popisuje lepšie vzťah medzi závislou a nezávislou premennou, je prostredníctvom testovej štatistiky definovanej vo vete 6. Keďže model prirodzených splinov je podmodelom modelu regresných splinov, je možné prostredníctvom F-testu rozhodnúť o zamietnutí resp. nezamietnutí platnosti modelu prirodzeného splinu.

Uvažujme bežnú hladinu významnosti $\alpha = 0.05$. Symbolom $p_s, s = 1, \ldots, S$, si označme p-hodnotu s-tej simulácie spomenutého testu. Podobne ako v predchádzajúcich dvoch častiach si náhodné veličiny $Z_s = g(p_s, \alpha), s = 1, \ldots, S$, tvoria

opäť náhodný výber z alternatívneho rozdelenia $\mathsf{Alt}(r), r \in (0,1)$, kde r predstavuje pravdepodobnosť, že na základe F-testu nezamietame nulovú hypotézu o platnosti podmodelu (pri stanovenej hodnote hladiny významnosti α). Pravdepodobnosť r odhadneme opäť relatívnou početnosťou $\hat{r} = \bar{Z}$.

V tabuľke nižšie môžeme vidieť relatívnu početnosť $\hat{r}$ pre rôzne kombinácie n_j a σ_l. Chybu odhadu $\hat{r}$ môžeme takisto ako v predchádzajúcich častiach tejto sekcie obmedziť zhora hodnotou 0.01 (pre počet $S = 2500$). Farebné odlíšenie hodnôt nachádzajúcich sa v tabuľke je opäť totožné s tabuľkou 3.1 .

n a $\boldsymbol{\sigma}$	**0.028**	**0.055**	**0.108**	**0.211**	**0.413**	**0.810**
512	0.00	0.00	0.00	0.00	0.00	0.41
256	0.00	0.00	0.00	0.00	0.10	0.67
128	0.00	0.00	0.00	0.01	0.43	0.84
64	**0.00**	0.00	0.00	0.13	0.71	**0.89**
32	0.00	0.00	0.01	0.47	0.83	**0.92**
16	0.00	0.00	0.26	0.76	**0.91**	**0.93**

Tabuľka 3.3: Tabuľka relatívnych početností výberu modelu prirodzeného splinu na základe F-testu pre rôzne kombinácie n_j a σ_l.

Zhrnutie výsledkov

Ako vidíme v tabuľkách 3.1, 3.2 a 3.3, relatívne početnosti majú pre rôzne voľby $n_j, j = 1, \ldots, 6$, a $\sigma_l, l = 1, \ldots, 6$, tendenciu stúpať smerom k pravému dolnému rohu tabuľky, tzn. k miestu kde sa počet pozorovaní n_j zmenšuje a smerodajná odchýlka chybovej zložky σ_l zväčšuje. Tvar náhodnej veličiny $Z_s, s = 1, \ldots, S$ bol v jednotlivých častiach tejto sekcie volený tak, aby jednotlivé pravdepodobnosti p, q a r odpovedali pravdepodobnosti toho, že odhad regresnej funkcie modelu prirodzeného splinu je (na základe $R^2_{adj}, PRESS$ štatistiky a p-hodnoty F-testu s hladinou $\alpha = 0.05$) lepší ako odhad regresnej funkcie modelu regresného splinu. Hodnoty v daných tabuľkách sú niekedy v kontraste so skutočnosťou, keďže niekedy by sme podľa hodnôt relatívnej početnosti mali rozhodnúť pre model, ktorý skutočne na základe priemernej MSE alebo hodnoty MSE nižšej v každom vyhodnotenom bode neplatí. Z týchto hodnôt môžeme teda pozorovať, že tieto ukazovatele nie sú vždy vhodné pre výber "správneho"modelu. Treba však zohľadniť, že nami získané výsledky môžeme posúdiť iba kvalitatívne, keďže sme uvažovali iba jednu pevnú regresnú funkciu s pevne zvolenými uzlovými bodmi, no zmenou hodnotou *seed*, teda generovaním inej regresnej funkcie, boli výsledky veľmi podobné. Naše výsledky by sme mohli prípadne zovšeobecniť uvažovaním rôznych regresných funkcií.

Kapitola 4

Záver

V tejto bakalárskej práci sme sa zaoberali parametrickými modelmi regresných splinov. Na začiatku práce sme zadefinovali základné pojmy regresnej analýzy neskôr využívané v kapitole 3. V danej kapitole sme následne diskutovali základne vlastnosti rôznych regresných funkcií parametrických modelov so zameraním na model regresných splinov. Predstavili sme si dve bázy slúžiace na reprezentáciu regresných splinov a to useknutú mocninovú bázu a tzv. B-spline bázu. Predstavili sme si model prirodzených splinov a jeho použitie zo zámerom možného zlepšenia presnosti odhadu regresnej funkcie. Následne sme samostatne odvodili tvar bázových funkcií daného modelu. Ďalej sme diskutovali základne prístupy voľby polohy a počtu uzlových bodov.

V sekcii 3.7 sme sa venovali porovnávaniu presnosti odhadov regresných funkcií modelu regresného a prirodzeného splinu. Odvodili sme vzorec strednej štvorcovej chyby odhadu a následne sme kvalitatívne diskutovali okolnosti, za ktorých je vhodné použiť model prirodzeného splinu a túto teóriu sme si ilustrovali na jednoduchom príklade.

V poslednej časti práce sme sa Monte Carlo simuláciou snažili skúmať, či ukazovatele, ktoré sú v praxi bežne používané pre výber modelu, dokážu správne rozhodnúť v prospech skutočne platného modelu. Na základe nami získaných výsledkov, ktoré sú zachytené v tabuľkách sme dospeli k záveru, že tieto ukazovatele nedokážu vždy správne zachytiť platnú skutočnosť.

Literatúra

[1] Zvára, K. (2008): *Regrese*. Matfyzpress, Praha. ISBN 978-80-7378-014-8

[2] Zvára, K. (1989): *Regresní analýza*. Academia, Praha. ISBN 9788020001252

[3] Anděl, J. (2005): *Základy matematické statistiky*. Matfyzpress, Praha. ISBN 80-86732-40-1

[4] Cipra, T. (2006): *Finanční a pojistné vzorce*. Prvé vydanie. GRADA, Praha. ISBN 80-247-1633-X

[5] Cipra, T. (2008): *Finanční ekonometrie*. Prvé vydanie. Ekopress. ISBN 9788086929439

[6] Abadir, K. M. & Magnus, J. R. (2005): *Matrix Algebra*. Cambridge University Press, New York. ISBN 978-0-521-82289-3

[7] Bican, L. (2009): *Lineární algebra a geometrie*. ACADEMIA, Praha. ISBN 978-80-200-1707-9

[8] Rencher, A.C. & Schaalje, G.B. (2008): *Linear Models in Statistics*. Druhé vydanie. John Wiley & Sons. ISBN 9780470192603

[9] de Boor, C. (2001): *A Practical Guide to Splines*. Springer, USA. ISBN 0-387-395366-3

[10] Dierckx, P. (1995): *Curve and Surface Fitting with Splines*. Clarendon Press. ISBN 9780198534402

[11] James, G. & Witten, D. & Hastie, T. & Tibshirani, R. (2013): *An Introduction to Statistical Learning (with Applications in R)*. Springer, USA. ISBN 978-1-4614-7137

[12] Frank E. Harrell, Jr. (2001): *Regression Modeling Strategies*. Springer, New York. ISBN 978-1-4419-2918-1

[13] Ruppert, D. & Wand, M. P. & Carroll, R. J. (2003): *Semiparametric Regression*. Cambridge University Press, New York. ISBN-13 978-0-511-06683-2

[14] Kagerer, K. (2013): *A short introduction to splines in least squares regression analysis*(online). Diskusný príspevok, University of Regensburg. Posledná aktualizácia: 28.3.2013 [cit. 16.3.2015], URL `http://epub.uni-regensburg.de/27968/1/DP472_Kagerer_introduction_splines.pdf`

[15] Ramsay, J.O. & Silverman, B.W. (2005): *Functional Data Analysis*. 2. vydanie. Springer ISBN-13 978-0387-40080-8

[16] Hastie, T. & Tibshirani, R. & Friedman, J. (2009): *The Elements of Statistical Learning (Data Mining, Inference, and Prediction)*. Springer. 2. vydanie. ISBN 978-0-387-84858-7

[17] Harris, T. J. & Wu, S. & McAuley, K. B. (2007) : *The Use of Simplified or Misspecified Models: Linear Case*. The Canadian Journal of Chemical Engineering. Zväzok 85, vydanie 4. Posledná aktualizácia: 19.5.2008 [cit. 20.4.2015], URL `http://onlinelibrary.wiley.com/doi/10.1002/cjce.5450850401/pdf`.

Zoznam obrázkov

Zoznam tabuliek