UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

BAKALÁŘSKÁ PRÁCE

Nelineární regrese pomocí GAM

Vedoucí bakalářské práce:
doc. RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2015

Vypracovala:
Bára Hronová
ME, III. ročník

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bára Hronová

Název práce: Nelineární regrese pomocí GAM

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2015

Abstrakt: U regresních modelů jako jsou například klasický lineární regresní model nebo logistická regrese je předpokládán lineární vztah mezi střední hodnotou závislé proměnné a vysvětlujícími proměnnými, které jsou lineární funkcí neznámých parametrů. V této práci popisuji zobecněné aditivní modely, které nahrazují lineární funkci součtem neznámých hladkých funkcí, které odhadneme pomocí použití vyhlazujících kubických splajnů v iteračním procesu, který se nazývá backfitting algoritmus. Tuto metodu můžeme použít na jakékoliv pravděpodobnostní rozdělení závislé proměnné. Zobecněné aditivní modely jsou tak vhodné pro analýzu datového souboru, potažmo vztahů mezi proměnnými s využitím podrobné informace o charakteru zkoumané závislosti, obsažené v odhadnutých funkcích. Práce obsahuje též možnosti zpracování zobecněných aditivních modelů v prostředí statistického softwaru R a jejich vlastnosti jsou demonstrovány na reálných datech.

Klíčová slova: splajn, B-splajn, regrese, neparametrická regrese, zobecněný aditivní model, software R

Počet stran: 38

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bára Hronová

Title: Nonlinear regression using GAM

Type of thesis: Bachelor's

Department:
Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2015

Abstract: With regression models such as standard linear regression model or logistic regression there is assumed a linear relationship between mean value of the explanatory variable and predictors, which are linear function of unknown parametres. In this thesis I decribe generalized additive models, which replace a linear function with the sum of unknown smooth functions, estimated by using smooth cubic splines in an iterative process called backfitting algorithm. This method can be used for any distributions of the response. Generalized additive models are suitable for an analysis of a data set, or more precisely of relationship between variables using detailed information about the character of the examined dependence contained in the estimated functions. This thesis also covers the possibility of processing generalized additive models in statistical software R and their properties are demonstrated on real data.

Key words: spline, B-spline, regression, nonparametric regression, generalized additive model, software R

Number of pages: 38

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem vytvořila tuto bakalářskou práci samostatně za vedení doc. RNDr. Karla Hrona, Ph.D. a že jsem v seznamu použité literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci dne 17. dubna 2015

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce doc. RNDr. Karlu Hronovi, Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích. Dále si zaslouží poděkování můj počítač, že vydržel moje pracovní tempo, a typografický systém TEX, kterým je práce vysázena. A v neposlední řadě děkuji mým přátelům, rodině a snoubenci, kteří mě neustále podporují.

Obsah

Úvod

K vyšetření závislosti mezi vysvětlující a vysvětlovanou proměnnou se nejčastěji používají lineární modely, které jsou oblíbené díky své jednoduché interpretaci a explicitním vztahům pro odhadované parametry. Ve skutečnosti se ale málokdy setkáváme s typickou lineární závislostí a též chyby modelu často nejsou normálně rozděleny. Proto je lepší v takových případech použít obecnější modely. Úkolem této práce je tedy popsat metody, které nejsou založeny na linearitě, zejména pak metodu GAM, neboli zobecněný aditivní model.

V první kapitole se seznámíme se splajnovou aproximací a splajny samotnými, které jsou základními kameny GAM a jsou tedy potřebné k jejich definování. V další kapitole si připomeneme základy z regresní analýzy, které by měl znát každý, který prošel základním kurzem matematické statistiky. I když se tedy jedná o pojmy známé, jejich uvedení je nezbytné pro ucelenost práce a její přehlednost. V této kapitole připomeneme i logistickou regresi, která je zároveň nejjednodušším příkladem GLM (zobecněného lineárního modelu), takže plynule přejdeme k poslední teoretické kapitole. Ta popisuje nelineární regresi, jež parametricky modeluje nelineární závislost mezi vysvětlující a vysvětlovanou proměnou. Klíčový je zde popis modelu GAM, u kterého si uvedeme i algoritmus pro odhad parametrů v tomto modelu.

V závěrečné kapitole je popsána aplikace GAM na konkrétních příkladech, které jsou řešené ve statistickém programu *R*. V tomto programu jsem zároveň prováděla všechny výpočty a tvořila i všechny obrázky, které jsou doplněny v mé práci. Tato část obsahuje tři příklady, první je vlastní a výsledky dalších dvou jsou převzaté z literatury.

1 Metody rozšiřující bázi

Při tvorbě této kapitoly jsem čerpala hlavně z [5], [8], [10]. Hlavní myšlenka spočívá v tom, že vstupní p-rozměrnou proměnou $\mathbf{x} = (x_1, ..., x_p)'$ pomocí transformací dále rozšiřujeme nebo nahrazujeme tak, abychom v novém prostoru mohli užít standartní lineární regresní model. Nechť $h_m(\mathbf{x}) : \mathbb{R}_p \to \mathbb{R}$ je m-tá transformace $\mathbf{x}$, kde $m = 1, \ldots, M$. Lineární rozšíření báze v $\mathbf{x}$ se modeluje takto:

$$f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m h_m(\mathbf{x})$$

Příklady pro rozšíření báze:

- $h_m(\mathbf{x}) = x_m \quad m = 1, \ldots, p \quad \ldots$ popisuje vícenásobný regresní model,
- $h_m(\mathbf{x}) = x_j^2 \quad \ldots$ kvadratická transformace,
- $h_m(\mathbf{x}) = \ln(x_i) \quad \ldots$ nelineární transformace,
- $h_m(\mathbf{x}) = I(L_m \leq x_k < U_m) \ldots$ indikátorová funkce vede k modelům s konstantním příspěvkem pro x_k v intervalu $[L_m, U_m)$ resp. po částech konstantním v případě, že bude definováno více se nepřekrývajících oblastí.

1.1 Interpolace pomocí splajnů

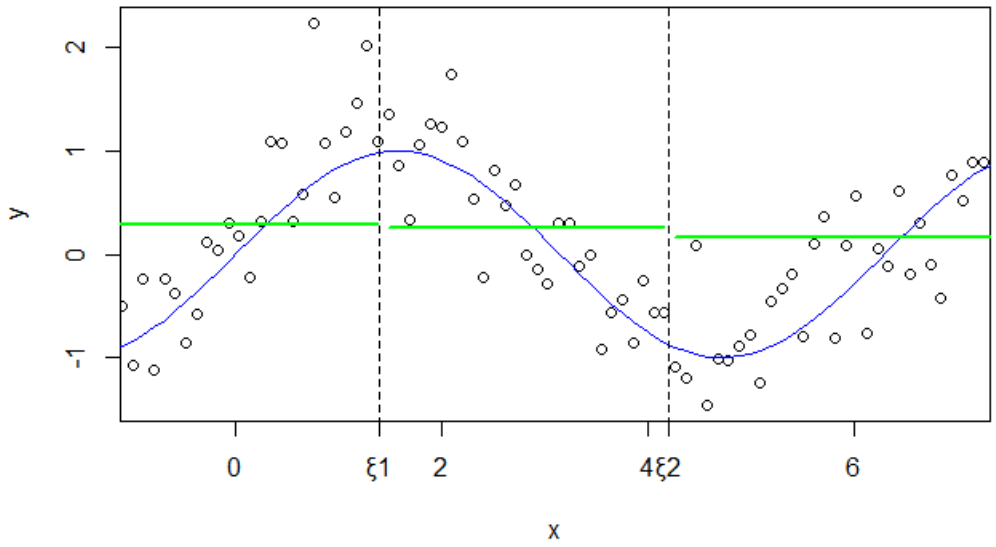
I když se díky své jednoduchosti a relativně rychlé konstrukci pro popis vztahu mezi vysvětlující a vysvětlovanou proměnnou (resp. proměnnými) nejvíce používá lineární regrese, někdy je potřeba přesnějšího popisu, kterého lze dosáhnout díky splajnům.

Název splajn pochází od speciálního pružného křivítka, které se používalo při konstrukci trupů lodí a později letadel. Je to úzká ohebná tyč, která se závažím zafixuje v určitých bodech a svou vlastní pružností se prohne tak, že vytvoří křivku takového tvaru, který vyhovuje konstruktérům. Její tvar je možno pak ještě regulovat pomocnými závažími.

Kdybychom tedy chtěli splajn popsat jednoduchou matematickou formulací, tak splajn je křivka, která vznikne spojením více na sebe navazujících polynomů. Díky své jednoduchosti jsou tedy polynomy stavebním kamenem pro tyto

splajny. Zakladatelem matematické teorie splajnů byl rumunsko-americký matematik Isaac Jacob Schoenberg (1903-1990).

Nejdříve popíšeme strukturu splajnových funkcí a pak si přiblížíme obvyklý systém používaný pro jejich konstrukci, a to B-splajnový systém.



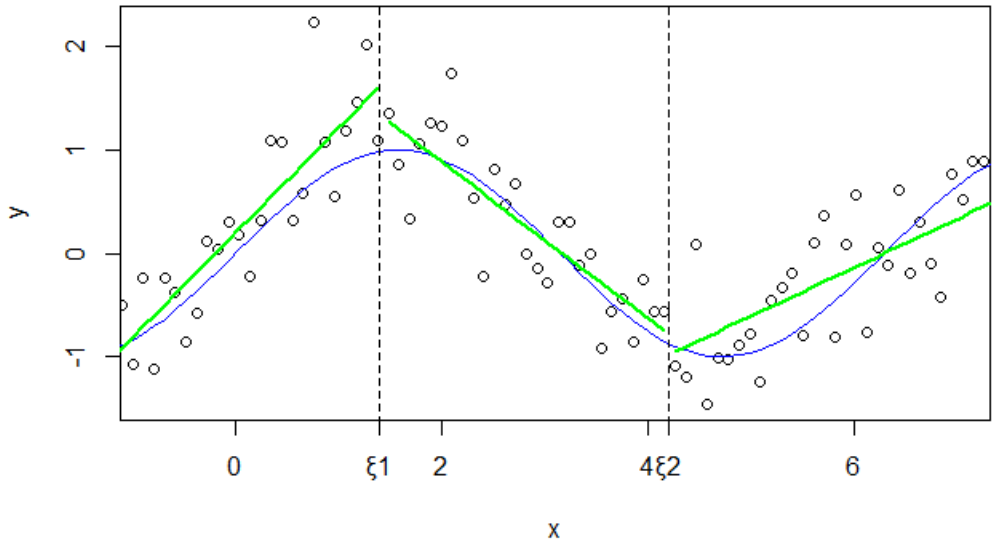
Obrázek 1: Po částech konstantní polynom (zeleně)

Prvním krokem k definici splajnu je rozdělení intervalu, přes který je funkce aproximována, do $k+1$ podintervalů, rozdělené hodnotami $\xi_i, i = 1, \dots k$, které se nazývají uzly. Na obrázku 1 vidíme, že dva uzly rozdělují interval do tří podintervalů. Přes každý interval $(-\infty, \xi_1), [\xi_1, \xi_2), \dots [\xi_{k-1}, \xi_k), [\xi_k, \infty)$ je definován polynom stupně nejvýše M. Po částech konstantní funkce jsou nejjednodušší složené polynomy. Složené polynomy stupně $M = 1, 2, 3$ se postupně nazývají složené lineární, kvadratické a kubické polynomy. Stupněm polynomu se myslí jeho nejvyšší mocnina a je tedy o jedno menší než řád, čímž se myslí počet konstant potřebných k jeho definování. K určení složeného polynomu stupně M s k uzly $\xi_1, \dots, \xi_k$ potřebujeme $(M+1)(k+1)$ neznámých parametrů, protože máme dohromady $k+1$ polynomů a každý má $M+1$ koeficientů. Například tedy na obrázku 3 jsou 3 polynomy stupně 1, to znamená, že v tomto případě máme $3 \times 2 = 6$ neznámých parametrů.

Obrázky 1-5 ukazují funkci sinus (modře) a k ní vytvořená simulovaná data. Na obrázku 1 se na každém podintervalu data aproximují konstantní funkcí a jejich bázové funkce mají tento tvar:

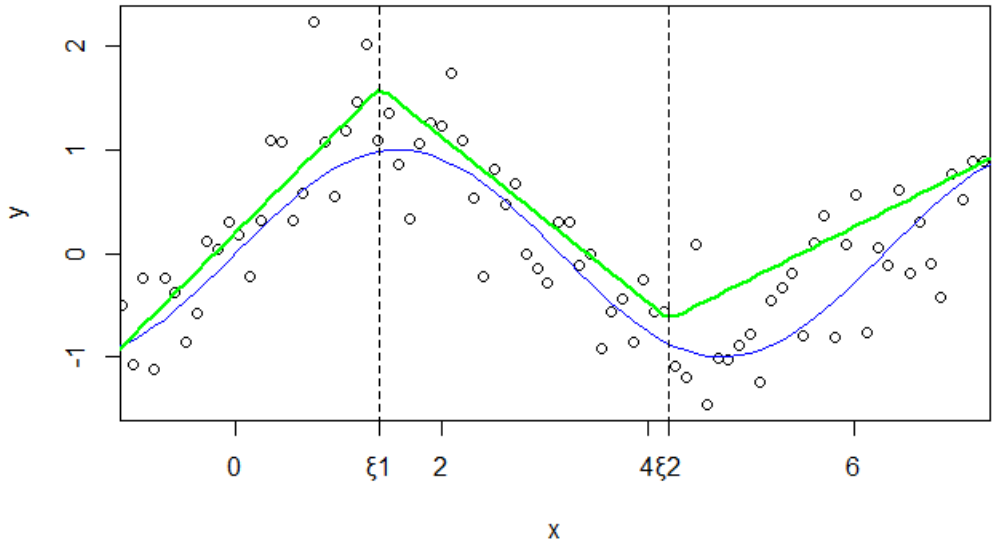
$$h_1(x) = I(x < \xi_1), \qquad h_2(x) = I(\xi_1 \leq x < \xi_2), \qquad h_3(x) = I(\xi_2 \leq x).$$

Vidíme, že metodou nejmenších čtverců, kterou si z důvodu následné lepší návaznosti podrobně popíšeme až v další kapitole, je pro model $f(x) = \sum_{m=1}^{3} \beta_m h_m(x)$ výsledkem odhadu parametrů aritmetický průměr $\hat{\beta}_m = \bar{y}_m$ hodnot y v každém podintervalu.



Obrázek 2: Po částech lineární nespojitá funkce (zeleně)

Když pokročíme dále k obrázku 2, vidíme, že daty je proložena po částech lineární funkce. To stejné je i na obrázku 3, s tím rozdílem, že je zde navíc podmínka, a to taková, že polynomy musí být v uzlech spojité. To znamená, že funkční hodnoty těchto polynomů se musí v bodě jejich spojení rovnat. Tato omezení vedou k omezením týkajícím se odhadů parametrů. Například pokud se u prvního uzlu požaduje, aby $f(\xi_1^-) = f(\xi_1^+)$, znamená to, že platí $\beta_1 + \xi_1\beta_4 = \beta_2 + \xi_1\beta_5$. Tím pádem se snižuje počet parametrů o jeden, tedy u dvou uzlů, které tu máme o dva. Pak z původních šesti volných parametrů nám zbývají



Obrázek 3: Po částech lineární spojitá funkce (zeleně)

pouze čtyři, o kterých uvažujeme jako ostupních volnosti.

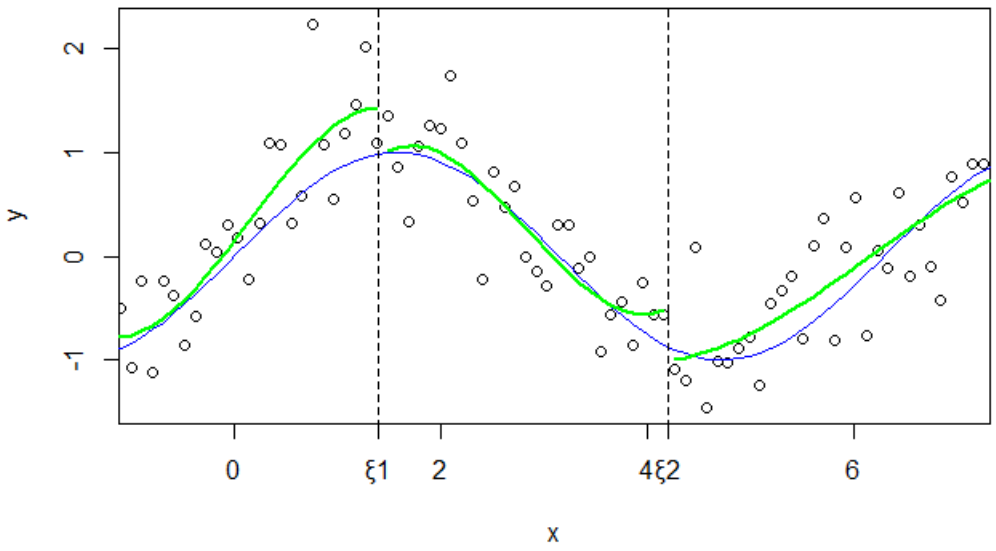
V praxi požadujeme, aby derivace až do řádu $M-1$ byly v uzlech také spojité. To můžeme vidět na kubickém splajnu (třetího stupně) na obrázku 5. Částečné polynomy, které jsou celkem tři, jsou kvadratické, tedy mají 4 koeficienty, což dává $3 \times 4 = 12$ koeficientů celkem. Nicméně od nich musíme ještě odečíst 6, protože v každém uzlu jsou 3 omezení (funkční hodnoty, první a druhá derivace jsou spojité) a máme dva uzly. Takže jsme dostali $12 - 3 \times 2 = 6$ stupňů volnosti.

Pravidlo je jednoduché: Celkový počet stupňů volnosti se rovná počtu konstant polynomu (řád polynomu) plus počet uzlů, tedy $M + 1 + k$.

Když zde nejsou uzly, splajn se znovu stává jednoduchým polynomem.

Jak si můžeme povšimnout, když porovnáme obrázky 1, 2, 3 a 4, 5, vidíme, že při zvyšování stupně se nám zlepšuje aproximace dat. Kdybychom ve skutečnosti zvýšili stupeň na čtyři nebo více, tak by to nevedlo k lepším výsledkům. Nejen z tohoto důvodu, ale i díky svým početním výhodám, jsou také kubické splajny nejpoužívanější.

Na obrázcích 4 a 5 vidíme složené kubické polynomy. Na prvním z nich je funkce v uzlech nespojitá, pak by následovala funkce, která by byla spojitá, po ní funkce se spojitou první derivací a funkci která má spojitou i druhou derivaci



Obrázek 4: Po částech kvadratická nespojitá funkce (zeleně)

vidíme na obrázku 5, kde už jde vlastně o kubický splajn. Poznamenejme, že oblíbeným typem kubických splajnů jsou přirozené splajny, využívající podmínky, aby byla funkce v krajních bodech intervalu lineární (tedy nulová druhá derivace). Toho dosáhneme za cenu čtyř stupňů volnosti (dvou na každém konci).

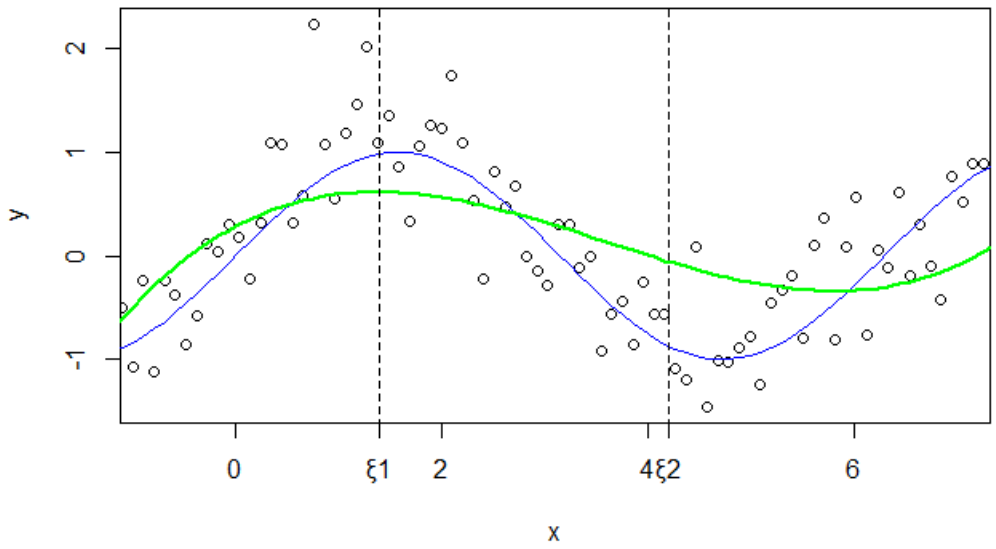
Ještě jednou si tedy shrneme, že splajn stupně M s uzly $\xi_i, i = 1, \ldots, k$ je složený polynom stupně M a má spojitou derivaci až do stupně $M - 1$. Bázové funkce splajnů jsou dány rovnicemi:

$$h_j(x) = x^{j-1}, \quad j = 1, \ldots, M+1,$$
$$h_{M+1+l}(x) = (x - \xi_l)_+^M, \quad l = 1, \ldots k.$$

Potom platí: Počet bázových funkcí = počet parametrů (=df, tj. stupňů volnosti).

U kubických splajnů ($M = 3$) máme tyto následující bázové funkce:

$$h_1(x) = 1, \qquad h_3(x) = x^2, \qquad h_5(x) = (x - \xi_1)_+^3,$$
$$h_2(x) = x, \qquad h_4(x) = x^3, \qquad h_6(x) = (x - \xi_2)_+^3.$$



Obrázek 5: Po částech kvadratická spojitá funkce (zeleně)

Předem musíme mít určeny tedy tyto parametry:

- stupeň splajnů M,
- počet uzlů,
- pozice uzlů: určené uživatelem, mohou být ekvidistantní či lépe např. na percentilech x-ových hodnot (pro $k = 3$ by se jednalo o dolní kvartil, medián a horní kvartil).

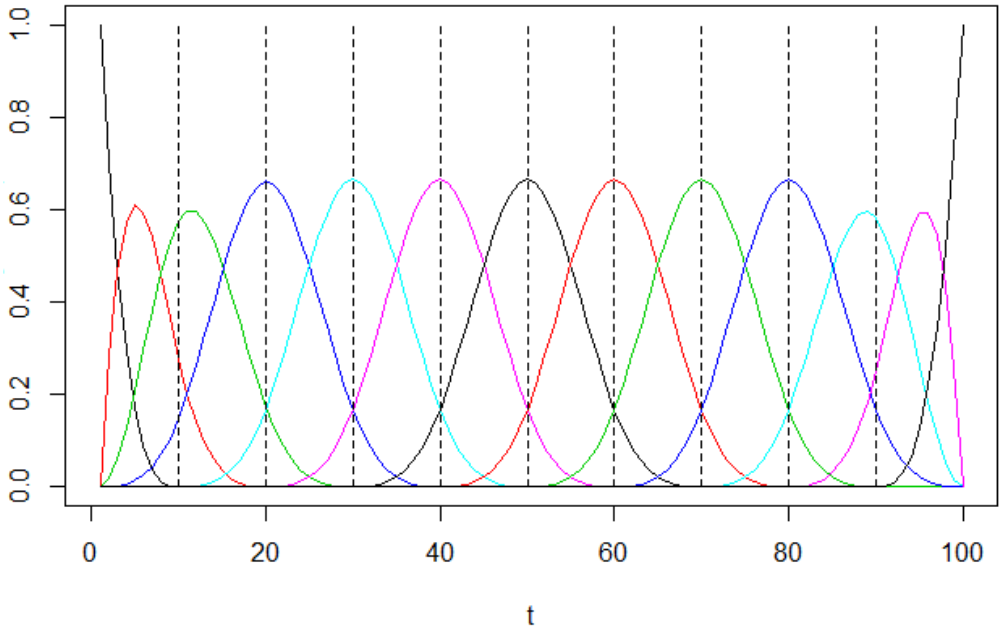
1.2 Bázové splajny (B-splajny)

Výše popsané splajnové báze nejsou v praxi tolik numericky užívané, narozdíl od B-splajnů, které jsou numericky vhodnější a poskytují ekvivalentní formu znázornění báze. Systém B-splajnů byl vyvinut americkým matematikem Carlem de Boorem (narozen 1937). Za tímto účelem určíme systém bázových funkcí $\phi_r(t), r = 1, ..., R$, který bude mít následující vlastnosti:

- každá bázová funkce $\phi_r(t)$ je sama o sobě splajnovou funkcí definovaná stupněm M a posloupností uzlů $\xi_i, i = 1, \ldots, k$;

- vzhledem k tomu, že násobek splajnové funkce je opět splajnová funkce a součet a rozdíl splajnů jsou také splajny, tak každá lineární kombinace těchto bázových funkcí je splajnová funkce;
- každá splajnová funkce definovaná stupněm M s uzly $\xi_i, i = 1, \ldots, k$ může být vyjádřena jako lineární kombinace těchto bázových funkcí.

Obrázek 6 ukazuje třináct B-splajnových funkcí třetího stupně definovaných devíti rovnoměrně rozloženými uzly, které jsou na obrázku znázorněny svislými přerušovanými čarami. Všimněme si, že každá ze sedmi bázových funkcí ve středu je kladná nad čtyřmi přiléhajícími podintervaly. Protože kubické splajny mají dvě spojité derivace, každá bázová funkce umožňuje hladký přechod do oblastí,kde jsou nulové. Tyto středové bázové splajny mají stejný tvar kvůli stejnému rozmístění uzlů; naopak nerovnoměrné rozmístění uzlů by nám vytvořilo splajny lišící se svým tvarem.



Obrázek 6: B-splajnové funkce třetího stupně

Tři levé bázové funkce a jejich tři pravé protějšky se liší tvarem, ale jsou nicméně opět kladné nejvýše nad čtyřmi přiléhajícími podintervaly. Je to vlast-

nost kompaktního nosiče, která spočívá v tom, že B-splajnová funkce řádu M je kladná nejvýše nad $M + 1$ sousedními podintervaly. Její význam spočívá v efektivním výpočtu. Protože když máme R B-splajnových funkcí, pak matice skalárního součinu těchto funkcí řádu R, která se za tímto účelem využívá, bude mít pouze M vedlejších úhlopříček umístěných nad a pod hlavní diagonálou, které obsahují nenulové hodnoty. To znamená, že pro výpočetní efektivitu nezáleží na tom, jak velké je R, čehož se v praxi hojně využívá.

Vidíme, že na obrázku 6 jsou tři bázové funkce nalevo a napravo rozdílné. Je to z toho důvodu, že když přecházíme od okrajů do středu, přechody splajnů do nulové oblasti mají vždy spojité druhé derivace, tedy jsou hladké. Na druhou stranu přechod ze středu do krajů se u tří krajních splajnů liší. Splajn nejvíce nalevo je nespojitý, další je jen spojitý a třetí má spojitou první derivaci. Je to logické, protože většinou nemáme přehled o tom, co se děje za hranicemi našich dat (např. včetně výskytu nespojitosti), takže to můžeme jen předpovídat.

Zápis $B_r(x, \xi)$ značí hodnotu B-splajnové funkce definované posloupností uzlů ξ v bodě x. Index r označuje největší uzel v x nebo bezprostředně vlevo od x. V souladu s tímto zápisem je splajnová funkce $S(x)$ s diskrétními vnitřními uzly definována jako:

$$S(x) = \sum_{r=1}^{M+1+k} c_r B_r(x, \xi).$$

Teď už si jen stačí určit, kam umístit uzly ξ. Jak již bylo zmíněno dříve, pokud máme rovnoměrně rozložená data, můžeme udržovat rovnost rozestupů. Pokud jsou data nerovnoměrná, je vhodné uzly umístit na j-tý datový bod, kde j je předem stanovené, tedy na kvantilech z rozdělení x-ových hodnot. Speciálním případem jsou hladké splajny, u kterých jsou uzly umístěné v každé hodnotě argumentu.

2 Regresní analýza

Jak už bylo řečeno v úvodu, regresní analýza je metoda, která studuje vzájemné závislosti mezi jednotlivými proměnnými. Jejím dalším cílem je predikce hodnot jedné závislé proměnné, za předpokladu, že již známe hodnoty několika jiných nezávislých proměnných. V dalším se zaměříme především na první z naznačených úkolů. V této části, pro kterou bylo čerpáno z [1], [6], [7], [12], si tak popíšeme a definujeme regresní model, odhad regresních koeficientů pomocí metody nejmenších čtverců a zmíníme též speciální logistickou regresi, která nám poslouží v nadcházející kapitole.

Regresní funkce je definována jako podmíněná střední hodnota určité náhodné veličiny vzhledem k různým hodnotám nezávisle proměnných, které jsou typicky nenáhodné. V nejjednodušším případě (přímková regrese) se jedná o jednu nezávisle proměnnou, která je lineární funkcí parametrů,

$$\mathsf{E}(Y|x) = \beta_0 + \beta_1 x.$$

Regresní model můžeme obecně pro dvojice pozorování nezávisle a závisle proměnné (x_i, y_i) uvažovat ve tvaru

$$Y_i = g(x_i, \boldsymbol{\beta}) + \epsilon_i, i = 1, 2, ..., n,$$

kde $g(x, \boldsymbol{\beta})$ je nějaká hladká funkce. Náhodné odchylky od ideální regresní závislosti souhrně označujeme ϵ_i a bereme je za rušivou složku, šum neboli chybu. Vektor $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ je vektor regresních parametrů (koeficientů).

Pro popis struktury dat neboli vysvětlení dat chceme odhadnout neznámé parametry $\boldsymbol{\beta}$, což se v případě lineární funkce parametrů provádí pomocí metody nejmenších čtverců. Výsledný odhad značíme následně $\hat{\boldsymbol{\beta}}$.

2.1 Lineární regresní modely

Lineární zjednodušení (tzn. vyjádření regresní funkce jako lineární funkce parametrů) je velmi oblíbené především při větším počtu vysvětlujících proměnných.

V tomto modelu se předpokládá součtový vliv všech činitelů a regresní funkce je tak v následujícím tvaru:

$$\mathsf{E}(Y|(x_1, x_2, ..., x_p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

nebo

$$Y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, i = 1, 2, ..., n,$$

kde β_0 je absolutní člen a $\beta_1, ..., \beta_p$ jsou tzv. dílčí regresní koeficienty. Souhrnně často hovoříme o vícenásobné regresi.

Model regresní přímky

$$\mathsf{E}(Y|x) = \beta_0 + \beta_1 x$$

nebo

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, ..., n,$$

maticově $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ pro

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

je pak speciálním případem pro jednu vysvětlující proměnnou. Přitom $\epsilon_i, i = 1, ..., n$ je posloupnost nekorelovaných náhodných veličin s nulovou střední hodnotou a konečným kladným rozptylem σ^2, tj. $\mathsf{E}(\epsilon_i) = 0$ a $var(\epsilon_i) = \sigma^2$. Za účelem statistické inference (intervaly spolehlivosti, testování hypotéz) pak navíc často předpokládáme, že veličiny ϵ_i mají stejné, a to normální rozdělení.

V tomto modelu lze linearitu chápat z hlediska neznámých koeficientů regresní funkce. Často se ale využívají i regresní modely, které jsou lineární z hlediska všech parametrů, ale nelineární z hlediska vysvětlujících proměnných. Nejznámější je kvadratický regresní model, který je v případě jedné vysvětlující proměnné ve tvaru:

$$\mathsf{E}(Y|x) = \beta_0 + \beta_1 x + \beta_2 x^2.$$

2.2 Metoda nejmenších čtverců

U přímkové regrese a analogicky následně i u vícenásobné regrese či obecné regresní funkce odhady $\hat{\beta}_0, \hat{\beta}_1$ neznámých parametrů β_0, β_1 určíme metodou nejmenších čtverců, kterou si definujeme.

Definice 2.1. Náhodné veličiny $\hat{\beta}_0, \hat{\beta}_1$, které pro dané $Y_1, \dots Y_n$ minimalizují výraz

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2,$$

nazýváme odhady parametrů β_0, β_1 určené metodou nejmenších čtverců.

Jinak řečeno, touto metodou požadujeme, aby součet čtverců odchylek pozorovaných hodnot Y_i a odhadnutých hodnot $\hat{\beta}_0 + \hat{\beta}_1 x_i$ byl minimální. S využitím diferenciálního počtu tedy provedeme parciální derivaci tohoto výrazu podle obou koeficientů a výsledky položíme rovny nule, tak dostaneme stacionární body. Vzniklé soustavě rovnic se často říká normální rovnice, které mají v maticovém zápise tvar

$$(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}.$$

Ty mají jediné řešení

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

které je současně bodem minima funkce $S(\beta_0, \beta_1)$.

Nevýhoda metody nejmenších čtverců spočívá v citlivosti na možná odlehlá pozorování ve vysvětlovaných i vysvětlujících proměnných.

Rezidua neboli chyby vyrovnání hodnot jsou dány jako rozdíl $e_i = Y_i - \hat{Y}_i$. Rezidua e_i lze považovat za odhady hodnot rušivé složky ϵ_i. Veličina

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

se nazývá reziduální součet čtverců. Pro odhady $\hat{\boldsymbol{\beta}}$ regresních koeficientů $\boldsymbol{\beta}$ platí, že minimalizují RSS.

Potíže při výpočtech a zhoršení kvality odhadu (kvůli velkému rozptylu odhadů koeficientů regresní funkce) mohou nastat v případě, že matice $\mathbf{X}'\mathbf{X}$ je regulární a tedy existuje $(\mathbf{X}'\mathbf{X})^{-1}$, ale je tzv. špatně podmíněná, tj. je na „pokraji“ singularity. To znamená, že sloupce $\mathbf{X}$ jsou „téměř“ lineárně závislé a mluvíme pak o problému kolinearity. Pro jeho eliminaci nám poslouží hřebenová regrese, při které se k diagonále matice $\mathbf{X}'\mathbf{X}$ přičte m-násobek jednotkové matice $\hat{\boldsymbol{\beta}} = ((\mathbf{X}'\mathbf{X}) + m\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$. S růstem konstanty m klesá kolinearita a rozptyl odhadnutých koeficientů se blíží nule; vhodnou volbou m se zabývá např. [13].

2.3 Logistická regrese

Doposud jsme předpokládali, že jak vysvětlovaná, tak vysvětlující veličina jsou spojité, ale setkáváme se i s proměnnými, které jsou diskrétní povahy. Níže uvedený model, představující nejjednodušší případ tzv. zobecněných lineárních modelů, nám pak bude inspirací pro úvahy v následující kapitole.

Uvažujeme tedy alternativní znak Y na n statistických jednotkách. Nezávislé veličiny $Y_1, \ldots, Y_n$ s alternativním rozdělením s parametry $p_i, i = 1, \ldots, n$ popisují hodnoty znaku Y. Platí

$$\mathsf{P}(Y_i = 0) = 1 - p_i, \quad \mathsf{P}(Y_i = 1) = p_i, \quad \mathsf{E}(Y_i) = p_i, \quad \mathrm{var}(Y_i) = p_i(1 - p_i).$$

Regresní závislost bude ve tvaru tzv. logistické funkce

$$p_i = \mathsf{E}(Y_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}, \quad x_i = 0, 1.$$

Odtud lze vyjádřit

$$1 - p_i = \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}},$$

díky čemuž můžeme z logistické funkce vytvořit funkci novou, která se nazývá šance a porovnává pravděpodobnosti realizace Y_i v jedničce a nule (nastoupení a nenastoupení jevu). Šance má tvar

$$\omega(x_i) = \frac{p_i}{(1 - p_i)} = e^{\beta_0 + \beta_1 x_i} = \frac{\mathsf{P}(Y_i = 1)}{\mathsf{P}(Y_i = 0)}.$$

Zlogaritmováním šance dostáváme opět novou funkci, která se nazývá logit a v další kapitole, která se bude týkat i zobecněných lineárních modelů, nám tato funkce bude představovat tzv. spojovací funkci. Spojovací funkce určuje vztah mezi vysvětlovanou proměnnou a regresní funkcí. Logit má tvar

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i.$$

Obecněji pro více vysvětlujících veličin se dá logit napsat jako

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p,$$

kde p značí analogicy jako dříve pravděpodobnost, že očekávaný jev nastane a může být popsán následovně

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}.$$

Poznamenejme, že v kontextu dalšího textu si p označíme jako $\mu(\mathbf{x})$.

3 Zobecněný aditivní model - GAM

V této kapitole si popíšeme vyhlazující splajny, které jsou základem pro odhad parametrů u GAM. Dále si popíšeme, jak se od zobecněného lineárního modelu dostaneme k zobecněnému aditivnímu modelu a naznačíme, jak funguje tzv. backfitting algoritmus pro odhad parametrů v modelu GAM. Pro sepsání této kapitoly jsem čerpala z [3], [5], [9].

3.1 Vyhlazující splajny

Vyhlazující splajny představují určitý ,,vyhlazovátor" jako užitečný nástroj, který nám shrnuje trend závislé proměnné Y jako funkce jedné nebo více nezávislých proměnných $x_1, \ldots, x_p$. Díky němu následně budeme schopni lépe analyzovat vztah mezi Y a vysvětlujícími proměnnými.

Nejdůležitější vlastností je jeho neparametrická podoba. Což znamená, že hladká funkce je také neparametrická, takže nemusí mít pevnou podobu. To nám umožní aproximaci se součtem funkcí a ne pouze jednou funkcí, což jak uvidíme dále, bude největší rozdíl od GLM.

Kubický vyhlazující splajn je taková funkce, která ze všech funkcí $f(x)$, které mají spojitou druhou derivaci, minimalizuje ,,penalizovaný" reziduální součet čtverců pro n-tici pozorování hodnot závislé a nezávislé proměnné (x_i, y_i)

$$PRSS(f, \lambda) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int [f''(t)]^2 dt, \tag{1}$$

kde λ je pevně zvolená konstanta. První člen nám představuje metodu nejmenších čtverců, takže pouze s touto částí bychom dostali ,,aproximační" křivku, která by nebyla vůbec hladká. Tento člen nám zaručuje co nejlepší přizpůsobení neboli těsnost k datům, zatímco druhý člen ,,penalizuje" zakřivení funkce. Integrál u tohoto členu měří křivost funkce, takže lineární funkce by měla hodnotu integrálu nulovou, zatímco nelineární funkce by měla nenulovou hodnotu.

Volba vyhlazujícího parametru λ, tak představuje kompromis mezi přesností aproximace dat a křivostí funkce. Máme zde dva krajní případy:

- $\lambda = 0$: penalizace se stává nedůležitou, takže dostaneme funkci, která prochází daty;
- $\lambda = \infty$: role penalizovaného výrazu je klíčová, takže výsledkem je lineární funkce, neboli přímka určená metodou nejmenších čtverců.

Větší hodnoty λ nám dávají vyhlazenější křivku, zatímco s menšími hodnotami dostáváme křivku křivější.

Jako řešení dostanu přirozený splajn, který můžeme popsat též takto,

$$f(x) = \sum_{j=1}^{R} N_j(x)\theta_j,$$

kde N_j jsou prvky R-dimenzionální množiny bázových funkcí, které reprezentují přirozené splajny. Vztah (1) ze začátku kapitoly tedy můžeme maticově napsat jako

$$PRSS(\theta, \lambda) = (\mathbf{y} - \mathbf{N}\theta)^{'}(\mathbf{y} - \mathbf{N}\theta) + \lambda\theta^{'}\mathbf{\Omega}_N\theta$$

s tím, že $\{\mathbf{N}\}_{ij} = N_j(x_i)$ a $\{\mathbf{\Omega}_N\}_{jk} = \int N_j^{"}(t)N_k^{"}(t)dt$. Výsledek získáme zobecněním hřebenové regrese

$$\hat{\boldsymbol{\theta}} = (\mathbf{N}^{'}\mathbf{N} + \lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^{'}\mathbf{y}.$$

Výsledný vyhlazující splajn pak získáme jako

$$\hat{f}(x) = \sum_{j=1}^{R} N_j(x)\hat{\theta}_j.$$

3.1.1 Výběr vyhlazujícího parametru

Zatím jsme si neřekli, jak se volí parametr λ ve vyhlazujících splajnech. Zde se využívá techniky křížové validace. Tato křížová validace je počítána softwarem, proto si zde jen přibližně vysvětlíme, jaký je základní princip.

Křížová validace spočívá v tom, že naše datová množina, kterou máme k dispozici, se rozdělí na několik podmnožin. Z nich se vybere jedna, která se nazývá

trénovací množinou a zbytek jsou testovací množiny. Pomocí trénovací množiny se ,,natrénuje“ model, potažmo jeho parametry, a pomocí testovací množiny testujeme přesnost a výkonnost tohoto modelu. Tento proces se několikrát opakuje, pokaždé s jiným výběrem trénovací a testovacích množin.

Křížová validace nám pak říká, jak dobře bude model s námi odhadnutým λ, které jsme natrénovali pomocí dat z trénovací množiny, pracovat na datech z testovací množiny.

3.2 Od GLM k GAM

Lineární regresní model uvažovaný z kapitoly 2.1. je vlastně též zobecněným lineární modelem (generalized linear model, GLM), kde předpokládáme, že vysvětlovaná proměnná Y má normální rozdělení a identickou spojovací funkci. Obecně GLM zahrnují i ostatní typy rozdělení z exponenciální rodiny rozdělení (exponenciální, gama, lineární, Poissonovo a další) a zahrnují spojovací funkci, která se vztahuje k příslušné střední hodnotě μ.

Jestliže uvažujeme vícenásobnou lineární regresi ve tvaru

$$\mathsf{E}(Y|x_1,\ldots,x_p) = \alpha + \sum_{j=1}^{p} \beta_j x_j,$$

pak zobecněný aditivní model nahrazuje součet vysvětlujících proměnných $\sum_{j=1}^{p} \beta_j x_j$ součtem nelineárních, ale hladkých funkcí $\sum_{j=1}^{p} f_j(x_j)$ a můžeme jej vyjádřit ve tvaru

$$\mathsf{E}(Y|x_1,\ldots,x_p) = \alpha + \sum_{j=1}^{p} f_j(x_j).$$

Jako další můžeme uvažovat logistickou regresi ve tvaru

$$\ln \frac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})} = \alpha + \sum_{j=1}^{p} \beta_j x_j,$$

kde $\mu(\mathbf{x}) = \mathsf{P}(y = 1|\mathbf{x})$. Zobecněním, což představuje opět nahrazení součtu vysvětlujících proměnných součtem hladkých funkcí, dostaneme takzvaný adi-

tivní logistický regresní model

$$\ln \frac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})} = \alpha + \sum_{j=1}^{p} f_j(x_j).$$

U zobecněného lineárního modelu, respektive u zobecněného aditivního modelu nahrazujeme levou stranu různými spojovacími funkcemi a pravá strana zůstává ve tvaru lineární kombinace vstupních vysvětlujících proměnných (u GLM), respektive nelineárních funkcí vstupních vysvětlujících proměnných (u GAM).

Obecně u GAM je $\mu(\mathbf{x})$ propojena s aditivní funkcí vysvětlujících proměnných díky spojovací funkci g:

$$g(\mu(\mathbf{x})) = \alpha + \sum_{j=1}^{p} f_j(x_j).$$

Příklady klasických spojovacích funkcí jsou:

- $g(\mu) = \mu$: u aditivních lineárních modelů s normální rozdělením se užívá identity, jak už jsme zmínili v úvodu této kapitoly,
- $g(\mu) = logit(\mu)$: logitová spojovací funkce se užívá u alternativního rozdělení,
- $g(\mu) = \log(\mu)$: log-lineárních nebo log-aditivních modelů se užívá u Poissonova rozdělení.

3.3 Odhad parametrů u GAM

Neparametrické funkce, kterými jsme nahrazovali součty vysvětlujících proměnných, můžeme odhadnout použitím kubického vyhlazujícího splajnu, pomocí iterační metody zvané backfitting algoritmus.

V následujícím textu budeme uvažovat aditivní lineární model v tomto tvaru po naměření hodnot vysvětlované a vysvětlujících proměnných,

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i, \quad i = 1, \ldots, n,$$

kde chyby měření ϵ_i jsou nezávislé na x_j a jejich střední hodnota je nulová. Dále f_j jsou libovolné funkce jedné proměnné. Pak pro pozorování, která máme dána jako uspořádaná dvojice (x_i, y_i), můžeme využít ,,penalizovaného" reziduálního součtu čtverců ($PRSS$) v tomto tvaru

$$PRSS(\alpha, f_1, \ldots, f_p) = \sum_{i=1}^{n} \left\{ y_i - \alpha - \sum_{j=1}^{p} f_j(x_{ij}) \right\}^2 + \sum_{j=1}^{p} \lambda_j \int f_j''(t_j)^2 dt_j.$$

Integrál $\int f_j''(t_j)^2 dt_j$ v posledním členu měří, o kolik je druhá derivace funkce větší než nula, tedy křivost dané funkce. Stejně, jako tomu bylo u vyhlazujících splajnů, tento integrál má pro lineární funkci nulovou hodnotu a nenulovou hodnotu pro nelineární funkci. $\lambda_j \geq 0$ jsou ladící parametry a čím větší je jejich hodnota, tím více se funkce blíží k linearitě. Vidíme, že aditivní model s kubickými vyhlazujícími splajny minimalizuje $PRSS$. Každá funkce f_j nám představuje kubický splajn pro danou proměnnou x_j s uzly v každé hodnotě x_{ij}, $i = 1, \ldots n$.

Jednoznačnost tohoto řešení lze dosáhnout zavedením dalšího omezení:

$$\sum_{i=1}^{n} f_j(x_{ij}) = 0 \quad \forall j \quad \Longrightarrow \quad \hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} y_i =: \overline{y_i}$$

a plnou sloupcovou hodností matice $\mathbf{X} = [(x_{ij})]$.

Iterační algoritmus pro odhad modelu GAM:

1. Zahájení $\hat{\alpha} = \overline{y_i}$, $\hat{f}_j \equiv 0 \quad \forall i, j$
2. Pro cyklus $j = 1, 2, \ldots, p, \ldots 1, 2, \ldots, p \ldots$

- $\hat{f}_j \leftarrow S_j \left\{ \left[y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ij}) \right] \right\}, i = 1, \ldots, n$
- $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{n} \sum_{i=1}^{n} \hat{f}_j(x_{ij})$,

pokračujeme do doby, dokud se odhady funkcí $\hat{f}_j$ nestabilizují. Funkce $S_j(x) = \sum_{j=1}^{R} N_j(x)\theta_j$ označují kubický vyhlazující splajn, který jsme měli v kapitole 3.1.

4 Příklady aplikací GAM na data v softwaru R

Praktická část je zpracována v softwaru R [11]. Tento software je volně šiřitelný, takže si ho může dovolit kdokoliv, kdo disponuje počítačem a navíc je uživatelsky přívětivý, což jsou jeho nesporné výhody oproti jiným statistickým softwarům.

V této části si ukážeme tři praktické příklady, z nichž na prvním si jednoduše ukážeme, jak GAM fungují. Jsou v něm použity data `bodyfat` převzatá z knihovny `mfp`. Další příklad je částěčně převzatý z [3]. Ten bude o něco komplexnější a týká se vědecké úrovně různých zemí, kde využijeme tato data [4] a knihovnu `mgcv`. V posledním příkladu budeme pracovat s daty `prostate` z knihovny `ElemStatLearn`, která se týkají rakoviny prostaty.

4.1 Příklad 1: Tělesné proporce mužů

Data, která použijeme v tomto příkladu, jsou z knihovny `mfp`, jak už jsem zmínila a obsahují odhady procent tělesného tuku a naměřené hodnoty obvodů různých částí těla 252 mužů. Nejdříve si otevřeme knihovnu a potom si načteme data pomocí těchto příkazů:

```
>library(mfp)
>data(bodyfat)
```

Tato data obsahují několik veličin, ale pro naše účely budeme využívat pouze některé z nich, a to konkrétně:

- `age` – věk v letech,
- `height` – výška uvedená v palcích (jeden palec je asi 2,54 centimerů),
- `weight` – hmotnost měřená v librách (jedna libra je asi 0,453 kilogramů),
- `brozek` – procento tělesného tuku uvedené dle [2].

Před další analýzou odstraníme odlehlá pozorování u veličin `height` a `weight` a vybereme si z původního datového souboru pouze čtyři potřebné veličiny. Teď, když máme definovaný nový datový soubor s názvem `nnbodyfat1`,

zadáme funkci pro použití GAM z knihovny `mgcv`, ale nejdříve vezmeme funkci pouze jedné vysvětlující proměnné, a to `height` (výška), zatímco naší vysvětlovanou proměnou zde bude `brozek` (procento tělesného tuku).

```
>gam1=gam(brozek~s(height),data=nnbodyfat1)
```

Funkce `s(.)`, která se používá ve specifickém modelu vzorce pro GAM, naznačuje, že veličina `height` má být vyrovnána pomocí vyhlazujícího splajnu. Když si pak zadáme do R pouze `gam1`, dostaneme základní informace o této funkci.

```
>gam1
Family: gaussian
Link Function: identity

Formula:
brozek~s(height)

Estimate degrees of freedom:
7.01 total = 8.01

GCV score:  6.655017
```

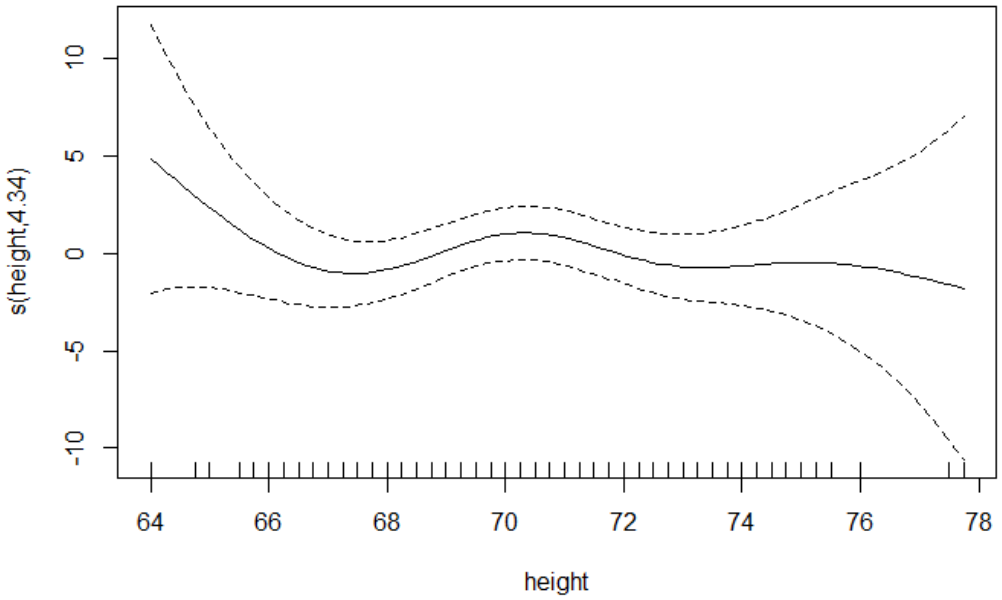
GCV, neboli skóre zobecněné křížové validace, můžeme brát jako odhad střední hodnoty chyby predikce na základě křížové validace. Samo o sobě nám toto číslo moc neříká, ale dá se srovnávat s jinými ukazateli z ostatních modelů, a pak je vhodnější ten s nižším skóre.

Stupně volnosti pro každou veličinu jsou nalezeny pomocí zobecněné křížové validace, pro tento příklad je použito 7,01 parametrů a celkový počet stupňů volnosti v modelu je pak obecně dán jako součet všech stupňů volnosti plus 1,

určený pro regresní konstantu α. V tomto příkladu 8,01.

Příkaz `plot()` pro funkce `gam` v R vytvoří následující graf zobrazený na obrázku 7, na kterém můžeme vidět čárkovaně vyznačený 95% interval spolehlivosti pro obdrženou splajnovou funkci.

```
>plot(gam1)
```



Obrázek 7: Regresní funkce při aplikaci zobeněné aditivní regrese pouze s jednou vysvětlující proměnnou

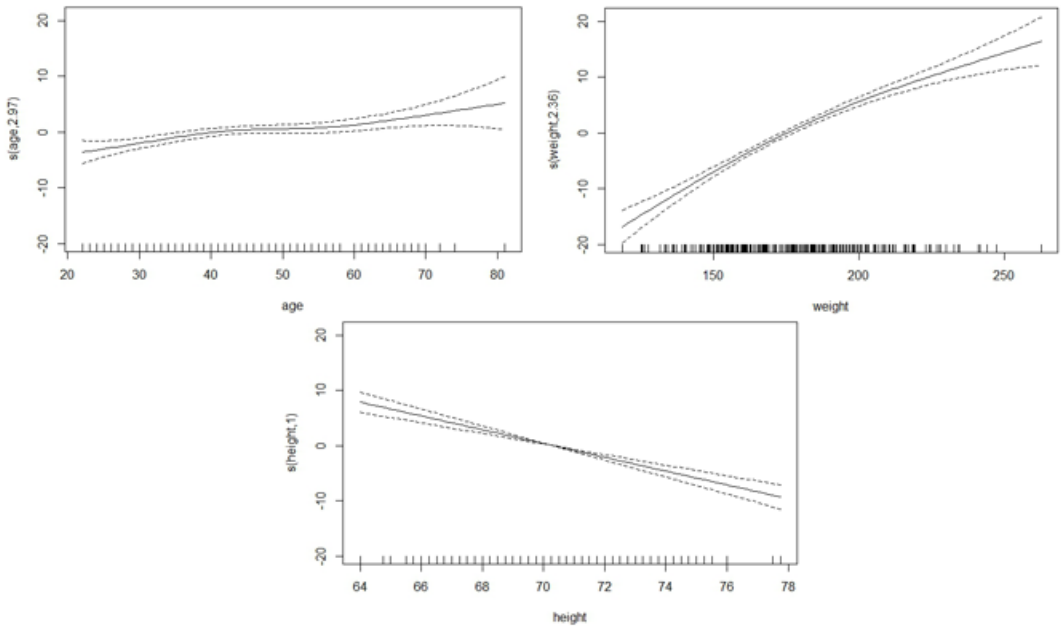
Nyní, když vezmeme více než jen jednu vysvětlující veličinu a vytvoříme novou funkci, kde vysvětlovaná veličina bude `brozek` a vysvětlující veličiny budou `age`, `weight` a `height`, bude vztah vypadat takto:

```
>gam2= gam(brozek~s(age)+s(weight)+s(height),data=nnbodyfat1)
>gam2


Family: gaussian
Link Function: identity
```

```
Formula:
brozek~s(age)+s(weight)+s(height)

Estimate degrees of freedom:
2.97 2.36 1 total = 7.33

GCV score:  24.42116
```

Z těchto informací opět můžeme vyčíst, jaká je hodnota stupňů volnosti a znovu si můžeme vykreslit grafy (obrázek 8), ale tentokrát dostaneme tři.

```
>plot(gam2,page=1)
```



Obrázek 8: Dílčí regresní funkce při aplikaci zobecněné aditivní regrese (normální rozdělení)

Ve výsledku vidíme, že procento tuku nezávisí na věku, zato roste s váhou a klesá s výškou.

Nyní změníme standardní nastavení *R*, co se týče pravděpodobnostního rozdělení a namísto normálního rozdělení použijeme gama rozdělení. Abychom to mohli

provést, musíme zajistit, že naše data budou kladná, proto odstraníme jednu nulovou hodnotu.

```
>telestuk = (nnbodyfat1[nnbodyfat1\$brozek>=0.1,])
```

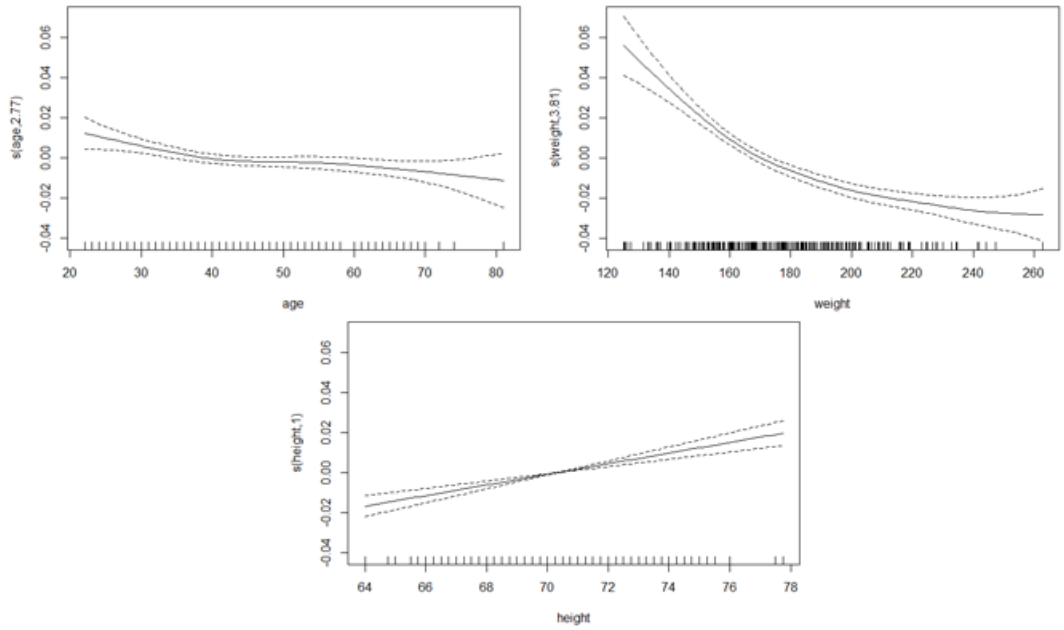
Teď pomocí daného příkazu změníme rozdělení:

```
>gam3= gam(brozek~s(age)+s(weight)+s(height),data=telestuk, +
+ family=Gamma)
>gam3

Family: Gamma
Link Function: inverse

Formula:
brozek~s(age)+s(weight)+s(height)

Estimate degrees of freedom:
2.77 3.81 1.00 total = 8.57

GCV score:  0.1072288
```

Vidíme, že změna rozdělení měla významný vliv na regresní závislost oproti situaci s normálním rozdělením. Podívejme se na příslušné grafy (obrázek 9) jednotlivých vysvětlujících funkcí.

```
>plot(gam3,page=1)
```

I z grafu je patrné, že volba rozdělení pravděpodobností závislé proměnné má také významný vliv na tvar dílčích regresních funkcí.



Obrázek 9: Dílčí regresní funkce při aplikaci zobecněné aditivní regrese (gama rozdělení)

4.2 Příklad 2: Vědecké skóre

Datový soubor v dalším příkladu se skládá z následujících veličin: průměrného vědeckého skóre podle jednotlivých zemí z programu pro mezinárodní hodnocení studentů (PISA), spolu s hrubým národním důchodem na obyvatele (HND), vzdělávacím indexem, indexem zdraví a indexem lidského rozvoje z databáze OSN. Načteme si potřebná data a otevřeme si knihovnu `mgcv`, která nám poskytuje funkce pro práci se zobecněnými aditivními modely:

```
>d = read.csv("http://www.nd.edu/~mclark19/learn/data/ +
+ pisasci2006.csv")
>library(mgcv)
```

Hlavní veličiny, které budeme potřebovat, jsou následující:

- `Overall` – vědecké skóre (průměrné skóre z výsledků vědeckého testu pro patnáctileté studenty),

- `Income` – veličina vyjadřující bohatství, která vychází z hrubého národního důchodu na obyvatele,
- `Edu` – představuje vzdělávací index, který se měří jako průměrný počet let strávených ve škole u dospělých ve věku 25 let a očekávaná délka školní docházky u děti v předškolním věku,
- `Health` – index zdraví.

Když už zhruba víme, jak s funkcí `gam` v Rku pracovat, aplikujeme je na tyto data s tím, že jako vysvětlovanou proměnnou zvolíme veličinu `Overall` a vysvětlující proměnné budou představovat tyto veličiny `Income`, `Edu` a `Health`. Nově teď použijeme příkaz `summary` a zjistíme, co můžeme vyčíst z obdrženého výstupu.

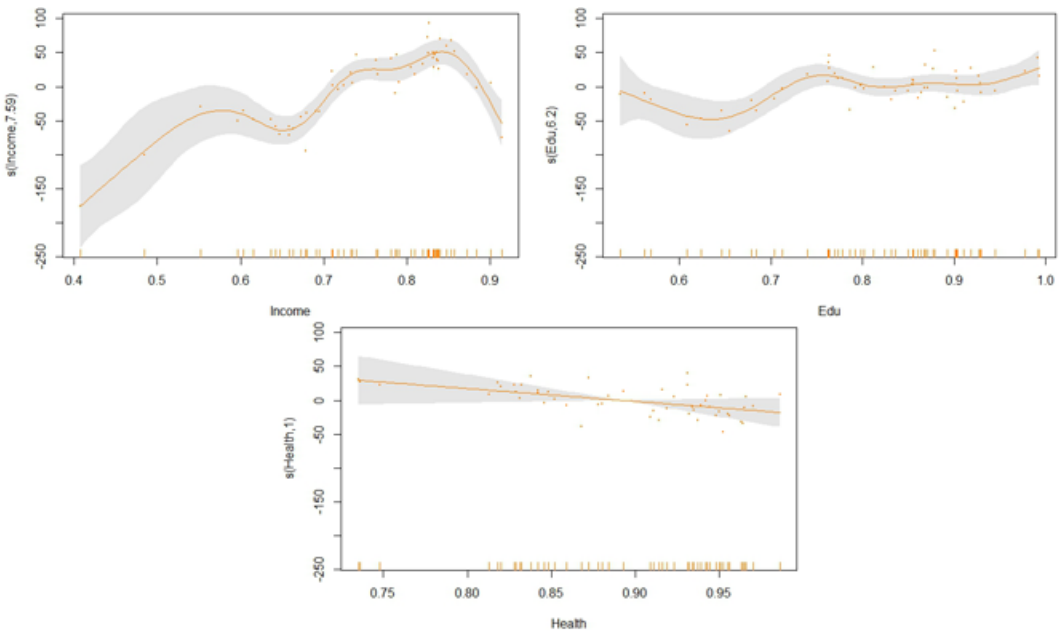
```
>mod_gam1 <- gam(Overall ~ s(Income)+s(Edu)+s(Health),data = d)
>summary(mod_gam1)
```

```
Family: gaussian
Link function: identity

Formula:
Overall ~ s(Income) + s(Edu) + s(Health)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  471.154      2.772     170   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
            edf Ref.df     F  p-value
s(Income) 7.593  8.415 8.826 1.33e-07 ***
s(Edu)    6.204  7.178 3.309  0.00733 **
s(Health) 1.000  1.000 2.736  0.10661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.863   Deviance explained = 90.3%
GCV score = 573.83  Scale est. = 399.5     n = 52
```

Jako první vidíme, že zde máme uvedené rozdělení pravděpodobností a spojovací funkci, které jsou v tomto případě normální a identita. Dále si povšimneme, že výstup je rozdělen na dvě části, z nichž první je parametrická, což zde představuje absolutní člen (`Intercept`), ale nemusí tomu tak vždy být. Druhá část se týká odhadnutých funkcí a zahrnuje veličiny `Income`, `Edu` a `Health` . Zdá se, že statisticky významný vliv má veličina `Income` a `Edu`, zatímco `Health` nemá. Hodnota efektivních stupňů volnosti (edf) u `Health` se rovná jedné, což nasvědčuje tomu, že má v podstatě tato proměnná jednoduchý lineární účinek. O tom se přesvědčíme i v nadcházejících grafech.

Nyní si vykreslíme grafy a podíváme se na vizuální efekty jednotlivých vlivů. Funkce jsou zobrazeny včetně 95% intervalů spolehlivosti (obrázek 10).

```
>plot(mod_gam1, pages=1, residuals=T, pch=19, cex=0.25, +
 + scheme=1, col='#FF8000', shade=T, shade.col='gray90')
```



Obrázek 10: Dílčí regresní funkce z aditivní regrese

Jsou tak vidět účinky jednotlivých vlivů a speciálně na třetím obrázku avi-

zovaný lineární efekt veličiny `Health`. Účinek `Income`, jehož vliv na vědcké skóre nejprve rapidně roste, se z jeho nejvyššího bodu potom postupně snižuje a `Edu` má celkově mírný pozitivní vliv na `Overall`. Vliv veličiny `Health`, jak jsem již zmínila, má lineární charakter, ale překvapivě mírně negativní.

4.3 Příklad 3: Rakovina prostaty

Poslední příklad se týká práce s daty, které mají zkoumat vztah mezi úrovní prostatického antigenu a několika klinickými měřeními u 97 mužů, kteří se chystají přijmout radikální prostatektomii. U tohoto příkladu použijeme jako vysvětlovanou proměnnou veličinu `lpsa`, která představuje zlogaritmovaný prostatický antigen a budu pozorovat vlivy ostatních veličin (použijeme celkem šest z přístupných dat). V roli vysvětlovaných proměnných budou veličiny:

- `lcavol` – zlogaritmovaný rozsah rakoviny,
- `lweight` – zlogaritmovaná váha prostaty,
- `age` – věk muže v letech,
- `lbph` – zlogaritmovaná velikost nezhoubného zvětšení prostaty,
- `lcp` – zlogaritmovaná kapsulární penetrace,
- `pgg45` – procento Gleasonova skóre stupně 4 nebo 5 (Gleasonovo skóre je systém používaný k hodnocení karcinomu prostaty na základě určitého nálezu).

Na začátku si opět otevřeme potřebné knihovny, a to `mgcv` a knihovnu `ElemStatLearn`, která obsahuje data `prostate`.

```
>library(mgcv)
>library(ElemStatLearn)
>data(prostate)
```

Teď už můžeme zadat správný tvar příkazu pro aplikaci GAM na data a opět zjistíme základní informace, pomocí příkazu `summary()`.

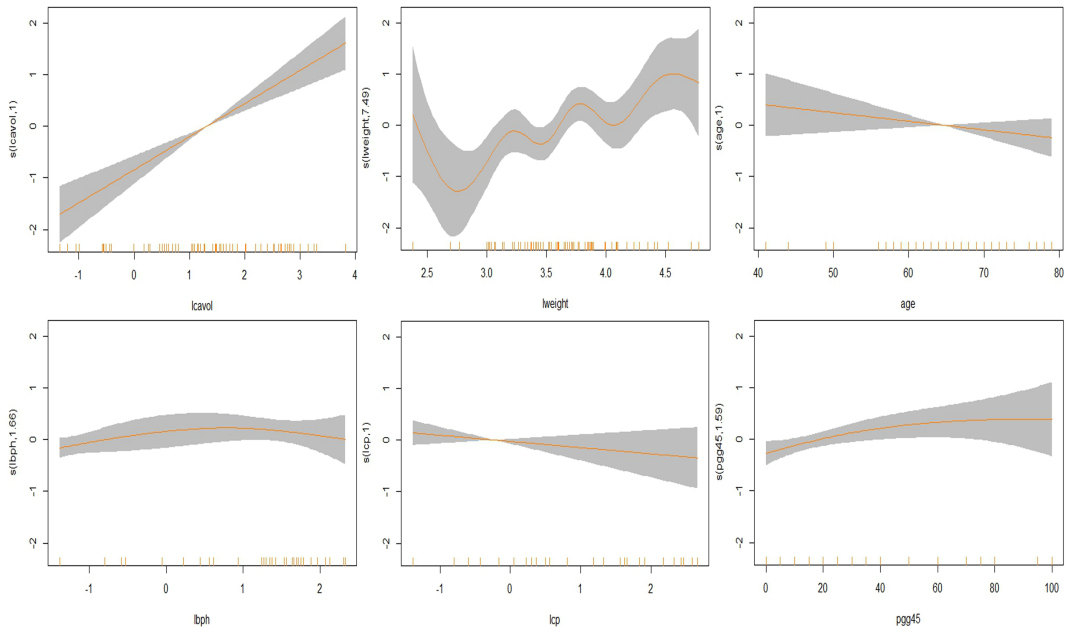
```
>mod_gam2<- gam(lpsa~s(lcavol)+s(lweight)+s(age)+s(lbph)+s(lcp)+
+ s(pgg45), data=prostate, subset=train)
>summary(mod_gam2)
```

Ze souhrnu je zřejmé, že statisticky významné jsou pouze dvě veličiny. První z nich je `lcavol`, tedy rozsah rakoviny, jehož vliv na prostatický antigen má lineární charakter, stejně jako u většiny zbývajících veličin, což odpovídá též intuitivní představě o situacích mezi proměnnými. Druhou veličinou je `lweight`, neboli váha prostaty, u které je spočtena hodnota 7,491 efektivních stupňů volnosti.

```
Family: gaussian
Link function: identity

Formula:
lpsa ~ s(lcavol) + s(lweight) + s(age) + s(lbph) + s(lcp) + s(pgg45)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.45235    0.08149   30.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
             edf Ref.df      F  p-value
s(lcavol)  1.000  1.000 38.961 4.33e-08 ***
s(lweight) 7.491  8.418  3.102  0.00531 **
s(age)     1.000  1.000  1.667  0.20231
s(lbph)    1.665  2.037  1.924  0.15489
s(lcp)     1.000  1.000  1.363  0.24820
s(pgg45)   1.589  1.952  3.151  0.05211 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.695   Deviance explained = 75.9%
GCV score = 0.57048  Scale est. = 0.44494   n = 67
```

Ještě si vykreslíme grafy pomocí příkazu:

```
>plot(mod_gam2, page=1, shade=TRUE, col='#FF8000', shade.col= +
+ 'grey')
```



Obrázek 11: Dílčí regresní funkce z aditivní regrese

Na obrázku 11 vidíme, že grafy odpovídají naznačenému shrnutí a u většiny z nich má jejich vliv téměř lineární charakter, ať už pozitivní nebo negativní. Jen veličina `lweight` je charakterizována větší křivostí a proměnlivostí. Zajímavá je zde role věku, který naznačuje spíše slabý negativní efekt na proměnnou `lpsa`, což by si jistě zasloužilo podrobnější odborné pozorování.

Závěr

Cílem mé bakalářské práce bylo seznámit čtenáře s nelineární regresí, konkrétně se zobecněným aditivním modelem. Vzhledem k tomu, že touto problematikou se česká literatura zabývá jen okrajově, může tato práce sloužit k širšímu seznámení s tímto modelem a následně pomoci při jeho dalším studiu, přestože jsem se vzhledem k rozsáhlosti problematiky musela dopustit četných zjednodušení při popisu modelu.

Toto téma jsem si vybrala také kvůli tomu, že mě zajímalo učivo regresní analýzy v kurzu matematické statistiky a tak jsem si chtěla rozšířit obzor v této problematice a jít více do hloubky. Největším oříškem pro mě bylo pochopit danou teorii, a to navíc z cizojazyčných zdrojů. Nicméně to byla výzva a díky pomoci mých přátel jsem zvládla jak příklady, tak snad i hlubší porozumění popisované problematiky.

Musela jsem se také naučit pracovat v softwaru R, což mě ale už od začátku bavilo. Hodně jsem s programem pracovala a zkoušela jeho různé funkce a možnosti. Myslím si, že mě to i dost obohatilo a beru jako velké plus, mít tyto dovednosti. K psaní mé práce jsem zvolila typografický systém LaTeX, protože je vhodnější pro sázení prací s matematickými vzorci než jiné systémy.

Doufám také, že má práce poslouží ostatním studentům, jako dobrý zdroj informací o zobecněných aditivních modelech a regresní analýze vůbec.

Literatura

[1] Anděl, J., *Matematická statistika*, 1.vydání, Praha: Nakladatelství technické literatury, 1978

[2] Brozek, J., Grande, F., Anderson, J. T., Keys, A., *Densitometric analysis of body composition: Revision of some quantitative assumptions*, Annals of the New York Academy of Sciences, 1963

[3] Clark, M., *Generalized Additive Models* [online], dostupné z: http://www3.nd.edu/ mclark19/learn/GAMS.pdf [citováno 5. 1. 2014]

[4] Data [online], dostupné z: http://www.nd.edu/ mclark19/learn/data/pisasci2006.csv [citováno 5. 1. 2015]

[5] Filzmoser, J., *Klassifikation und Diskriminanzanalyze*, Wien: Technische Universität Wien, 2009

[6] Hebák, P., *Regrese I.část*, 1.vydání, Praha: Vysoká škola ekonomická v Praze, 1998

[7] Hron, K., Kunderová, P., *Zálady počtu pravděpodobnosti a metod matematické statistiky*, 1.vydání, Olomouc: Univerzita Palackého v Olomouci, 2013

[8] Kobza, J., *Splajny*, 1.vydání, Olomouc: Univerzita Palackého v Olomouci, 1993

[9] Liu, H., *Generalized Additive Model* [online], dostupné z: http://www.d.umn.edu/math/Technical%20Reports/Technical%20Reports%202007-/TR%202007-2008/TR_2008_8.pdf [citováno 5. 1. 2014]

[10] Ramsay, J. O., Silverman, B.W., *Functional Data Analysis*, 2.vydání, New York: Springer Science+Business Media, Inc., 2005

[11] The R Project for Statistical Computing [online], dostupné z: http://www.rproject.org/ [citováno 5. 1. 2015]

[12] Víšek, J. A., *Statistická analýza dat*, 1.vydání, Praha: ČVUT, 1998,

[13] Zvára, K., *Regrese*, Praha: MATFYZPRESS, 2008