Úloha VI.S ... nelineární

10 bodů; průměr 8,06; řešilo 16 studentů

a) *Zkuste vlastními slovy popsat, k čemu a jak se používá nelineární regrese (postačí vlastními slovy popsat následující: model nelineární regrese, způsob odhadu regresních koeficientů, vyjádření nejistot odhadů regresních koeficientů a hodnot prokládané funkce, statistické testy hodnot regresních koeficientů, identifikovatelnost parametrů a způsob volby prokládané funkce). Není potřeba uvádět přesná matematická odvození, stačí požadované pojmy a vlastnosti stručně popsat.*

b) *V přiloženém datovém souboru* regrese1.csv *naleznete dvojice hodnot* (x_i, y_i)*. Těmito daty chceme proložit teoretickou funkční závislost, kterou je v tomto případě sinusoida, tedy funkce tvaru*

$$f(x) = a + b \cdot \sin(cx + d)\,.$$

Vykreslete graf naměřených hodnot a proložené funkce a stručně ho okomentujte (takovýto graf musí mít všechny náležitosti). Není potřeba dělat regresní diagnostiku.
Nápověda: *Dejte si pozor na identifikovatelnost parametrů v tomto modelu a vhodné omezující podmínky na parametr c.*

c) *V přiloženém datovém souboru* regrese2.csv *naleznete dvojice hodnot* (x_i, y_i)*. Těmito daty chceme proložit teoretickou funkční závislost, kterou je v tomto případě exponenciála, tedy funkce tvaru*

$$f(x) = a + \mathrm{e}^{bx+c}\,.$$

Určete hodnoty odhadů všech regresních koeficientů včetně nejistot měření.
Nápověda: *Grafickou metodou ověřte předpoklad homoskedasticity a v případě potřeby pro určení nejistot měření regresních koeficientů použijte Whiteův (sendvičový) odhad kovariančí matice.*

d) *V přiloženém datovém souboru* regrese3.csv *naleznete dvojice hodnot* (x_i, y_i)*. Těmito daty chceme proložit teoretickou funkční závislost, kterou je v tomto případě hyperbola, tedy funkce tvaru*

$$f(x) = a + \frac{1}{bx + c}\,.$$

Vykreslete graf naměřených dat v podobě průměrů a chybových úseček a proložené funkce a stručně ho okomentujte (takovýto graf musí mít všechny náležitosti). Proveďte regresní diagnostiku.

Bonus: *V přiloženém datovém souboru* regrese4.csv *naleznete dvojice hodnot* (x_i, y_i)*. Těmito daty chceme proložit teoretickou závislost, která je ovšem příliš složitá na analytické vyjádření. Proložte těmito daty regresní spliny (s vhodně zvolenými uzly a vhodně zvoleným stupněm).*

Pro práci s daty použijte výpočetní prostředí R*. Pro vyřešení těchto úkolů postačí drobně upravit přiložený skript, ve kterém je pomocí komentářů v kódu vysvětlena potřebná syntaxe jazyka* R*.* *Michal chtěl udělat poslední sérii co možná nejtěžší.*

a) Detailní odpověď na tuto otázku dostaneme jen přečtením pátého dílu seriálu. Na tomto místě popíšeme jen opravdu nejzákladnější věci.
Nelineární regrese se použije v případech, kdy chceme naměřenými daty prokládat funkci, která není lineární v neznámých regresních koeficientech. Matematický model nelineární regrese je tedy takový, že uvažujeme měřená data ve tvaru

$$y_i = f(x_i, \beta_0, \beta_1, \ldots, \beta_k) + \varepsilon_i\,,$$

kde y_i jsou naměřené hodnoty závisle proměnné, x_i jsou hodnoty nezávisle proměnné a ε_i je náhodná nepřesnost měření, o které předpokládáme, že má rozdělení $N(0, \sigma^2)$.
Na odhad regresních koeficientů z naměřených dat používáme metodu maximální věrohodnosti, která je v tomto modelu ekvivalentní metodě nejmenších čtverců. Odhady regresních koeficientů tedy volíme tak, aby celkový součet čtverců

$$\sum_{i=1}^{n} (y_i - f(x_i, \beta_0, \ldots, \beta_k))^2$$

byl co možná nejmenší (odtud název metoda nejmenších čtverců). V případě nelineární regrese už neexistují explicitní vzorce na výpočet hodnot odhadnutých koeficientů, tyto odhady musíme hledat za pomoci numerických metod a matematického softwaru.
Podobně jako v lineární regresi můžeme konstruovat intervaly spolehlivosti pro hodnoty neznámých regresních koeficientů i pro hodnoty prokládané funkce. V případě nelineární regrese už jsou ale vzorce na vyjádření těchto nejistot značně komplikované, proto jsme je v textu seriálu neuváděli. Stačilo nám, že nám tyto nejistoty poskytne matematický software na výstupu. Zmíněné intervaly spolehlivosti mají pro dostatečný počet měření (alespoň 4-5 krát více měření než neznámých regresních koeficientů) následující vlastnosti

$$P\left(\beta_i \in \left(\widehat{\beta}_i \pm u_{1-\frac{\alpha}{2}} s_n^{K_i}\right)\right) \doteq 1 - \alpha\,,$$
$$P\left(f(x, \beta_0, \ldots, \beta_k) \in \left(f(x, \widehat{\beta}_0, \ldots, \widehat{\beta}_k) \pm u_{1-\frac{\alpha}{2}} s_n^f(x)\right)\right) \doteq 1 - \alpha\,,$$

kde $s_n^{K_i}$ je nejistota měření regresního koeficientu a $s_n^f(x)$ je nejistota měření funkční hodnoty prokládané funkce v bodě x.
V případě lineární i nelineární regrese můžeme provádět statistické testy o hodnotách regresních koeficientů. Můžeme tedy testovat hypotézy a alternativy typu

$$H : \beta_j = \vartheta\,,$$
$$A : \beta_j \neq \vartheta\,,$$

kde ϑ je nějaká předem zvolená konstanta. Jako testová statistika se v tomto případě použije následující transformace naměřených dat

$$T = \frac{\widehat{\beta}_j - \vartheta}{s_n^{K_j}}\,,$$

kde $s_n^{K_j}$ je nejistota měření regresního koeficientu. Kritický obor takovéhoto testu má potom tvar

$$C = (-\infty, u_{\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}, \infty).$$

Jednoduchou úpravou by se dal tento test upravit k testování jednostranných modifikací zmíněné hypotézy a alternativy.
Stejně jako v lineární regresi je nutné volit za prokládané funkce jen takové funkce, které mají nějaké fyzikální opodstatnění. V opačném případě se vystavujeme velkému riziku špatné volby prokládané funkce. Všechny závěry by potom byly chybné.

V nelineární regresi navíc ještě vyvstává podmínka identifikovatelnosti parametrů našeho modelu. Jednoduše řečeno musí platit, že pro různé volby regresních koeficientů musí být prokládaná funkce jiná. V opačném případě bychom totiž nemohli jednoznačně určit odhady hodnot regresních koeficientů. Pokud máme model s neidentifikovatelnými parametry, tohoto problému se dá vždy zbavit přeparametrizováním pomocí menšího počtu parametrů.

b) Naměřenými daty chceme prokládat teoretickou funkci tvaru

$$f(x) = a + b \cdot \sin(cx + d)\,.$$

V tomto momentě si musíme uvědomit, že je to přesně případ, který jsme popisovali v textu seriálu jako problémový z hlediska omezení na hodnoty parametrů a identifikovatelnost parametrů.

Když si vykreslíme jednoduchý graf naměřených hodnot bez prokládané funkce, vidíme, že měřená data mají přibližně tvar sinusoidy s periodou řádově π. Pokud bychom nijak neomezili možnou hodnotu parametru c, který ovlivňuje periodu sinusoidy, velmi pravděpodobně bychom dostali na výstupu z matematického softwaru velmi velkou hodnotu tohoto koeficientu (neboť pro velkou hodnotu periody by funkce sinus prošla naměřenými daty velmi přesně). Jelikož ale z grafu vidíme, že perioda měřených dat bude poměrně nízká, přidáme ještě omezení na hodnotu parametru c. Přidáme si podmínku, že odhad parametru c musí ležet v intervalu

$$(0, 2)\,,$$

což odpovídá periodě sinusoidy větší než $\frac{\pi}{2}$ (toto je z obrázku jistě splněno). Dále si přidáme omezení na možné hodnoty parametru d. Budeme chtít, aby odhad tohoto parametru ležel v intervalu

$$(-\pi, \pi)\,.$$

Toto děláme z důvodu identifikovatelnosti parametrů v našem modelu.

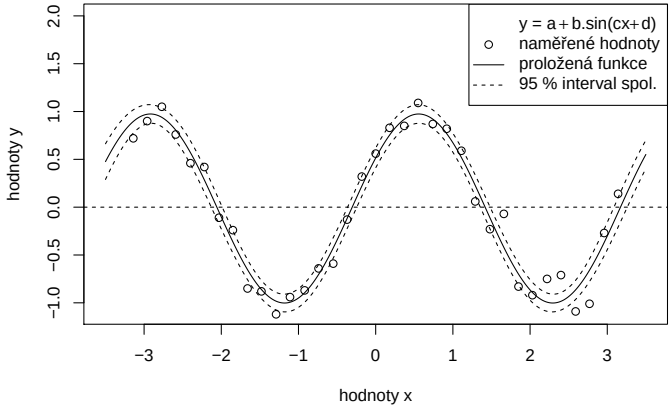
S těmito dvěma omezeními jsou už všechny parametry v našem modelu identifikovatelné a můžeme nechat matematický software spočítat jejich odhady. Po dosazení našeho modelu včetně těchto dvou požadavků do výpočetního prostředí R dostáváme na výstupu odhady regresních koeficientů, pomocí kterých můžeme vykreslit graf prokládané funkce. Takovýto graf společně s intervalovým odhadem pro prokládanou funkci můžeme vidět na Obr. 1. Z tohoto grafu je vidět, že proložená funkce sedí na naměřená data velmi dobře, takže nemusíme přidávat žádné další omezující podmínky a můžeme takovýto model prohlásit za finální. Zároveň je vidět, že jsme prokládanou funkci určili poměrně přesně, jelikož interval spolehlivosti je poměrně úzký.

c) Naměřenými daty budeme chtít prokládat teoretickou funkci tvaru

$$f(x) = a + \mathrm{e}^{bx+c}\,.$$

Všechny regresní koeficienty v tomto modelu jsou identifikovatelné a není žádný problém s omezeností některého koeficientu. Vše je tedy připraveno na vložení do výpočetního prostředí R.

Matematický software nám na výstupu poskytne hodnoty odhadů regresních koeficientů, na základě kterých můžeme vykreslit graf naměřených hodnot a proložené funkce, který můžeme vidět na Obr. 2. Z tohoto grafu dostáváme velké podezření na porušení předpokladu homoskedasticity. Toto ještě pro jistotu ověříme na grafu residuí oproti hodnotám nezávisle proměnné, který můžeme vidět na Obr. 3. Z těchto dvou grafů je nám jasné, že předpoklad



Obr. 1: Graf naměřených hodnot a proložené funkce z příkladu *b)*.

homoskedasticity byl porušen a že pro určení nejistot regresních koeficientů budeme muset použít Whiteův (sendvičový) odhad kovarianční matice.
Matematický software *R* poskytl následující odhady regresních koeficientů a příslušných nejistot (vypočítaných na základě Whiteova odhadu kovarianční matice)

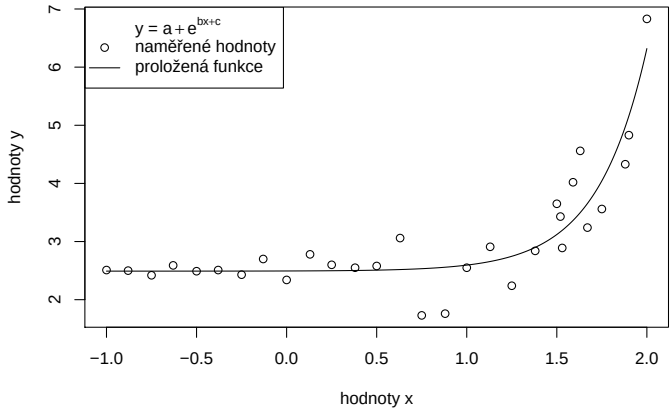
$$a = (2{,}49 \pm 0{,}08)\,,$$
$$b = (3.6 \pm 0{,}7)\,,$$
$$c = (-5{,}9 \pm 1{,}4)\,.$$

Na závěr tohoto příkladu můžeme konstatovat, že nejistoty měření regresních koeficientů spočítané klasickým postupem (které v případě heteroskedasticity nejsou správné) se od nejistot měření regresních koeficientů spočítaných pomocí Whiteova odhadu kovarianční matice liší maximálně o 45 %.

d) Naměřenými daty budeme prokládat funkci

$$f(x) = a + \frac{1}{bx + c}\,.$$

Je vidět, že všechny parametry jsou identifikovatelné a že nemůže nastat problém s omezeností jednotlivých parametrů, jelikož prokládaná funkce není periodická. Můžeme tedy rovnou pomocí nelineární regrese proložit tuto funkci naměřenými daty. Graf proložené funkce a naměřených dat ve formě chybových úseček můžeme vidět na Obr. 4. Je vidět, že proložená funkce dobře sedí na naměřená data. Zároveň nejsou vidět žádné známky heteroskedasticity. Můžeme se tedy domnívat, že byly všechny předpokladu nelineárního regresního modelu splněny. Toto ovšem ještě ověříme podrobnější regresní diagnostikou.



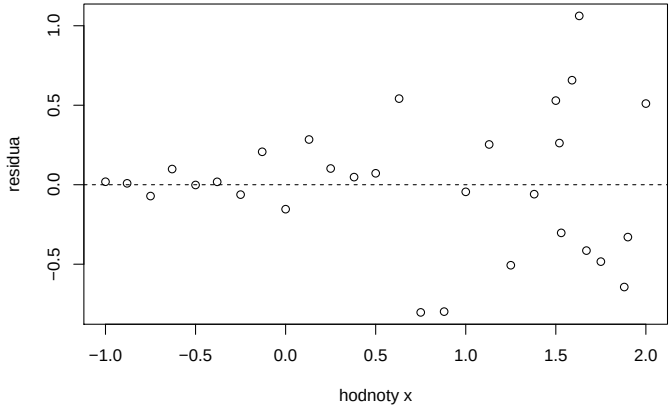
Obr. 2: Graf naměřených hodnot a proložené funkce z příkladu *c)*.

Na Obr. 5, resp. 6 můžeme vidět graf residuí oproti hodnotám nezávisle proměnné, resp. graf residuí oproti hodnotám proložené funkce. Na obou grafech vidíme náhodný shluk bodů kolem osy x bez žádné tendence shlukování nad nebo pod osou x ani různého rozptylu residuí v různých částech grafu. Můžeme tedy usoudit, že jsme prokládanou funkci volili správně a že naměřená data splňují předpoklad homoskedasticity. Dále můžeme na Obr. 7 vidět graf residuí oproti posunutým residuím. Na tomto grafu vidíme jen náhodný shluk bodů kolem počátku souřadnic bez známek shlukování v jednotlivých kvadrantech. Můžeme tedy usoudit, že i předpoklad nezávislosti měření byl pravděpodobně splněn.

Na závěr tedy můžeme říci, že regresní diagnostika neodhalila žádné závažné porušení předpokladů nelineárního regresního modelu. Toto proložení tedy můžeme považovat za splněné a konstatovat, že aplikace nelineárního regresního modelu na tato data byla v pořádku.

Bonus: Naším úkolem bude proložit zadanými daty regresní spliny. Než začneme zadávat nějaké příkazy do matematického softwaru, musíme si uvědomit, jakým omezením čelíme. Předně si musíme uvědomit, že náš datový soubor má celkem 89 měření. Z toho vyplývá omezení na počet regresních koeficientů, který by neměl překročit 15, aby na jeden regresní koeficient připadalo alespoň 5 měření. V ideálním případě bychom se měli snažit volit co nejmenší počet regresních koeficientů, ale zároveň musí být proložená funkce dostatečně přesná. Dalším omezením je stupeň prokládaných polynomů. Musíme si uvědomit, že vysoký stupeň prokládaných polynomů zvyšuje počet potřebných regresních parametrů. Musíme tedy opět vyvážit přesnost prokládané funkce a požadavek na co nejnižší stupeň prokládaných polynomů. Podle teorie vyložené v textu seriálu začneme se stupněm 3 a případně budeme dále, pokud to situace bude vyžadovat, stupeň zvyšovat.

Nyní už můžeme začít s prokládáním samotné funkce. Budeme postupovat přesně podle návodu v textu seriálu. Nejprve tedy jako uzly našich regresních splinů zvolíme ty body, ve kterých se nejvíce mění chování měřených dat, následně budeme tuto volbu upravovat (případně



Obr. 3: Graf residuí oproti hodnotám nezávisle proměnné z příkladu *c)*.

přidávat více uzlů), abychom docílili správné podoby proložené funkce.

Tímto postupem jsme po několika iteracích dospěli k modelu, kde prokládáme polynomy třetího stupně a uzly máme zvolené v bodech

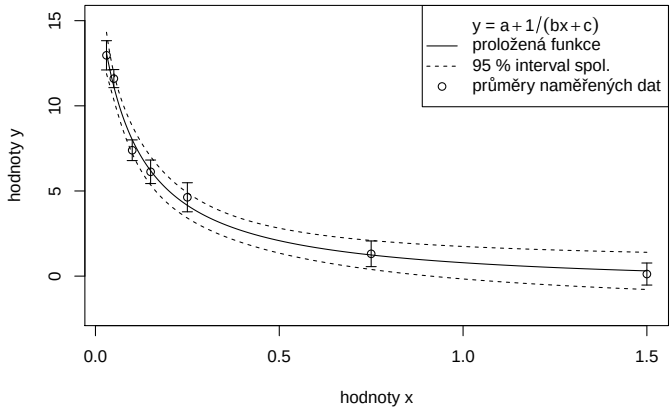
$$1; 1{,}8; 2{,}2; 3; 4{,}5; 5; 6; 6.5\,.$$

Na Obr. 8 můžeme vidět naměřená data společně s proloženou funkcí. Je vidět, že takto zvolená funkce velice dobře aproximuje naměřená data, můžeme tedy takovýto model prohlásit za finální.

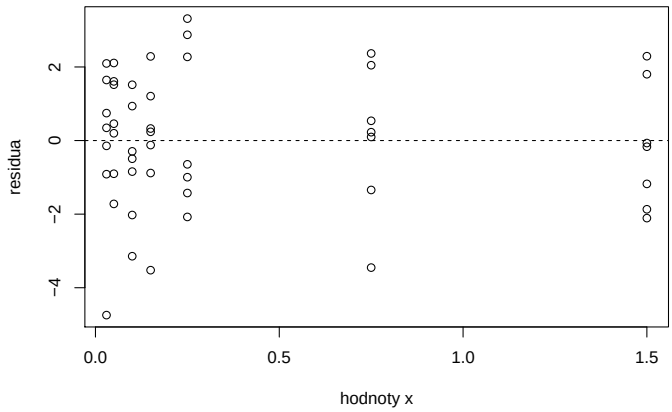
Na závěr jen poznamenejme, že výše popsaná volba uzlů rozhodně není jediná správná. Kdybychom některý z uzlů o trochu posunuli, tvar prokládané funkce by se téměř nezměnil a výsledný model by byl stále správný. Je navíc možné, že podobný graf bychom dostali i pro zcela jinou volbu uzlů.

Michal Nožička
`nozicka@fykos.cz`

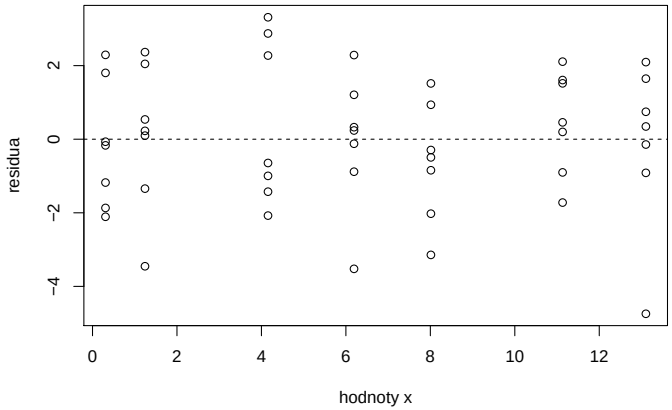
Fyzikální korespondenční seminář je organizován studenty MFF UK. Je zastřešen Oddělením pro vnější vztahy a propagaci MFF UK a podporován Ústavem teoretické fyziky MFF UK, jeho zaměstnanci a Jednotou českých matematiků a fyziků.



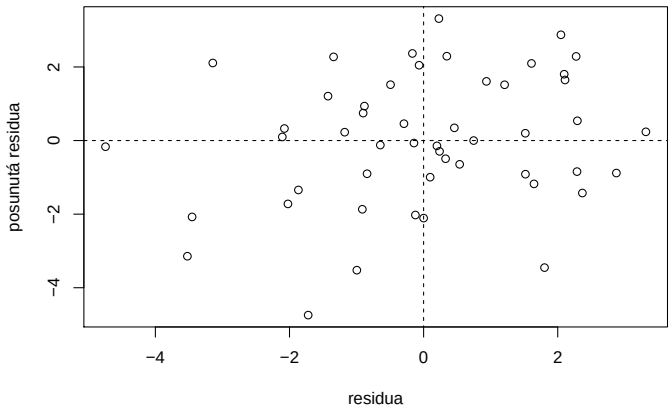
Obr. 4: Graf naměřených hodnot a proložené funkce z příkladu *d)*.



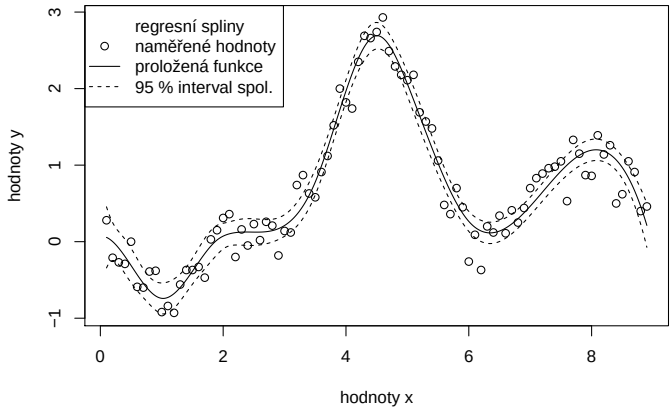
Obr. 5: Graf residuí oproti hodnotám nezávisle proměnné z příkladu *d)*.



Obr. 6: Graf residuí oproti hodnotám proložené funkce z příkladu *d)*.



Obr. 7: Graf residuí oproti posunutým residuím z příkladu *d)*.



Obr. 8: Graf proložených regresních splinů naměřenými daty z bonusového příkladu.