Kubický spline

Menu:	QCExpert	Regrese	Kubický spline

Modul Kubický spline slouží proložení prakticky libovolných regresních křivek naměřenými daty s jednorozměrnou nezávisle proměnnou *x* a jednorozměrnou náhodnou závisle proměnnou *y*. Regresní model $y = f(x) + \varepsilon$ se zde skládá z *p* kubických křivek definovaných na *p* navazujících úsecích osy $x, \left(-\infty, x_{u,1}\right) \cup \left\langle x_{u,1}, x_{u,2}\right) \cup ... \cup \left\langle x_{u,p-1}, +\infty\right)$. Hodnoty $x_{u,i}$ se nazývají uzly a mohou být definovány uživatelem. Analytické vlastnosti (hladkost, spojitost, křivost, apod.) a statistické vlastnosti (reziduální rozptyl, rozptyly a významnosti parametrů, rozptyl predikce apod.) těchto křivek jsou předmětem modelování. Model polynomického spline lze zapsat ve tvaru

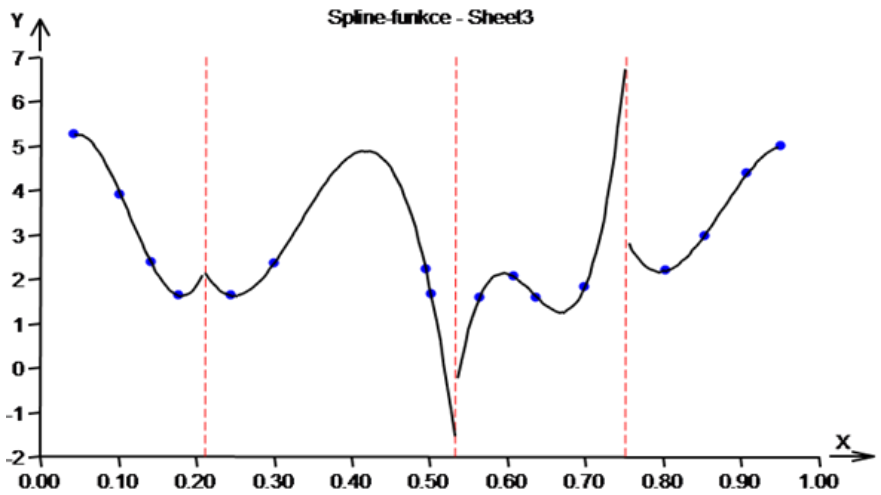
$$y(x) = \left[S^h(x)\right]^T \mathbf{a}_k ,$$

$$\left[S^h(x)\right]^T = \left[1, x, x^2, ..., x^h\right]; h \geq 0$$

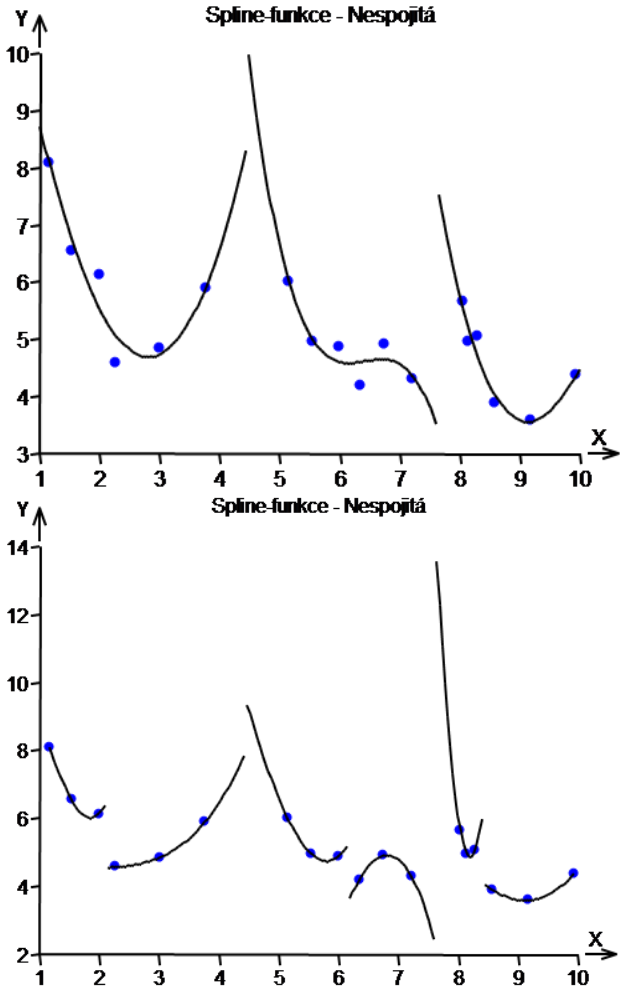
kde *h* je stupeň polynomu, pro *h*=3 se tedy jedná o kubický spline, kterým se zde budeme zabývat. $S^h(x)$ je vektor prediktorů, a $\mathbf{a}_k$ je vektor regresních koeficientů pro *k*-tý úsek. Pokud se naměřená dvojice (x_i, y_i) nachází v *k*-tém úseku (tedy v intervalu ($x_{u,k-1}$; $x_{u,k}$)), je hodnota predikovaná modelem daná

$$f(x_i) = a_{k,1} + a_{k,2} x_i + a_{k,3} x_i^2 + a_{k,4} x_i^3$$

Parametry $a_{k,i}$ se získají pomocí lineární regrese metodou nejmenších čtverců. Pokud bychom netrvali na spojitosti v uzlových bodech, vedl by tento model na *p* nezávislých regresních modelů, viz Obrázek 1. Je zřejmé, že tvar průběhu výsledné křivky velmi závisí na zvoleném počtu úseků a poloze uzlových bodů, jak ilustruje Obrázek 2.



Obrázek 1 Proložení dat nezávislými kubickými polynomy bez požadavku spojitosti



Obrázek 2 a, b Vliv počtu úseků na tvar průběhu, počet úseku *p* = 3 a *p* = 6

Tyto modely jsou ovšem nespojité a mají jen omezené použití. Požadujeme-li spojitost křivky, zařadíme do modelu podmínku spojitosti ve funkční hodnotě

$$\left[S^h(x)\right]^T \mathbf{a}_{k-1} = \left[S^h(x)\right]^T \mathbf{a}_k .$$

Podobně lze pokračovat dále a požadovat spojitost prvních a druhých derivací (zde $h = 3$).

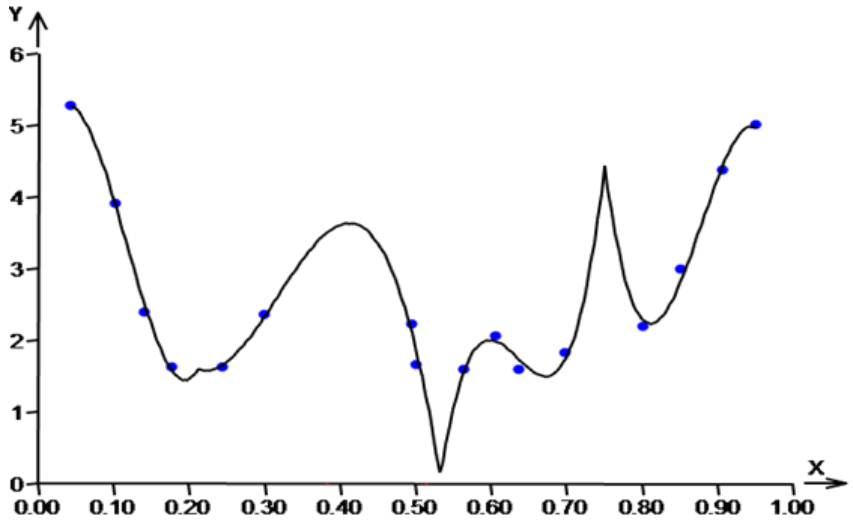
$$\left[S^{h'}(x)\right]^T \mathbf{a}_{k-1} = \left[S^{h'}(x)\right]^T \mathbf{a}_k , \left[S^{h''}(x)\right]^T \mathbf{a}_{k-1} = \left[S^{h''}(x)\right]^T \mathbf{a}_k$$

přičemž první a druhá derivace polynomu

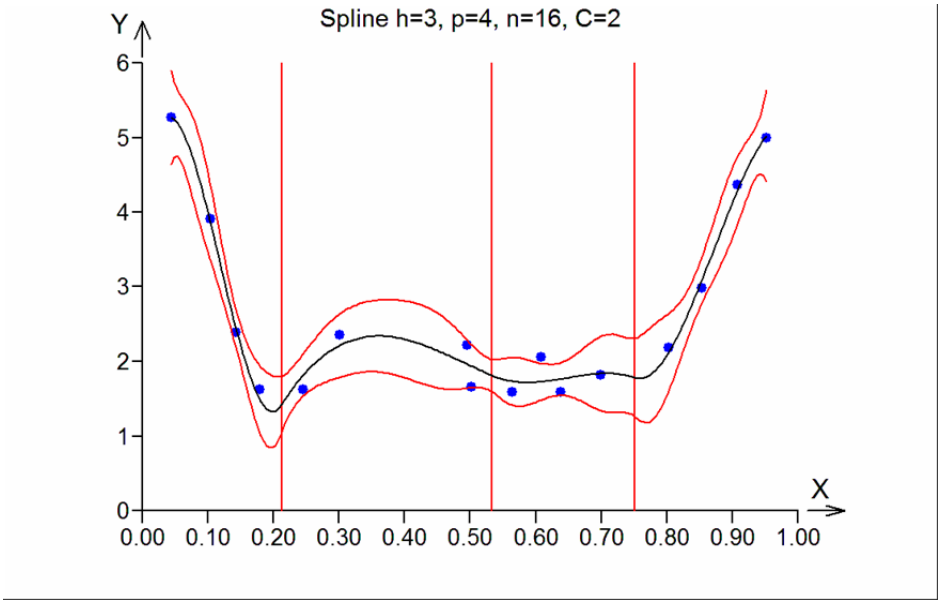
$$\left[S^{h'}(x)\right]^T = \frac{\partial\left[S^h(x)\right]^T}{\partial x} = \left[0,1,2x,3x^2,\ldots,hx^{h-1}\right]$$

$$\left[S^{h''}(x)\right]^T = \frac{\partial^2\left[S^h(x)\right]^T}{\partial x^2} = \left[0,0,2,6x,\ldots,h(h-1)x^{h-2}\right].$$

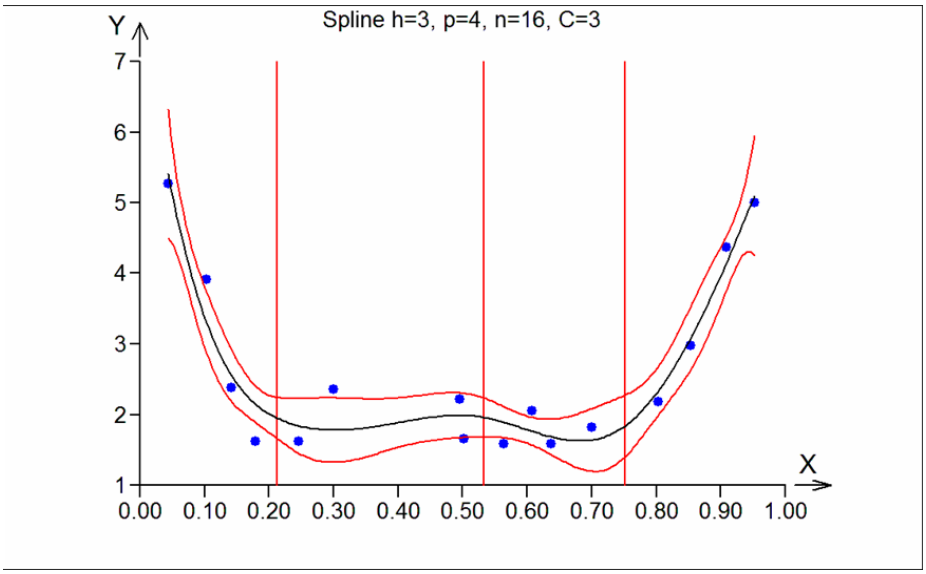
Tak získáme křivku hladkou do první derivace a spojitou do druhé derivace. Taková křivka spolu s možností volby počtu a polohy uzlových bodů $x_{u,i}$ bude vyhovovat prakticky všem požadavkům na proložení dat neparametrickou regresní křivkou. Výhodou regresního kubického spline proti jádrovému vyhlazení nebo klouzavému průměru je nevychýlenost v oblasti maxima nebo na okrajích závislosti. Navíc je možné získat statistické vlastnosti proložení (hodnoty a směrodatné odchylky (resp. kovarianční matici) odhadu regresních koeficientů, intervaly spolehlivosti predikce, atd.). Příklad vždy stejných dat se stejnými uzly proložených splinem spojitým pouze v hodnotách $f(x)$ uvádí Obrázek 3, v hodnotách $f(x)$ a prvních derivacích $f'(x)$ Obrázek 4 a spline spojitý v hodnotách $f(x)$, prvních derivacích $f'(x)$ a druhých derivacích $f''(x)$ uvádí Obrázek 5. Okrajové podmínky na $f(x)$, $f'(x)$ a $f''(x)$ nejsou kladeny.



Obrázek 3 Proložení dat s požadavkem spojitosti funkčních hodnot

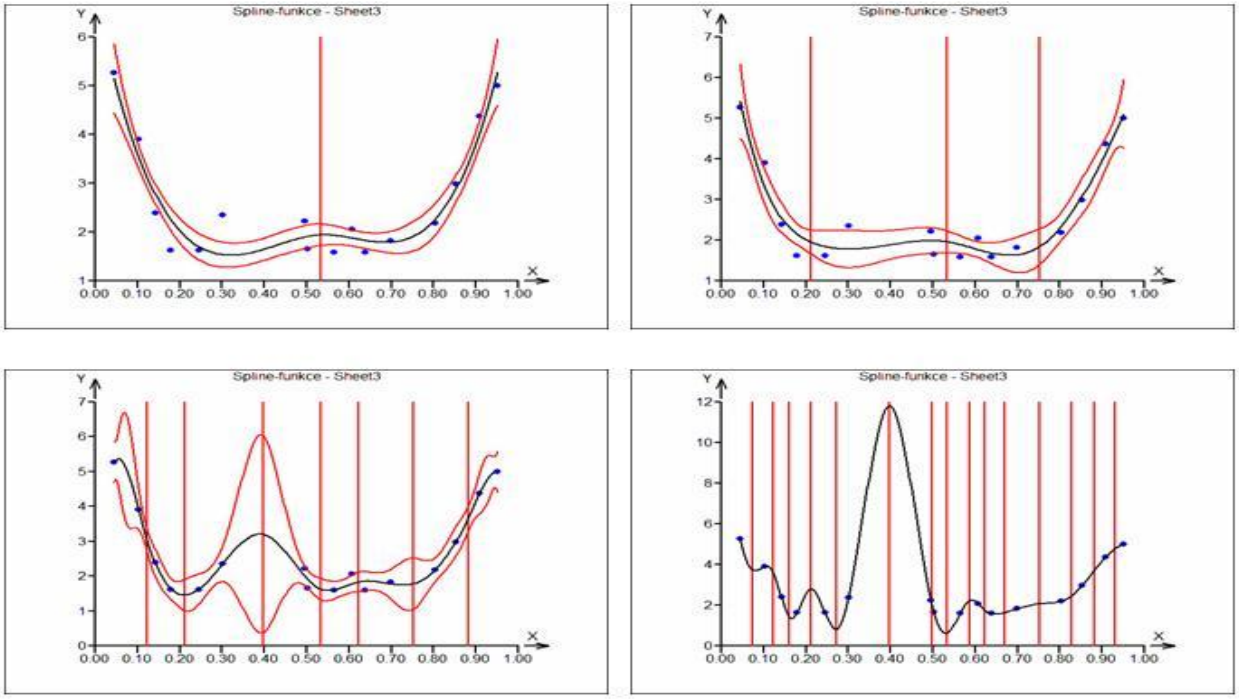


Obrázek 4 Proložení dat s požadavkem spojitosti funkčních hodnot a prvních derivací



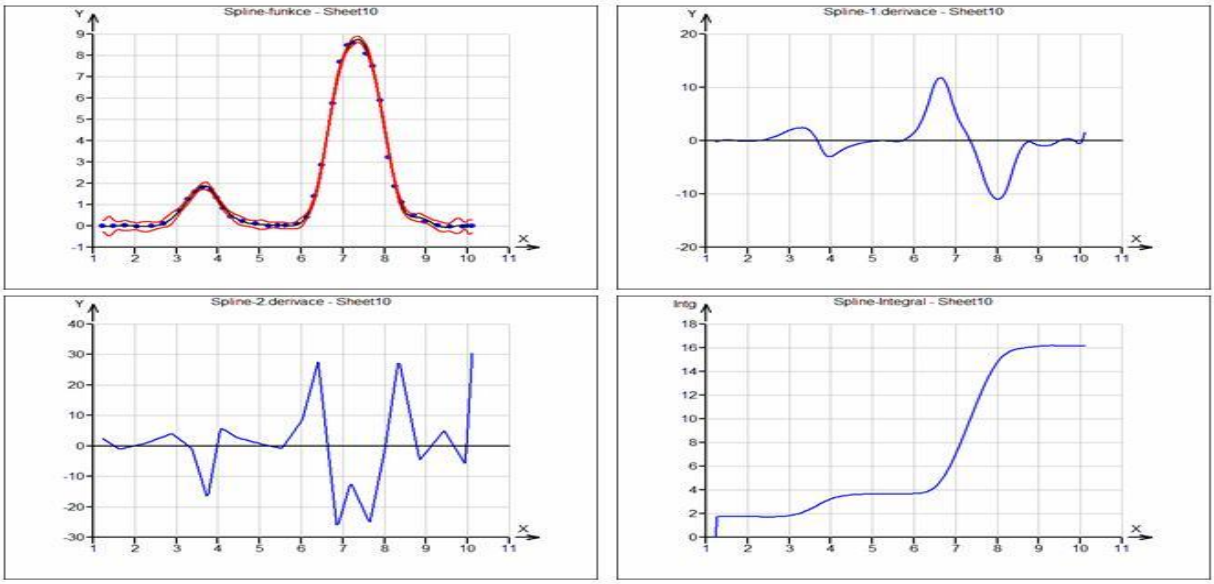
Obrázek 5 Proložení dat s požadavkem spojitosti funkčních hodnot, prvních a druhých derivací

Vliv počtu uzlů ilustruje Obrázek 6. Je-li počet bodů v jednotlivých úsecích příliš malý (například 1), je možné získat interpolační spline, který prochází všemi body. Pro takovou funkci ovšem není možné získat statistické vlastnosti (interval spolehlivosti) a často nelze ani jednoznačně vypočítat regresní koeficienty, neboť model je přeurčený, počet neznámých koeficientů je větší, než počet dat. Zde lze však využít techniku obecné inverze (např. Moore-Penroseovy pseudoinverze), pomocí níž lze řešit i singulární lineární systémy.

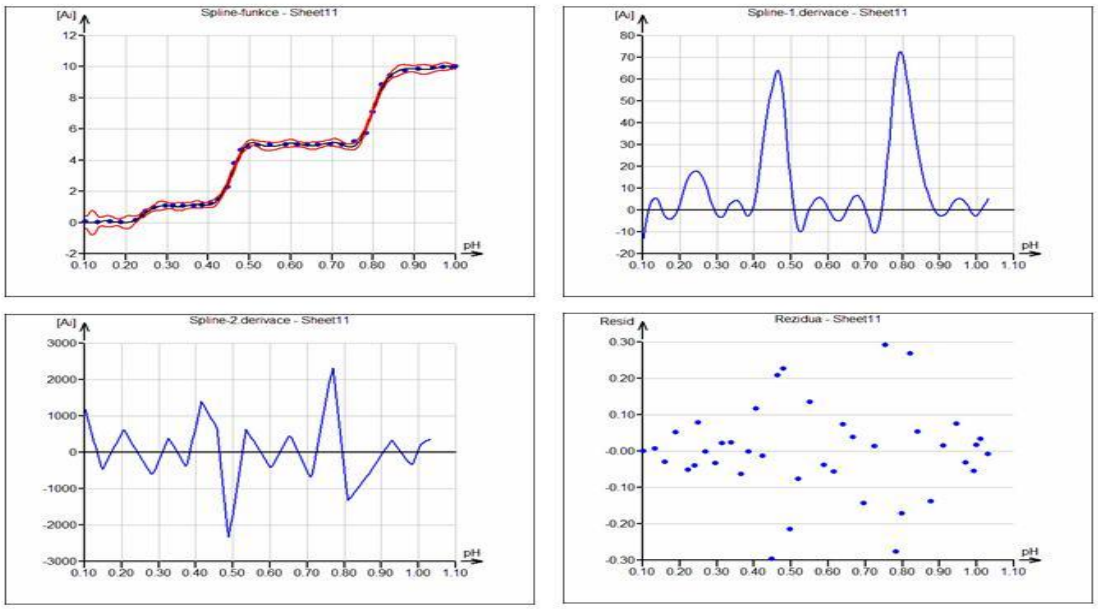


Obrázek 6 Příklad vlivu počtu uzlů na tvar průběhu spline (data jsou vždy stejná)

Analytický tvar kubického spline dovoluje snadno konstruovat průběhy prvních a druhých derivací a integrálu. Jejich využití ilustruje Obrázek 7 a Obrázek 8.



Obrázek 7 Příklad využití spline pro nalezení inflexních bodů, maxim a integrálu



Obrázek 8 Příklad využití spline pro analýzu kumulativních dat (např. titrace, apod.)

Data a parametry

Standardní data pro kubický spline jsou ve dvou sloupcích, jeden sloupec obsahuje hodnoty nezávisle proměnné, druhý hodnoty závisle proměnné (na jejich fyzickém pořadí v datové tabulce samozřejmě nezáleží). Navíc je možné použít další sloupce, v nichž budou uživatelem definované polohy uzlů, body pro výpočet predikce a váhy jednotlivých měření. Strukturu dat ilustruje Obrázek 9 a Obrázek 10. Pokud volíme polohu uzlů ručně a ovlivnit tak průběh křivky, je užitečné si uvědomit, že každý úsek spline může obsahovat nejvýše jedno maximum, jedno minimum a jeden inflexní bod. Váhy se používají nejčasěji při

opakovaném měření se stejným výsledkem. Například dva řádky s hodnotami x=3, y=10; x=3, y=10 lze nahradit jedním řádkem s váhou 2: x=3, y=10; w=2. Jsou-li zadány hodnoty X pro predikci, vypočítá se v těchto bodech predikovaná hodnota y, její interval spolehlivosti a hodnota první a druhé derivace v tomto bodě.

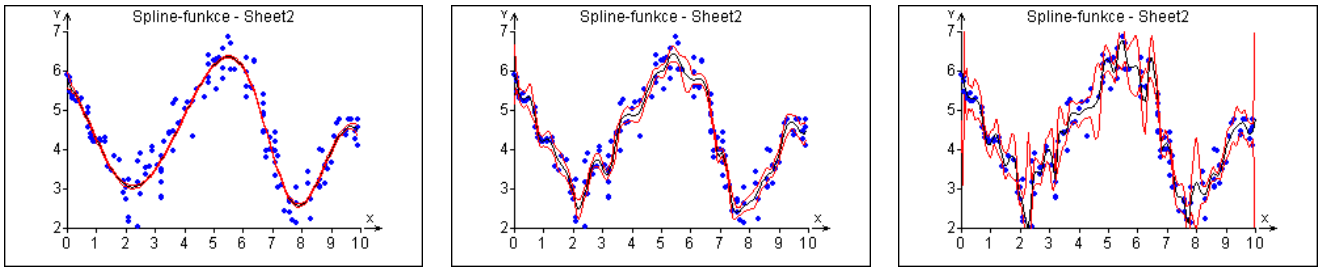
Upozornění: Šířka m charakteristické matice regrese je v případě kubického spline m=4p. Při příliš velkém počtu úseků nelze proto zajistit numerickou stabilitu, doporučuje se proto volit malý počet úseků, typicky p = 2 až 10, ne však více než 50. Příklad nevhodné volby p ilustruje Obrázek 11.

X	Y
0.2	5.91
0.7	4.77
1.4	2.93
1.9	2.93
2.6	2.31
3.5	2.67
3.9	2.46
4.7	2.9
5.7	4.71
6.7	5.35
7.1	5.5
8.6	5.97

Obrázek 9 Základní data pro kubický spline

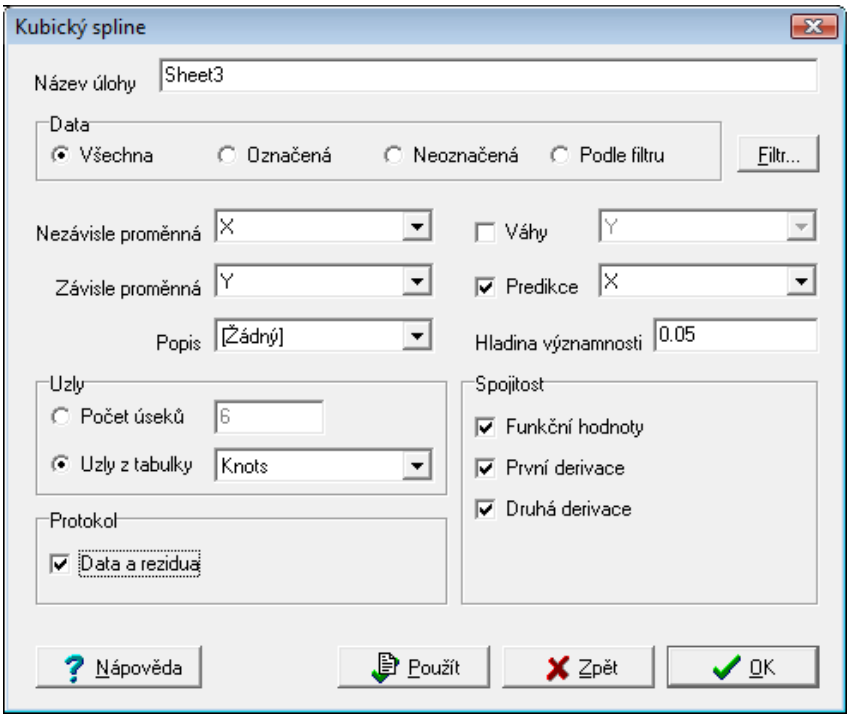
X	Y	knots	predict	weight
0.2	5.91	2	0	1
0.7	4.77	4	1	1
1.4	2.93	6	2	2
1.9	2.93		3	4
2.6	2.31		4	5
3.5	2.67		5	5
3.9	2.46		6	4
4.7	2.9		7	4
5.7	4.71		8	3
6.7	5.35		9	2
7.1	5.5			2
8.6	5.97			2

Obrázek 10 Úplná data pro kubický spline



A: p=5 B: p=30 C: p=50

Obrázek 11 Vhodná (A, B) a nevhodná (B, C) volba počtu úseků p při počtu dat n=150



Obrázek 12 Dialogový panel pro kubický spline

V dialogovém panelu se zvolí sloupec hodnot nezávisle a závisle proměnné, sloupec vah a hodnot pro predikci. Zadáme hladinu významnosti pro intervaly spolehlivosti a testy významnosti, obvykle 0.05. Sloupec *Popis* je použit k označení bodů v grafech, je-li vybrán. Polohu uzlů lze zadat buď v tabulce a příslušný sloupec se vybere v poli *Uzly z tabulky*. Pokud se vybere možnost *Počet úseků* a zadá číselná hodnota p, vypočítají se polohy uzlů pro n dat automaticky tak, aby byl v každém úseku stejný počet bodů n/p. Pokud není tento podíl celé číslo, je počet bodů v prvních $(p - 1)$ úsecích roven int(n/p) a v posledním úseku jsou zbývající body. Je tedy vhodné volit pokud možno počet úseků tak, aby byl zbytek po dělení n/p co nejmenší, Zvolíme-li například pro 12 bodů 7 úseků, získáme 6 úseků s jedním bodem a jeden úsek se šesti body. Ve skupině *Spojitost* vybereme typy požadované spojitosti ve funkčních hodnotách, prvních a druhých derivacích. Nemáme-li zvláštní důvody, obvykle označíme všechny spojitosti, čímž získáme hladší průběh funkce a jejích intervalů spolehlivosti. Podmínky spojitosti navíc zvyšují počet statistických stupňů volnosti, zvyšují stabilitu řešení a obvykle zpřesňují predikci. Ve skupině *Data* vybereme případnou podmnožinu dat. Je-li označena položka *Data a rezidua*, vypíší se do protokolu tabulka predikovaných hodnot a reziduí. Po kliknutí na tlačítko *OK* nebo *Použít* se spustí výpočet.

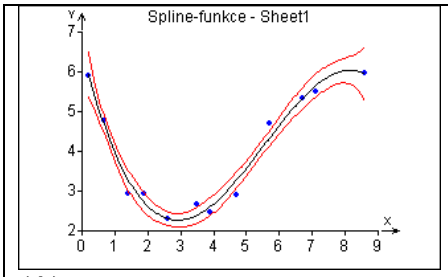
Protokol

Název úlohy	Zadaný název úlohy
Data	Podmnožina vybraných dat (Všechna, označená, neoznačená, případně zadaný filtr)
Nezávisle proměnná	Sloupec nezávisle proměnné
Závisle proměnná	Sloupec závisle proměnné
Počet dat	Počet řádků n
Počet uzlů	Počet uzlů $(p - 1)$
Počet úseků	Zadaný počet úseků p

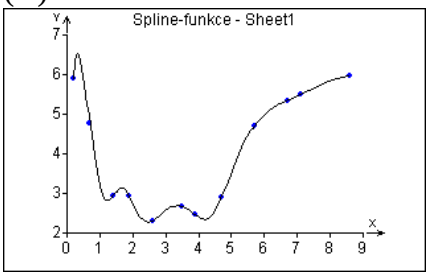
Spojitost ve funkční hodnotě	Požadovaná podmínka spojitosti ve funkčních hodnotách (ANO / NE)
Spojitost v první derivaci	Požadovaná podmínka spojitosti v prvních derivacích (ANO / NE)
Spojitost ve druhé derivaci	Požadovaná podmínka spojitosti v druhých derivacích (ANO / NE)
Polohy uzlů	Číslo uzlu, Poloha uzlu
Parametry modelu	V této tabulce jsou uvedeny počty dat v jednotlivých úsecích a vypočítané hodnoty koeficientů regresního polynomu $y = a_0 + a_1x + a_2x^2 + a_3x^3$, s označením A(0)=absolutní člen, A(1)=lineární člen, A(2)=kvadratický člen, A(3)=kubický člen
Tabulka predikovaných hodnot	Byl-li vybrán sloupec pro predikci, uvádí tato tabulka pro každou zadanou hodnotu X: predikovanou hodnotu Y(Y-před), spodní a horní mez spolehlivosti této predikce(Spodní mez, Horní mez) a hodnotu první a druhé derivace (1.derivace, 2.derivace)
Res. součet čtverců	Reziduální součet čtverců odchylek
Res. směr. odchylka	Reziduální směrodatná odchylka
Reziduální rozptyl	Reziduální rozptyl
Prům. abs. odchylka	Průměrná absolutní odchylka
Stupňů volnosti eff.	Počet skutečných stupňů volnosti, $\nu_{eff} = n + 3*(p - 1) - 4p$ Je-li počet stupňů volnosti menší než 1, není možné vypočítat intervaly spolehlivosti. Hodnoty zadaných vah se neberou v úvahu.
Tabulka extrémů a inflexů	Obsahuje dvě tabulky: tabulku extrémů a tabulku inflexů v intervalu (x_{min}, x_{max}). Pokud extrémy nebo inflexy nejsou přítomny, tabulka se nevytvoří.
Tabulka extrémů	Tabulka uvádí analytické hodnoty extrémů (tedy minim a maxim) regresní křivky: počet extrémů, pořadí extrému, typ extrému (minimum nebo maximum), hodnota X, hodnota Y, a 2.derivace v extrému (první derivace v extrému je vždy nula).
Tabulka inflexů	Tabulka uvádí analytické hodnoty inflexů (tedy bodů přechodu z konkávního na konvexní tvar, nebo opačně) regresní křivky: počet inflexů, pořadí inflexu, hodnota X, hodnota Y, a 1.derivace v inflexu (druhá derivace v inflexu je vždy nula).
Tabulka dat a reziduí	Byla-li v dialogovém panelu označena položka *Data a rezidua*, jsouv tabulce hodnoty X a Y, predikovaná hodota Y-před a odchylka hodnoty Y od regresní křivky (reziduum).

Grafy

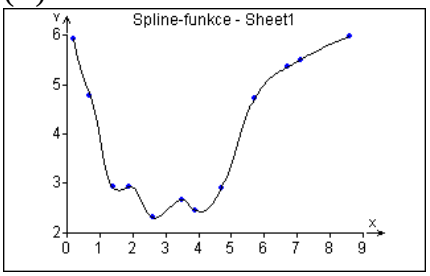
Spline – funkce	Graf regresní funkce, je-li počet stupňů volnosti větší než 0, kolem regresní funkce vyznačen červeně (1 – α) pás spolehlivosti na



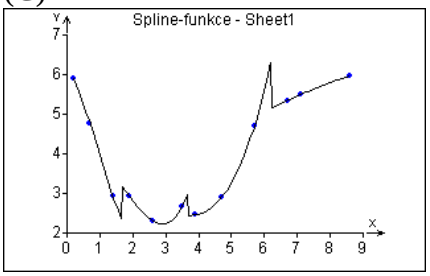
(A)



(B)



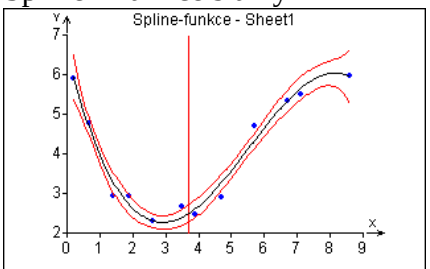
(C)

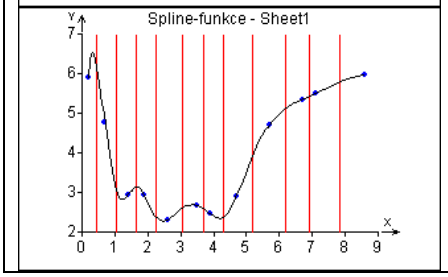


(D)

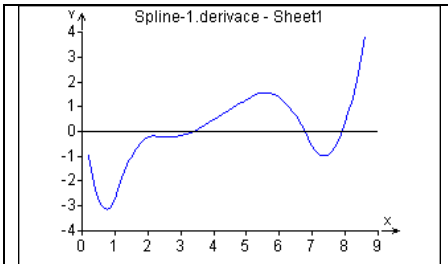
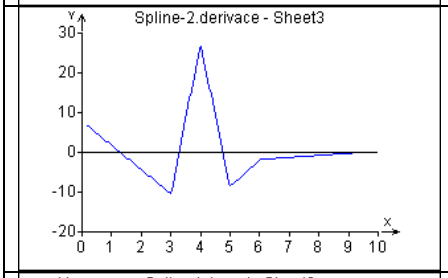
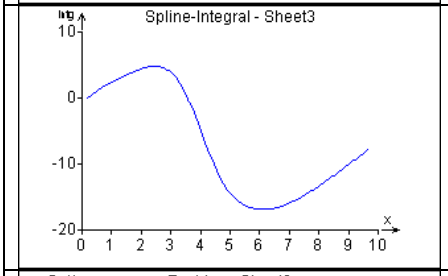
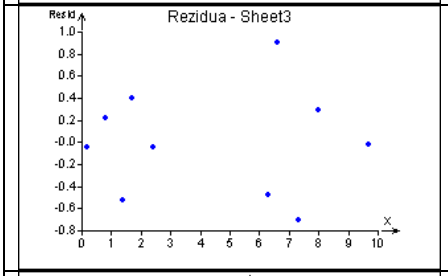
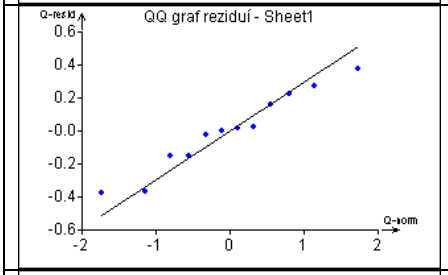
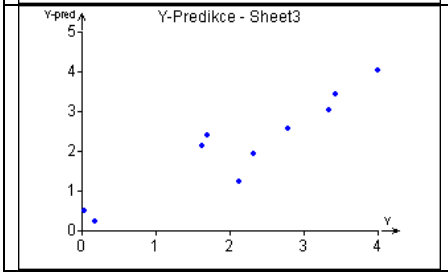
zadané hladině významnosti α (graf A). Pokud zadáme počet úseků shodný s počtem dat, získáme při všech podmínkách spojitosti interpolační spline, který prochází všemi body (graf B). Při snížení nároků na spojitost, např. požadujeme-li pouze spojitost v $f(x)$ a $f^{(1)}(x)$, můžeme získat interpolační spline i při $p = n/2$ (graf C). Graf (D) ilustruje proložení nespojitým splinem s p=4 a n=12. Ve většině reálných případů bude však smysluplný spline se všemi podmínkami spojitosti a vhodně volenými uzly p<<n.

Spline – funkce s uzly





Graf regresní funkce shodný s předchozím grafem s vyznačenými polohami uzlů.

	Graf první derivace spline-funkce $f^{(1)}(x)$. Nulové body (průsečíky s osou x) odpovídají maximu nebo minimu funkce $f(x)$. Průběh této křivky a její nulové body jsou rovněž v protokolu v tabulce „Tabulka predikovaných hodnot“ a „Tabulka extrémů a inflexů“.
	Graf druhé derivace spline-funkce $f^{(2)}(x)$. Nulové body (průsečíky s osou x) odpovídají inflexním bodům funkce $f(x)$. Protože f(x) je polynom 3. stupně, má druhá derivace tvar p lineárních úseků se zlomy v uzlových bodech. Průběh této křivky a její nulové body jsou rovněž v protokolu v tabulce „Tabulka predikovaných hodnot“ a „Tabulka extrémů a inflexů“.
	Graf integrálu regresní funkce $\int_{x_{\min}}^{x} f(z)\,\mathrm{d}z$.
	Graf reziduí $y_i - f(x_i)$, je-li v reziduích patrný zřetelný trend, je možné, že je regresní spline chybně zvolený a je třeba přidat uzly, nebo změnit jejich polohu.
	QQ-graf reziduí, tento graf slouží k posouzení normality reziduí. Jsou-li body rozptýleny kolem přímky, mají rezidua přibližně normální rozdělení.
	Graf Y-Predikce, posouzení kvality proložení dat. Čím těsněji leží data na přímce, tím těsnější je proložení. Cílem regrese však obecně není co nejlepší proložení (jako je interpolační spline), ale co „nejrozumnější“ proložení přiměřeně komplikovaným modelem.