Nelineární regrese

Menu:	QCExpert	Nelineární regrese

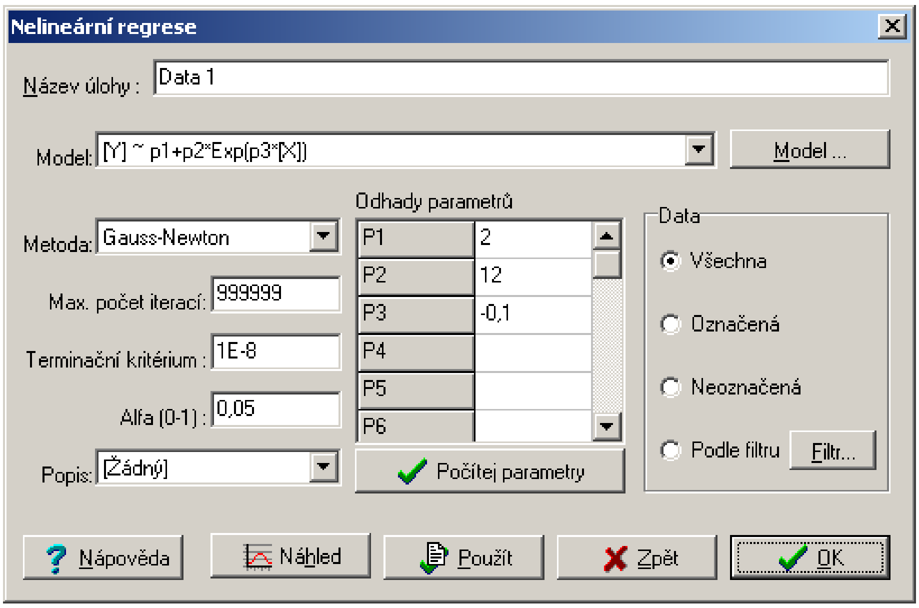
Modul nelineární regrese slouží pro tvorbu a analýzu explicitních nelineárních regresních modelů v obecném tvaru

$y = F(\mathbf{x},\mathbf{p})$ **(1-1)**

kde y je nezávisle proměnná, $\mathbf{x} = (x_1, x_2, \ldots x_q)$ jsou nezávisle proměnné, q je počet proměnných, $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ jsou parametry, m je počet parametrů, $F(\mathbf{x},\mathbf{p})$ je libovolná funkce nezávisle proměnné a parametrů. Maximální počet parametrů je 32, maximální počet proměnných je 254. Předpokládá se, že $\mathbf{x}$ je pokud možno deterministická (nenáhodná) nezávisle nastavená nebo jinak zjištěná veličina. Veličina y na $\mathbf{x}$ závisí, ale její hodnota je zatížena náhodnou chybou ε. Parametry $\mathbf{p}$ se odhadují na základě dat a daného modelu nelineární metodou nejmenších čtverců. Uživatel definuje model v základním dialogovém panelu nelineární regrese (Obrázek 1) resp. v okně *Tvorba modelu*.
Poznámka: Lze-li model vyjádřit ve tvaru lineárním vzhledem k parametrům, použijte lineární regresi (viz předchozí kapitola), kde je výpočet parametrů jednoznačný a není potřeba zadávat první odhady parametrů. Typické příklady takových modelů jsou: $y = p_1x + p_2\ln(x)$, $y = p_1 + \exp(p_2x)$, který lze linearizovat na $\ln y = \ln p_1 + p_2x$ (zde je třeba použít kvazilinearizaci), polynomické modely jako $y = p_1x + p_2x^2 + p_3x^3 + p_4$, atd.

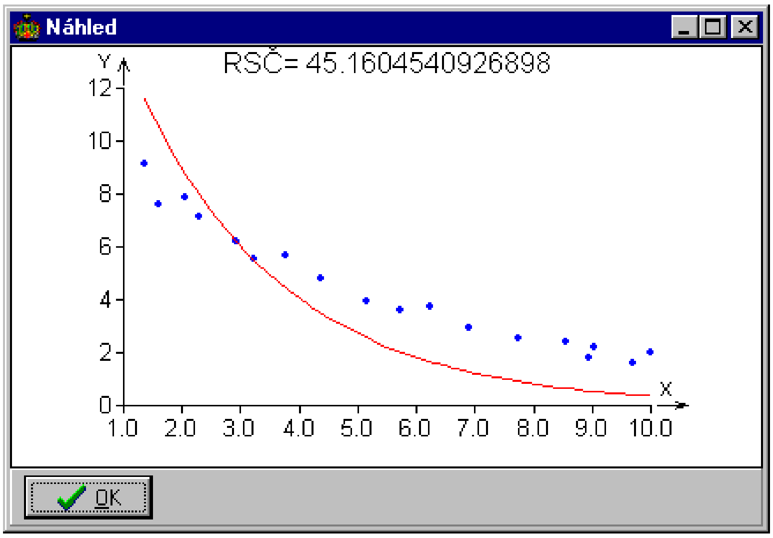
Data a parametry

Výpočet hodnot parametrů $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ probíhá na základě dat, která jsou uspořádána ve sloupcích datové tabulky. Sloupec reprezentuje hodnoty jedné proměnné. K identifikaci proměnných se používá záhlaví sloupce. Proměnné a parametry jsou součástí modelu, který se zadává po stisku tlačítka *Model…*, případně se vybere ze seznamu přímo v okně *Nelineární regrese*, pokud byl model definován již dříve. Zadání modelu je podrobněji popsáno dále. Zadaný model se objeví v okně *Model* dialogového okna Nelineární regrese.



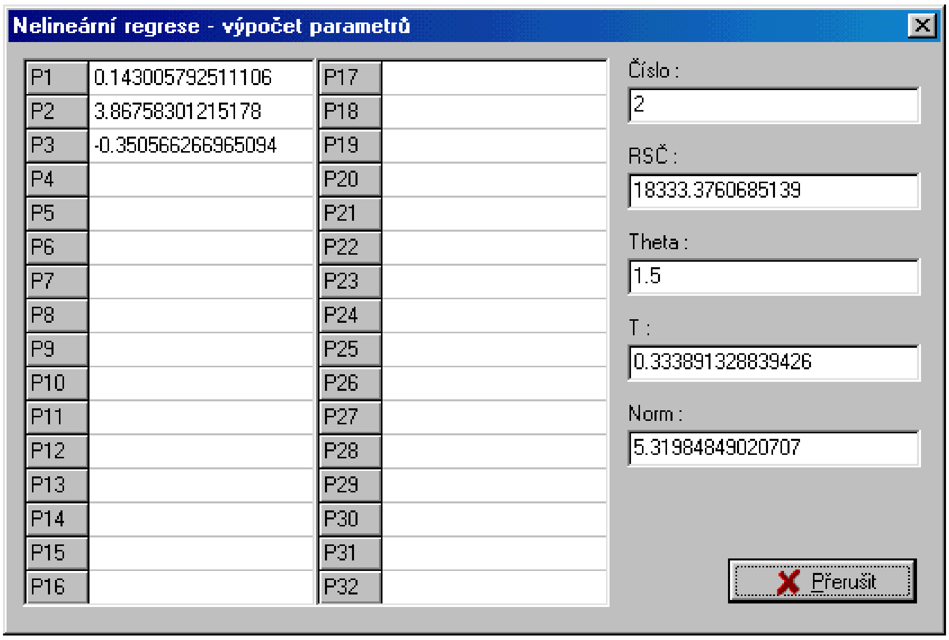
Obrázek 1 Základní dialogový panel pro Nelineární regresi

V poli *Název úlohy* je možno zadat text, který bude uveden v záhlaví protokolu i grafů. Ze seznamu *Metoda* vybereme optimalizační metodu (*Gauss-Newton, Marquardt, gradient, dog-leg, simplex*), v poli *Max počet iterací* lze omezit počet iterací. Terminační kritérium je maximální hodnota gradientu, popř. norma kroku parametrů v jedné iteraci, při níž výpočet končí. Alfa je pravděpodobnost (hladina významnosti) použitá při výpočtu intervalů spolehlivosti a při statistických testech. Skupina tlačítek *Data* určuje, zda se použijí pro výpočet všechna data, pouze označená data, nebo pouze neoznačená data. V poli *Odhady parametrů* je vždy nutno zadat počáteční odhady parametrů p_1, … p_m. Tyto odhady by měly být pokud možno co nejblíže optimálním hodnotám a jejich volbě je nutné věnovat pozornost. Příliš hrubé nebo zcela chybné odhady mohou vést k nenalezení správných hodnot případně k neúměrnému zvýšení počtu iterací a výraznému prodloužení výpočtu. Aby mohl uživatel ověřit, popřípadě doladit počáteční odhady, je na panelu *Nelineární regrese* tlačítko *Náhled*, které zobrazí data a průběh modelu s parametry uvedenými v poli *Odhady parametrů*, v záhlaví je uvedena hodnota součtu čtverců, *RSČ*, (**Obrázek** 2). Je žádoucí, aby čára probíhala v blízkosti bodů. Stiskem tlačítka *OK* se vrátíme do panelu *Nelineární regrese*, kde lze parametry pozměnit a znovu zobrazit náhled. Je-li v modelu více než jedna nezávisle proměnná, zobrazí se graf predikce proti hodnotám závisle proměnné s přímkou y=x, na níž při ideálním proložení (s nulovou chybou) data leží. Okno *Náhled* neobsahuje žádné interaktivní prvky a nelze v něm použít příkaz Kopíruj (*Ctrl-C*).



Obrázek **2** Okno ***Náhled (pro první odhad parametrů)***

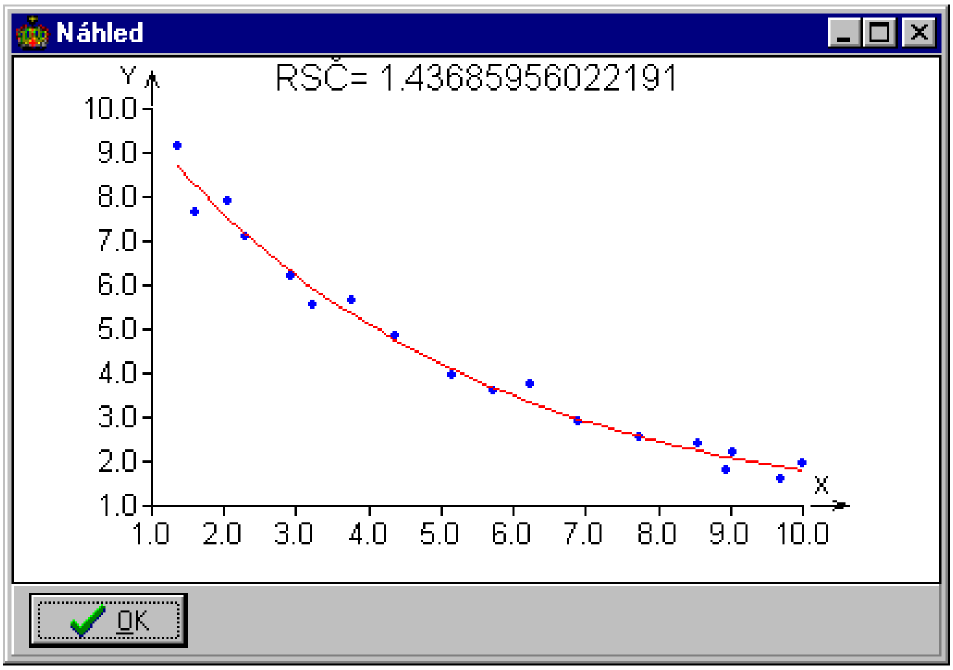
Po zadání odhadů parametrů se spustí výpočet stiskem tlačítka *Počítej* parametry. Zobrazí se panel *Výpočet parametrů* (Obrázek 3).



Obrázek **3** Panel ***Výpočet parametrů***

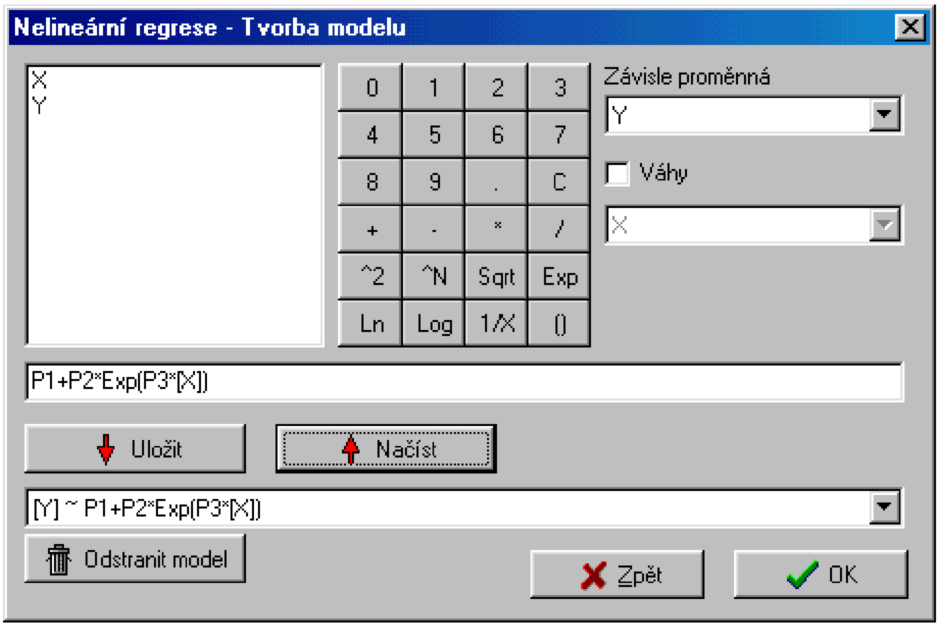
V něm se průběžně vypisují hodnoty jednotlivých parametrů, číslo iterace, aktuální součet čtverců *RSČ*, další parametry související s použitou optimalizační metodou a norma změny parametrů *Norm*, která se porovnává se zadanou hodnotou terminačního kritéria. Je-li norma menší než terminační kritérium, výpočet je ukončen. Výpočet lze rovněž kdykoliv ukončit stiskem tlačítka *Přerušit*. V takovém případě se jako výsledné berou hodnoty parametrů z poslední iterace.

Po ukončení výpočtu se program vrací zpět do panelu *Nelineární regrese*, pole *Odhady parametrů* obsahuje hodnoty parametrů z poslední provedené iterace. V případě normálního ukončení výpočtu jsou to hodnoty optimální. V případě pochybností lze opět zobrazit náhled (Obrázek 4), popřípadě opakovat výpočet stejnou nebo jinou metodou. Výpočet jinou metodou lze provést i po předčasném přerušení například z důvodu příliš pomalé konvergence či divergence hodnot parametrů. Jako počáteční odhady se v tom případě použijí hodnoty, ke kterým dospěla předcházející metoda. Tyto odhady lze navíc před spuštěním nového výpočtu upravit. Tento postup (úprava odhadů parametrů, výpočet libovolnou metodou, případné předčasné přerušení výpočtu, náhled) lze v libovolném pořadí opakovat. Stiskem tlačítka *OK* v hlavním panelu *Nelineární regrese* potvrdí uživatel hodnoty parametrů jako konečné, provede se výpočet statistických charakteristik, vytvoří se protokol a grafy. ***Pozor***: stiskne-li se OK před provedením výpočtu parametrů, žádná optimalizace parametrů se neprovede a zadané parametry jsou považovány za optimální!



Obrázek **4** Náhled po výpočtu parametrů

Tvorba modelu: Tlačítkem *Model…* se otevře panel pro tvorbu modelu (Obrázek 5). Pokud jsme již dříve nějaké modely vytvořili, je možné z nich jeden pouze vybrat bez otevření panelu *Tvorba modelu*, je ale nutno dbát, aby se shodovaly názvy proměnných v modelu a v tabulce s daty. Panel *Tvorba modelu* nabízí v levé části seznam proměnných v aktuálním listu tabulky s daty, z nichž se tvoří regresní model. V pravé části nahoře zvolíme nezávisle proměnnou. Po zaškrtnutí políčka Váhy lze zvolit sloupec s vahami w_i jednotlivých hodnot závisle proměnné (jedná se o koeficient, jímž se násobí příslušné reziduum, nikoli jeho čtverec). Zadané váhy se normují tak, aby jejich součet byl roven počtu dat n. Není-li políčko Váhy zaškrtnuté, uvažují se jednotkové váhy $w_i=1$. Uprostřed panelu jsou pomocná tlačítka pro tvorbu modelu. Ve spodní části panelu je editační řádek, v němž se model sestavuje a seznam dříve vytvořených modelů. Tlačítkem *Uložit* se převede hotový model do seznamu a nastaví se jako aktuální, tlačítkem *Načíst* se aktuální model ze seznamu převede do editačního řádku, kde jej lze modifikovat.



Obrázek 5 Dialogový panel pro tvorbu nelineárního modelu

Pokyny pro sestavování modelu:

Dvojitým kliknutím na proměnnou v seznamu proměnných opíšeme tuto proměnnou do editačního řádku. Název proměnné se uvádí vždy v hranatých závorkách. Parametry musí být označeny symboly P1, P2, …, které korespondují s hodnotami v poli *Odhady parametrů* v hlavním panelu *Nelineární regrese*. Při psaní složitějších výrazů je možno výhodně použít pomocných tlačítek s funkcemi. Je-li v editačním řádku označena část výrazu, stisknutím tlačítka funkce se tato funkce aplikuje na označenou část. Například výraz ln([*x*]+1) sestavíme takto: dvojitým kliknutím přepíšeme proměnnou *x* (v datech musí být sloupec tohoto jména): `[x]`; připíšeme + 1; celý výraz označíme: `[x]+1` a klikneme na tlačítko *Ln*, výsledkem bude: `ln([x]+1)`. Podobně použijeme tlačítka *^2*, *^A*, *Sqrt*, *Exp*, *Log*, *1/X*, *( )*. Tlačítko *C* smaže editační řádek. Další funkce je nutno psát ručně. Když je model sestaven, tlačítkem *Uložit* jej uložíme do seznamu modelů ve spodní části panelu. Tlačítkem *Načíst* načteme aktuální model ze seznamu modelů a můžeme jej modifikovat. Tlačítkem *Odstranit model* vymažeme aktuální model v seznamu modelů: pozor, tuto operaci nelze vrátit zpět! Tlačítkem *OK* sestavení modelu ukončíme.

Při dodržení běžných syntaktických konvencí a umisťování názvu proměnných kdo hranatých závorek lze model zapsat i bez použití myši a pomocných tlačítek. Hotové modely můžeme ze seznamu modelů vybírat přímo v hlavním panelu *Nelineární regrese* bez otevření panelu *Tvorba modelu*, pozor na souhlas názvů proměnných.

Metody výpočtu parametrů

Úkolem nelineární regrese je nalézt takové parametry daného modelu, které minimalizují nějakou vzdálenost tohoto modelu od naměřených hodnot,

$$\min_{\mathbf{p}} S(\mathbf{p}) = \min_{\mathbf{p}} D(\mathbf{y},\hat{\mathbf{y}}), \qquad \textbf{(1-2)}$$

kde D značí vzdálenost, $\mathbf{y}$ je vektor daných hodnot závisle proměnné a $\hat{\mathbf{y}}$ je vektor hodnot predikovaných modelem pro dané hodnoty nezávisle proměnné. Nejčastěji se používá Eukleidovská vzdálenost

$$S(\mathbf{p}) = \|\mathbf{y} - \hat{\mathbf{y}}\| = \|\mathbf{e}\| = \sqrt{\sum_{i=1}^{n} [y_i - \hat{y}_i]^2} = \sqrt{\sum_{i=1}^{n} e_i^2} \tag{1-3}$$

Pro účel minimalizace není třeba uvažovat odmocninu, výraz (1-3) tak přechází na součet čtverců odchylek, tedy na podmínku nejmenšího součtu čtverců. Minimalizace součtu čtverců se provádí iterativně numerickou nelineární optimalizací. Vzhledem k tomu, že obecně neexistuje ideální algoritmus nelineární optimalizace, nabízí QCExpert™ šest různých metod (algoritmů) pro nalezení optimálních hodnot parametrů $\mathbf{p}^*$ na základě uživatelem zadaných prvních odhadů $\mathbf{p}_0$. Prvních pět patří mezi derivační metody, které využívají informace o prvních, případně druhých derivacích kritéria $S(\mathbf{p})$ podle parametrů, poslední simplexová metoda je nederivační, používá pouze samotných hodnot $S(\mathbf{p})$. Obecně lze říci, že derivační metody jsou efektivnější jsou-li počáteční odhady parametrů $\mathbf{p}_0$ blízké optimálním odhadům $\mathbf{p}^*$, případně není-li nelinearita modelu příliš veliká. V opačném případě mohou být derivační algoritmy málo efektivní, případně zcela selhat a výhodnější může být nederivační simplex. Nejsou-li známy dostatečně přesně odhady parametrů, nebo derivační metody selhávají, lze parametry nejprve zpřesnit metodou simplex a v druhé fázi (po eventuálním přerušení výpočtu) použít metodu derivační. Modul *Nelineární regrese* obsahuje následující metody:

Gauss-Newton: Klasický derivační algoritmus vycházející z linearizace modelu v okolí $\mathbf{p}$. Pro modely s nízkou nelinearitou a odhadem $\mathbf{p}_0$ blízkým $\mathbf{p}^*$ vede nejrychleji k cíli. V opačném případě často diverguje. V případě nesprávného kroku optimalizace používá tlumicího parametru $Damp \le 1$, který je vypisován během výpočtu. Základní hodnota parametru je 1.
Marquardt: Smíšený derivační algoritmus kombinující metodu Gauss-Newtonovu a gradientovou. Obecně je spolehlivější, než obě tyto metody zvlášť.

Gradient-Cauchy: Derivační metoda používající směru maximálního spádu $S(\mathbf{p})$ s Cauchyho krokem určeným minimalizací ve směru gradientu. Cauchyho bod je určován heuristickým postupem, aby se zabránilo „zamrznutí“ algoritmu v zahnutém údolí. V případě nesprávného kroku optimalizace používá tlumicího parametru *Damp*, který je vypisován během výpočtu. Základní hodnota parametru je 1. V případě zahnutého údolí kritéria $S(\mathbf{p})$ (tzv. banánový tvar) může být tento algoritmus pomalý.
Dog Leg: Derivační metoda vycházející podobně jako Marquadtova metoda z kombinace gradientu a linearizace navíc se využívá historie optimalizace ke zpřesnění Hessiánu (matice druhých derivací), viz Denis Mei v seznamu literatury. V případě nesprávného směru optimalizace používá tlumicího parametru *Norm*, který je vypisován během výpočtu. Základní hodnota parametru je 1. Během výpočtu jsou dále vypisovány hodnoty pomocných parametrů *Theta* a *T*.

Gradient s pevným krokem: Derivační metoda využívající pouze informace o gradientu kritéria $S(\mathbf{p})$. Tato metoda je vhodná zvláště pro zpřesnění parametrů v počáteční fázi optimalizace. U silně nelineárních modelů je v blízkosti minima pomalá. V případě nesprávného kroku optimalizace používá tlumicího parametru *Damp*, který je vypisován během výpočtu. Základní hodnota parametru je 1.
Simplex: Nederivační metoda využívající postupu překlápění mnohostěnu (simplexu) s m+1 vrcholy. Implementace v QCExpertu™ používá znáhodněného heuristického postupu a mutací při konstrukci simplexu. Vzhledem k tomu, že nepoužívá derivace, je vhodná ke zpřesnění a optimalizaci parametrů i pro silně nelineární modely. Její nevýhodou je obvykle pomalost ve srovnání s derivačními metodami. Během výpočtu se vypisuje koeficient zvětšení simplexu *Norm*.

Protokol

Název úlohy	Název úlohy z dialogového panelu.
Hladina významnosti	Hodnota a zadaná v dialogovém panelu, která se používá pro výpočet intervalů spolehlivosti a všechny testy.
Počet stupňů volnosti	Počet dat zmenšený o počet parametrů, n–m.

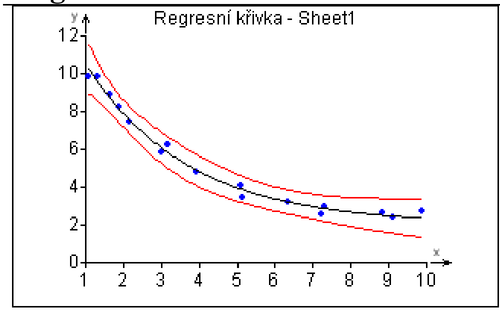
Kvantil t(1-alfa/2,n-m)	Kvantil t-rozdělení.
Kvantil F(1-alfa,m,n-m)	Kvantil F-rozdělení.
Metoda	Použitá metoda (metoda nejmenších čtverců)
Počet platných řádků	Počet řádků s platnými hodnotami všech proměnných.
Počet parametrů	Počet parametrů v regresním modelu.
Metoda	Zvolená numerická metoda optimalizace parametrů.
Nezávisle proměnné	Seznam nezávisle proměnných použitých v regresi.
Závisle proměnná	Závisle proměnná.
Model	Použitý regresní model, vlevo od vlnovky „~“ je závisle proměnná.
Počáteční hodnoty parametrů	Počáteční odhady parametrů pro poslední výpočet.
Výpočet	
Počet iterací	Počet iterací posledního výpočtu.
Ukončení výpočtu	Způsob ukončení výpočtu, v případě úspěšného výpočtu je uvedeno *Konvergence*, při ukončení tlačítkem *Přerušit* je uvedeno *Přerušení uživatelem*, při překročení maximálního počtu iterací je uvedeno *Divergence*, Nebyl-li žádný výpočet parametrů proveden, je uvedeno *Bez výpočtu*. ***Pozor***: slovo *Konvergence* nemusí vždy nutně znamenat úspěšný výpočet! Nutno zkontrolovat i grafický výstup, popř. korelační matici parametrů.
Doba výpočtu	Procesorový čas posledního výpočtu v sekundách.
Max. počet iterací	Maximální povolený počet iterací, jehož překročení se považuje za divergenci výpočtu.
Terminační kritérium	Použité terminační kritérium pro normu změny parametrů.
Odhady parametrů	Nalezená optimální hodnota parametrů, asymptotické odhady směrodatné odchylky a mezí intervalu spolehlivosti pro zadané α.
Korelační matice parametrů	Asymptotické odhady párových korelačních koeficientů parametrů regresního modelu. Na diagonále jsou vždy jedničky. Parametry většiny nelineárních modelů jsou obvykle korelovány. Jsou-li však hodnoty mimo diagonálu velmi blízké +1, resp. –1, je nutno model reparametrizovat (zapsat v jiném algebraickém tvaru), neboť odhady i jejich směrodatné odchylky budou zřejmě nespolehlivé.
Analýza klasických reziduí	Název proměnné, zde má význam pouze pro poslední sloupec (Vícenás. kor.), vlastní čísla nelze jednoznačně přiřadit k jednotlivým proměnným.
Y naměřené	Vlastní čísla korelační matice nezávisle proměnné.
Y vypočítané	Index (číslo) podmíněnosti κ je poměr největšího a nejmenšího vlastního čísla. Maximální hodnota $\kappa_{max} > 1000$ se považuje za indikaci silné multikolinearity.
Směr. odch. Y	Faktor vzrůstu rozptylu v důsledku multikolinearity, hodnoty VIF > 10 se považují za indikaci silné multikolinearity.
Reziduum	Vícenásobný korelační koeficient mezi danou proměnnou a všemi ostatními nezávisle proměnnými.
Reziduum [%Y]	
Statistické charakteristiky regrese	
Vícenásobný korel. koef. R	Vícenásobný korelační koeficient vyjadřuje relativní těsnost proložení (nikoli kvalitu modelu). Korelační koeficient vždy roste (resp. neklesá) s počtem proměnných!

Koeficient determinace R^2	Čtverec vícenásobného korelačního koeficientu.
Predikovaný korel. koef. Rp	Predikovaný korelační koeficient je citlivější na vybočující hodnoty než klasický koeficient.
Stř. kvadratická chyba predikce MEP	Chyba predikce *i*-té hodnoty závisle proměnné spočítaná regresí s vyloučením *i*-tého bodu. Citlivá na vybočující hodnoty a multikolinearitu, důležitá míra kvality regrese.
Akaikeho informační kritérium	AIC, kritérium kvality regrese vycházející z reziduálního součtu čtverců penalizovaného počtem proměnných.
Reziduální součet čtverců	Součet čtverců reziduí.
Průměr absolutních reziduí	Aritmetický průměr absolutních hodnot reziduí
Reziduální směr. odchylka	Směrodatná odchylka reziduí.
Reziduální rozptyl	Rozptyl reziduí
Šikmost reziduí	Šikmost reziduí
Špičatost reziduí	Špičatost reziduí
Cook-Weisbergův test heteroskedasticity	Testuje konstantnost rozptylu chyb. Je-li přítomna heteroskedasticita, je nutno uvažovat o použití vhodných vah.
Hodnota kritéria CW	Vypočítaná testační statistika.
Kvantil Chi^2(1-alfa,1)	Příslušný kvantil χ^2-rozdělení.
Pravděpodobnost	p-hodnota testu, je-li menší než zadaná hladina významnosti, je model statistiky významný.
Závěr	Verbální závěr testu.
Jarque-Berrův test normality	Testuje normalitu rozdělení chyb pomocí rozdělení reziduí.
Hodnota kritéria JB	Vypočítaná testační statistika.
Kvantil Chi^2(1-alfa,2)	Příslušný kvantil χ^2-rozdělení.
Pravděpodobnost	p-hodnota testu, je-li menší než zadaná hladina významnosti, je model statistiky významný.
Závěr	Verbální závěr testu.
Waldův test autokorelace	Testuje přítomnost autokorelace chyb na základě vypočítaných reziduí.
Hodnota kritéria WA	Vypočítaná testační statistika.
Kvantil Chi^2(1-alfa,1)	Příslušný kvantil χ^2-rozdělení.
Pravděpodobnost	p-hodnota testu, je-li menší než zadaná hladina významnosti, je model statistiky významný.
Závěr	Verbální závěr testu.
Znaménkový test reziduí	Neparametricky ověřuje přítomnost závislostí, které nejsou postihnuty modelem.
Hodnota kritéria Sg	Vypočítaná testační statistika.
Kvantil N(1-alfa/2)	Příslušný kvantil normálního rozdělení.
Pravděpodobnost	p-hodnota testu, je-li menší než zadaná hladina významnosti, je model statistiky významný.
Závěr	Verbální závěr testu.
Indikace vlivných dat	
Standardní	Klasické reziduum dělené svojí směrodatnou odchylkou $1/s_r$.sqrt(1-H_{ii}),

	někdy nazýváno studentizované, s_r je reziduální směrodatná odchylka.
Jackknife	Jackknife reziduum, jako Standardní, místo s_r je pro i-tý bod použita směrodatná odchylka získaná vynecháním i-tého bodu. Toto reziduum citlivěji indikuje vybočující body.
Predikované	Predikované reziduum, rozdíl i-té hodnoty nezávisle proměnné od modelu získaného po vynechání i-tého bodu. Toto reziduum citlivěji indikuje vybočující body.
Diag(Hii)	Diagonální prvky projekční matice, velké hodnoty naznačují velký vliv daného bodu na regresi. Součet H_{ii} je roven počtu parametrů. Příliš vlivné body jsou zvýrazněny červeně.
Atkinsonova vzdál.	Příliš vlivné body jsou zvýrazněny červeně.

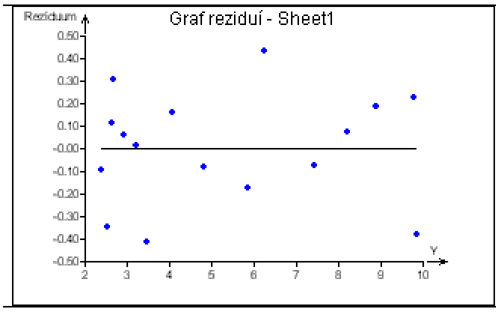
Grafy

Regresní křivka



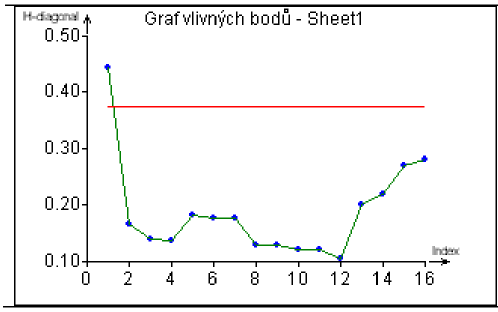
Pokud je vybráno více nezávisle proměnných, tento graf se nekreslí. Je-li v datech pouze jedna nezávisle proměnná, představuje graf průběh regresního modelu. Červeně je vyznačen pás spolehlivosti modelu na zadané hladině významnosti. Je nutné mít na paměti, že pás spolehlivosti predikce, zvláště mimo interval dat, je reálný pouze pokud zvolený model odpovídá skutečnosti. Vhodným zmenšením měřítka (zoom) lze získat detail, nebo naopak průběh i mimo interval měřených dat.

Rezidua



Graf normovaných reziduí, na ose X je hodnota závisle proměnné. vodorovná přímka odpovídá průměru reziduí. Nelineární průběh bodů svědčí o nevhodném nebo neúplném modelu, popř. o nesprávných hodnotách parametrů.

Vlivná data



Diagonální prvky projekční matice $\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, které vyjadřují míru vlivu jednotlivých dat na regresi ($\mathbf{X}$ je matice prvních parciálních derivací modelu podle jednotlivých parametrů v jednotlivých hodnotách nezávisle proměnné). Body nad vodorovnou přímkou se považují za silně vlivné a je třeba jim věnovat pozornost.