

# MATEMATICKÁ STATISTIKA II.

**P5****2006-10-30**

## NELINEÁRNÍ REGRESNÍ MODELY:

### Metoda nejmenších čtverců pro vybrané nelineární funkce:

- ✓ Výpočet parametrů vychází z podmínky minimálnosti čtverců  $\sum_{i=1}^n (y_i - y'_i)^2 = \min$ .
- ✓ Dosazením do výrazu za  $y'_i$  a derivováním podle jednotlivých parametrů funkce lze dospět k soustavě normálních rovnic, ze kterých se parametry vypočítají.
- ✓ Normální rovnice lze sestavovat mechanicky, aniž by jejich vyvození muselo být praktikováno prostřednictvím partiálních derivací. Sestavují se tak, že se každý člen rovnice postupně násobí příslušnou simultánní funkcí nezávisle proměnné u jednotlivých parametrů regresní rovnice a vždy po vynásobení jednotlivými simultánními funkcemi se provede součet.
- ✓ Předpokladem však je, aby regresní rovnice byla aditivního typu a simultánní funkce nezávisle proměnné bez neznámých parametrů.
- ✓ U závisle proměnné se uvádějí empirické hodnoty.
- ✓ Tak první normální rovnice pro funkci  $y'_i = a + \frac{b}{x_i}$  se získá vynásobením jedničkou, neboť při parametru  $a$  je simultánní funkce rovna 1 ( $= x_0$ ), a součtem, tedy  $\sum_{i=1}^n y_i = na + b \sum_{i=1}^n \frac{1}{x_i}$ .
- ✓ Druhá normální rovnice se obdrží vynásobením  $\frac{1}{x_i}$  a následným součtem, tedy  $\sum_{i=1}^n \frac{y_i}{x_i} = a \sum_{i=1}^n \frac{1}{x_i} + b \sum_{i=1}^n \frac{1}{x_i^2}$ .
- ✓ Podobným způsobem lze vytvořit soustavu normálních rovnic pro všechny ostatní regresní funkce aditivního tvaru:

$y'_i = a + b \log x_i$	$\sum_{i=1}^n y_i = na + b \sum_{i=1}^n \log x_i$ $\sum_{i=1}^n y_i \log x_i = a \sum_{i=1}^n \log x_i + b \sum_{i=1}^n \log^2 x_i$
$y'_i = a + bx_i + cx_i^2$	$\sum y_i = na + b \sum x_i + c \sum x_i^2$ $\sum x_i y_i = a \sum x_i + b \sum x_i^2 + c \sum x_i^3$ $\sum x_i^2 y_i = a \sum x_i^2 + b \sum x_i^3 + c \sum x_i^4$ <p><b>Polynommická regrese:</b></p> $\sum y_i = nb_0 + b_1 \sum x_i + \dots + b_p \sum x_i^p$ $\sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2 + \dots + b_p \sum x_i^{p+1}$ <p>.....</p> $\sum x_i^p y_i = b_0 \sum x_i^p + b_1 \sum x_i^{p+1} + \dots + b_p \sum x_i^{2p}$
$y'_i = a + bx_i^3 + c\sqrt{x_i} + \frac{d}{x_i^2}$	$\sum y_i = na + b \sum x_i^3 + c \sum \sqrt{x_i} + d \sum \frac{1}{x_i^2}$ $\sum x_i^3 y_i = a \sum x_i^3 + b \sum x_i^6 + c \sum x_i^3 \sqrt{x_i} + d \sum \frac{x_i^3}{x_i^2}$ $\sum \sqrt{x_i} y_i = a \sum \sqrt{x_i} + b \sum x_i^3 \sqrt{x_i} + c \sum x_i + d \sum \frac{\sqrt{x_i}}{x_i^2}$ $\sum \frac{y_i}{x_i^2} = a \sum \frac{1}{x_i^2} + b \sum \frac{x_i^3}{x_i^2} + c \sum \frac{\sqrt{x_i}}{x_i^2} + d \sum \frac{1}{x_i^4}$



### Exponenciální funkce:

- ✓ Odhad parametrů, které nejsou lineární v parametrech, neprovádíme MNČ přímo, protože její použití vede k soustavě nelineárních rovnic, z nichž zpravidla nedokážeme odhadnout přímo parametry ve formě vhodných výpočetních vzorců. Proto se při odhadu parametrů nelineárních regresních funkcí většinou postupuje tak, že se najde jejich vhodný počáteční odhad a postupným zlepšováním řešení nalezneme odhad s požadovanou přesností.
- ✓ Používá se tedy způsob, kdy určitou regresní funkci, která je nelineární z hlediska parametrů, převedeme pomocí linearizující transformace na funkci lineární v parametrech.
- ✓ Transformace spočívá v tom, že pomocí logaritmů, převrácením hodnot apod. dojdeme k takovému tvaru regresní funkce, že její parametry bude už možné odhadovat MNČ.

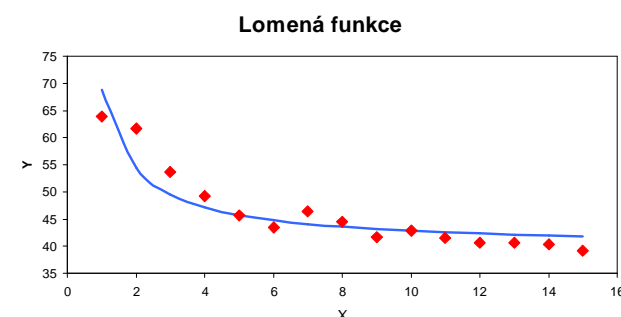
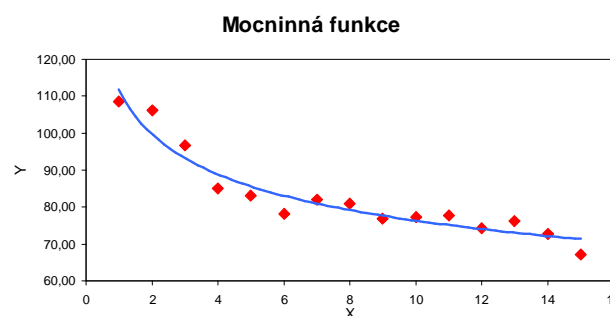
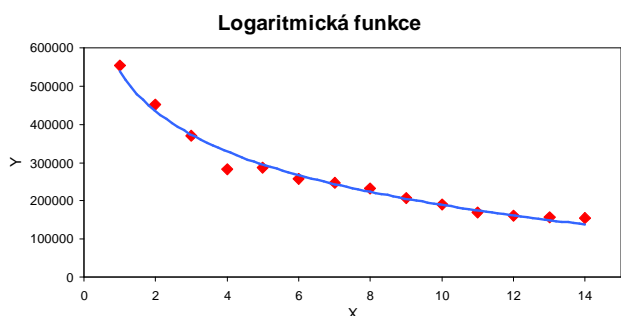
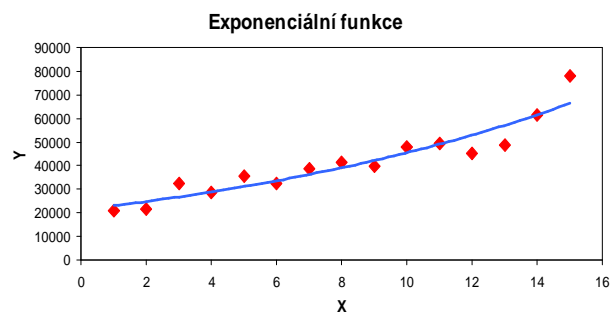
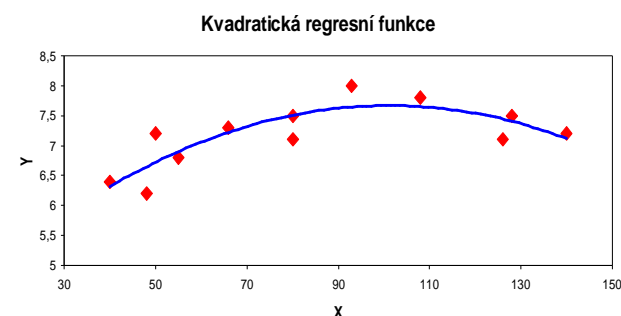
$$y'_i = a \cdot b^{x_i}$$

$$\log y'_i = \log a + x_i \log b$$

$$\sum \log y_i = n \log a + \log b \sum x_i$$

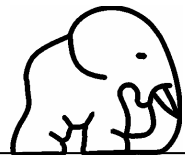
$$\sum x_i \log y_i = \log a \sum x_i + \log b \sum x_i^2$$

- ✓ Řešením jsou parametry ve tvaru  $\log a$  a  $\log b$ . Pokud chceme exponenciální funkci vyjádřit v původním tvaru, je potřeba provést odlogaritmování funkcí  $10^x$ .
- ✓ Funkce:



### Charakteristiky korelace u nelineární regrese:

- ✓ Pomáhají nám při posouzení kvality regresní funkce a ke zjištění síly závislosti.
- ✓ Posuzovaný vztah je tím silnější a regresní funkce tím lepší, čím více jsou empirické hodnoty vysvětlované proměnné soustředěné kolem odhadnuté regresní funkce, a naopak tím slabší, čím více jsou empirické hodnoty vzdáleny hodnotám vyrovnaným.



- ✓ Umožňuje také posoudit přesnost regresních odhadů – čím více se jednotlivé napozorované hodnoty soustředí kolem zvolené regresní čáry, tím je závislost těsnější a odhad přesnější.
- ✓ Při konstrukci míry ukazující na sílu závislosti vycházíme ze vztahu empirických a vyrovnaných hodnot, kdy pomocí těchto hodnot můžeme konstruovat **tři rozptyly s různou vypovídací schopností**:

- Rozptyl empirických (skutečně zjištěných) hodnot  $y$ :  $s_y^2 = \frac{1}{n} \sum (y_i - \bar{y})^2$ .
- Rozptyl vyrovnaných hodnot (teoretický rozptyl):  $s_{y'}^2 = \frac{1}{n} \sum (y'_i - \bar{y})^2$ .
- Rozptyl skutečně zjištěných hodnot kolem regresní čáry, tj. rozptyl empirických hodnot od hodnot vyrovnaných (reziduální rozptyl):  $s_{(y-y')}^2 = \frac{1}{n} \sum (y_i - y'_i - \overline{y - y'})^2 = \frac{1}{n} \sum (y_i - y'_i)^2$

- ✓ Lze dokázat, že při použití metody nejmenších čtverců mezi uvedenými rozptyly platí vztah  $s_y^2 = s_{y'}^2 + s_{(y-y')}^2$ .
- ✓ Rozptyl empirických hodnot lze tedy rozložit na rozptyl vyrovnaných hodnot a rozptyl reziduálních hodnot.
- ✓ **Podíl složek na empirickém rozptylu**:

- Teoretický rozptyl  $s_{y'}^2 = 0$ , takže  $s_y^2 = s_{(y-y')}^2$ . Jde o krajní případ, kdy je  $y'_i$  nezávislé na  $x_i$ , kdy jde vlastně o regresní přímku rovnoběžnou s osou  $x$ . v daném případě jde o nezávislost.
- Reziduální rozptyl  $s_{(y-y')}^2 = 0$ , takže  $s_y^2 = s_{y'}^2$ . Druhý krajní případ, kdy je každé  $y'_i$  shodné s  $y_i$ . Všechna empirická pozorování vyhovují teoretickým hodnotám na regresní přímce. Jde o pevnou závislost.
- Teoretický rozptyl  $s_{y'}^2 \neq 0$  a  $s_{(y-y')}^2 \neq 0$ , takže  $s_y^2 = s_{y'}^2 + s_{(y-y')}^2$ . V daném případě jde o volnou závislost.

- ✓ Závislost proměnné  $Y$  na proměnné  $X$  bude zřejmě tím silnější, čím větší bude podíl rozptylu vyrovnaných hodnot na celkovém rozptylu, a tím slabší, čím bude podíl tohoto rozptylu menší. Sílu závislosti je tedy možné

měřit poměrem  $I_{yx}^2 = \frac{s_{y'}^2}{s_y^2}$ .

- ✓ Tento poměr se nazývá **index determinace**. V případě funkční závislosti nabude hodnoty 1, v případě nezávislosti hodnoty 0. Čím více se bude blížit jedné, tím se závislost považuje za silnější, a tedy dobře vystiženou zvolenou regresní funkcí.
- ✓ Index determinace v procentickém vyjádření udává, jaké procento rozptýlení empirických hodnot závisle proměnné je důsledkem rozptylu teoretických hodnot závisle proměnné odhadnutých na základě příslušné regresní funkce.
- ✓ Kvalitu regresní funkce a intenzitu závislosti můžeme hodnotit podle toho, jak se podílí na rozptylu skutečně zjištěných hodnot rozptyl vyrovnaných hodnot, příp. rozptyl odchylek kolem regresní čáry.
- ✓ Je třeba mít na zřeteli, že velikost indexu determinace je zcela ovlivněna tím, zda se podařilo nalézt vhodný typ regresní funkce pro popis dané závislosti. Nízká hodnota indexu determinace nemusí ještě znamenat nízký stupeň závislosti mezi proměnnými, ale může to signalizovat chybnou volbu regresní funkce.

- ✓ Index determinace lze také konstruovat nepřímou, tj. ve tvaru  $I_{yx}^2 = \frac{s_{y'}^2}{s_y^2} = 1 - \frac{s_{(y-y')}^2}{s_y^2}$ .

- ✓ K měření těsnosti závislosti se v praxi častěji používá odmocnina indexu determinace, která se nazývá **index**

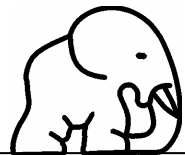
**korelace**:  $I_{yx} = \sqrt{\frac{s_{y'}^2}{s_y^2}}$ . Index korelace poskytuje stejné informace o těsnosti závislosti jako index determinace,

jinak však má menší vypovídací schopnost.

- ✓ Dosadíme-li do vzorce indexu korelace za oba rozptyly, dostaneme výpočetní vzorec ve formě

$$I_{yx} = \sqrt{\frac{\sum (y'_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}}.$$

- ✓ Index korelace se používá k měření těsnosti závislosti pro libovolnou regresní funkci, jejíž parametry byly odhadnuty metodou nejmenších čtverců.



- ✓ Pro dosazení do uvedených vzorců indexu korelace je potřebné vypočítat pro každou hodnotu  $x_i$  podle konkrétní regresní funkce teoretické hodnoty  $y'_i$  a pak teprve počítat příslušné součty čtverců pro teoretický či lépe reziduální rozptyl.

- ✓ Snadnější a výhodnější je následující postup výpočtu: 
$$I_{yx} = \sqrt{\frac{s_{y'}^2}{s_y^2}} = \sqrt{\frac{\sum (y'_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}} = \sqrt{\frac{\sum y_i'^2 - \frac{1}{n}(\sum y_i')^2}{\sum y_i^2 - \frac{1}{n}(\sum y_i)^2}},$$

přičemž  $\sum y_i'^2 = \sum y_i y'_i$ . Např. v případě kvadratické funkce lze psát:  

$$\sum y_i'^2 = \sum y_i (a + bx_i + cx_i^2) = a \sum y_i + b \sum x_i y_i + c \sum x_i^2 y_i.$$

### Korelační poměr:

- ✓ Pokud nelze z jakýchkoliv důvodů určit konkrétní tvar vyrovnávající regresní funkce, používá se k určení těsnosti závislosti míry, která se nazývá **korelační poměr**. V určitém smyslu je to obecnější míra závislosti než index či koeficient korelace, protože na rozdíl od nich nezávisí na tvaru regresní funkce.
- ✓ Z definice **korelační závislosti** vyplývá, že se změnami hodnot vysvětlující proměnné se systematicky mění podmíněné průměry závisle proměnné. V takovém případě se v podmíněných průměrech projevuje určitá variabilita, kterou lze měřit **rozptylem podmíněných průměrů**  $s_{\bar{y}}^2$ .
- ✓ Vliv ostatních činitelů na závisle proměnnou se pak projevuje tím, že v podmíněných rozděleních závisle proměnné dochází ke kolísání jednotlivých hodnot závisle proměnné okolo podmíněných průměrů. Toto kolísání se měří průměrem z podmíněných rozptylů  $\overline{s^2}$ .
- ✓ Závislost Y na X lze tedy zřejmě považovat za tím silnější, čím větší je variabilita podmíněných průměrů ve srovnání s variabilitou hodnot v podmíněných rozděleních.
- ✓ Protože platí  $s_y^2 = s_{\bar{y}}^2 + \overline{s^2}$ , je zřejmé, že lze tuto míru těsnosti závislosti konstruovat jako poměr

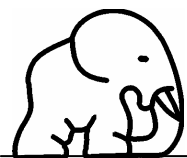
$$\frac{s_{\bar{y}}^2}{s_y^2} = \frac{s_y^2 - \overline{s^2}}{s_y^2} = 1 - \frac{\overline{s^2}}{s_y^2}.$$

- ✓ Tento poměr udávaný v % se nazývá **poměr determinace** a udává, jaké % rozptylu závisle proměnné lze vysvětlit vlivem nezávisle proměnné X. Doplněk do 100 % pak udává vliv blíže nespecifikovaných činitelů. Čím více se blíží poměr determinace jedné, tím je závislost proměnné Y na proměnné X silnější.
- ✓ V případě, že variabilita hodnot v podmíněných rozděleních je nulová, je poměr determinace roven 1 a jde tedy o úplnou závislost mezi oběma proměnnými. Naopak v případě, že jsou všechny podmíněné průměry stejné, je poměr determinace nulový a jde tedy o korelační nezávislost Y na X.
- ✓ K měření těsnosti závislosti se pak používá odmocnina z poměru determinace, která se nazývá **korelační**

**poměr**  $\eta_{yx} = \sqrt{\frac{s_{\bar{y}}^2}{s_y^2}}$ . Korelační poměr lze také vypočítat nepřímou ve tvaru  $\eta_{yx} = \sqrt{\frac{s_y^2 - \overline{s^2}}{s_y^2}} = \sqrt{1 - \frac{\overline{s^2}}{s_y^2}}.$

- ✓ Za předpokladu, že závislost mezi proměnnými byla zkoumána na dostatečně velkém počtu pozorování, kdy podmíněné průměry závisle proměnné Y nemohou být výrazněji ovlivňovány nahodilými vlivy, lze pak pozorováním velikosti korelačního poměru a indexu korelace (příp. koeficientu) usuzovat na vhodnost použité funkce. Čím více se budou hodnoty obou měř k sobě přibližovat, tím se bude použitá regresní funkce považovat za vhodnější zobrazení dané závislosti.
- ✓ Maticový způsob stanovení parametrů nelineárních funkcí  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ .

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$



Kvadratická funkce:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

Hyperbola (lomená):

$$\mathbf{X} = \begin{bmatrix} 1 & \frac{1}{x_1} \\ 1 & \frac{1}{x_2} \\ \vdots & \vdots \\ 1 & \frac{1}{x_n} \end{bmatrix}$$

Odmocninná funkce:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & \sqrt{x_1} \\ 1 & x_2 & \sqrt{x_2} \\ \vdots & \vdots & \vdots \\ 1 & x_n & \sqrt{x_n} \end{bmatrix}$$

Logaritmická funkce:

$$\mathbf{X} = \begin{bmatrix} 1 & \log x_1 \\ 1 & \log x_2 \\ \vdots & \vdots \\ 1 & \log x_n \end{bmatrix}$$

Exponenciální funkce:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \log y_1 \\ \log y_2 \\ \vdots \\ \log y_n \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} \log a \\ \log b \end{bmatrix}$$

✓ Maticově lze stanovit i hodnotu korelačního indexu: 
$$I = \sqrt{\frac{\mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{1}{n} \sum (y_i)^2}{\mathbf{y}'\mathbf{y} - \frac{1}{n} \sum (y_i)^2}}$$

**Statistická analýza v nelineárním modelu:**✓ **Intervalové odhady parametrů:**

- Bodové odhady  $\mathbf{b}$  regresních parametrů  $\beta$  jsou ze statistického hlediska bezcenné, protože nic neuvádějí o tom, v jakých mezích lze očekávat výskyt skutečných hodnot  $\beta$ . Odhady  $\mathbf{b}$  jsou náhodné veličiny určené na základě výběru dat o velikosti  $n$ .
- U nelineárních regresních modelů se při konstrukci intervalů spolehlivosti používá převážně linearizace, která je však použitelná pouze v případech, kdy model není silně lineární a míry nelinearity, asymetrie a vychýlení odhadů jsou malé.
- Postup pro stanovení intervalových odhadů jednotlivých parametrů je analogický intervalovému odhadu regresního koeficientu v případě lineárních modelů. Zanedbává se zde vliv ostatních parametrů. Protože jsou však většinou prvky vektoru  $\mathbf{b}$  (vektor regresních parametrů) vzájemně korelované, bývají intervaly takto stanovené podceněné, tj. příliš úzké.
- Pro nelineární modely je možné také stanovit intervaly spolehlivosti predikce, vyčíslené v celém rozmezí hodnot nezávisle proměnné, tzn. stanovit pásy spolehlivosti.

✓ **Testy hypotéz o odhadech parametrů:**

- Testování hypotéz souvisí úzce s konstrukcí oblastí spolehlivosti. Pokud parametry  $\beta_0$  leží v 95% oblasti spolehlivosti kolem  $\mathbf{b}$ , lze na hladině významnosti  $\alpha = 0,05$  považovat rozdíly  $(\beta - \beta_0)$  za statisticky nevýznamné.
- Samotné testy pak lze konstruovat stejně jako v lineárním modelu (za předpokladu alespoň přibližné normality odhadu metodou nejmenších čtverců).
- Individuální testy o nulových hodnotách parametrů však nemají v nelineární regresní analýze dobrý význam, protože známe-li vhodnou regresní funkci, jsou případné zjednodušené modely těžko interpretovatelné. V jiných případech je třeba testovat jiné hodnoty parametrů než nulové.

✓ **Těsnost proložení regresní křivky:**

- U lineárních regresních modelů slouží analýza reziduí k ověřování některých předpokladů o chybách  $\epsilon$ , u nelineárních modelů pak především k posouzení dosažené těsnosti proložení vypočtené regresní křivky danými experimentálními body.
- Analýzou vlivných bodů se identifikují body, které silně ovlivňují odhadované regresní parametry v modelu, což umožňuje určit vybočující pozorování nebo extrém.

✓ **Statistická analýza reziduí:**

- Pro aditivní modely měření a užívanou NMČ jsou rezidua definována vztahem  $e_i = y_i - f(x_i, \mathbf{b})$ .
- K analýze reziduí se užívá jednak názorného grafického zobrazení vektoru reziduí a jednak numerické analýzy směřující ke statistickému testování.



- ✓ **Grafická analýza reziduí** – grafickou (předběžnou) analýzou reziduí spočívající v prostém zobrazení vektoru reziduí, lze snadno odhalit:
- Odlehlé (extrémní) hodnoty v souboru reziduí.
  - Trend v reziduích.
  - Nedostatečné střídání znaménka u reziduí.
  - Chybný model nebo vzájemnou závislost reziduí.
  - Heteroskedasticitu (nekonstantnost rozptylu) závisle proměnné veličiny Y.
  - Náhlou změnu podmínek při měření hodnoty y.
- ✓ **Statistická (numerická) analýza reziduí:**
- Analýza reziduí je hlavní diagnostickou pomůckou při hledání a rozlišení regresního modelu a navíc těsnost dosaženého proložení experimentálními body je mírou věrohodnosti nalezených odhadů.
  - Mezi nejčastěji užívané statistiky patří především střední hodnota reziduí  $E(e)$ , která by se měla rovnat nule, dále průměrné reziduum, směrodatná odchylka střední hodnoty reziduí a konečně koeficient šikmosti a koeficient špičatosti reziduí.
  - Pro normální rozdělení reziduí by se měl koeficient šikmosti rovnat nule a koeficient špičatosti třem.
  - Pozn. Diagnostické metody nejsou vždy spolehlivé, protože rezidua nemají nulovou střední hodnotu, jsou vychýlená, jsou přibližně lineární kombinací chyb a navíc závisejí na skutečných hodnotách parametrů  $\beta$  (které jsou uživateli neznámé).

✓ **Příklad:**

Proměnná X	3	5	6	5	8	3	7	4	6	5	7	2
Proměnná Y	6	2,5	2	3	1,5	4,5	2	5,5	3	3,5	2,5	7

Comparison of Alternative Models

Model	Correlation	R-Squared
Logarithmic-X	-0,9341	87,25%
Exponential	-0,9315	86,77%
Square root-X	-0,9287	86,25%
Square root-Y	-0,9269	85,91%
Multiplicative	-0,9186	84,38%
Reciprocal-X	0,9145	83,63%
Linear	-0,9141	83,56%
Reciprocal-Y	0,9092	82,67%
S-curve	0,8680	75,34%
Double reciprocal	-0,7856	61,71%
Logistic	<no fit>	
Log probit	<no fit>	

Regression Analysis - Logarithmic-X model:  $Y = a + b \cdot \ln(X)$

Dependent variable: promenna Y

Independent variable: promenna X

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	9,78235	0,773336	12,6495	0,0000
Slope	-3,98656	0,481886	-8,27283	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	30,029	1	30,029	68,44	0,0000
Residual	4,38766	10	0,438766		
Total (Corr.)	34,4167	11			

Correlation Coefficient = -0,934084

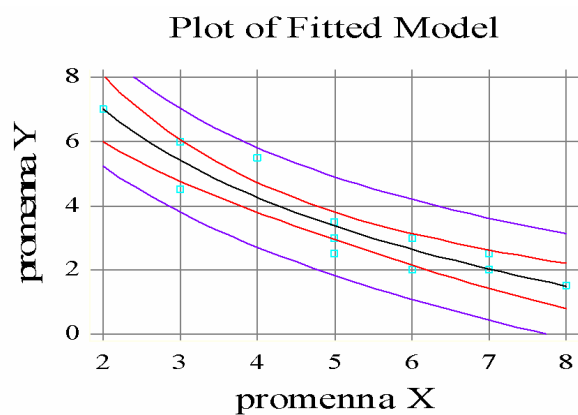
R-squared = 87,2513 percent

Standard Error of Est. = 0,662394



T Statistics – testové kritérium

P-Value – hladina významnosti



Černá čára – regresní funkce

Červené čáry – intervalový odhad regresní funkce

Fialová čára – pás spolehlivosti.

