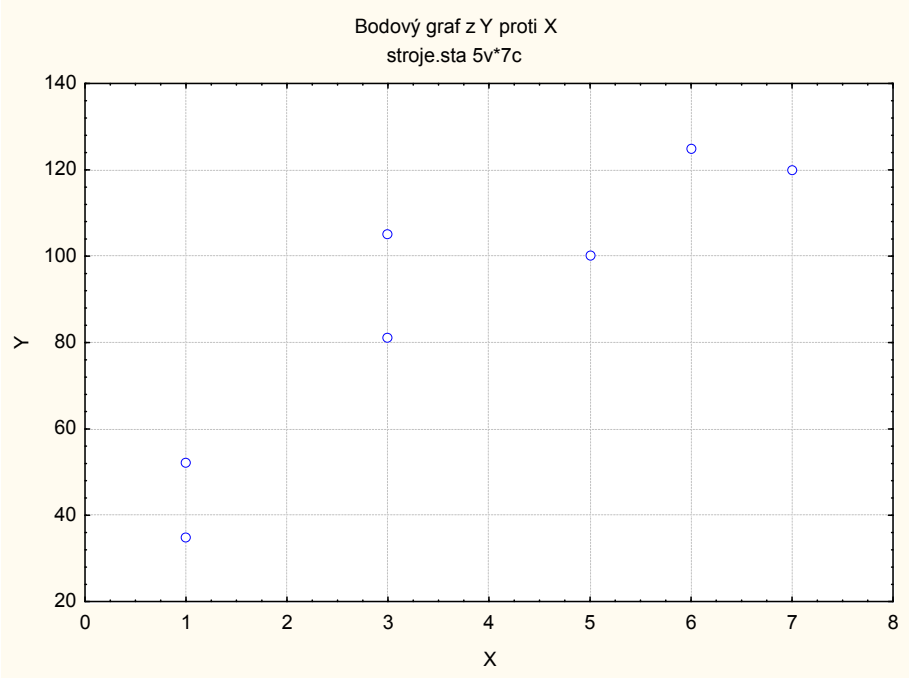
Téma č. 8.: Jednoduchá lineární regrese II

Příklad 1.: U sedmi náhodně vybraných strojů v určitém podniku se zjišťovalo stáří stroje v letech (proměnná X) a týdenní náklady v Kč na údržbu stroje (proměnná Y). Data:
(1,35), (1,52), (3,81), (3,105), (5,100), (6,125), (7, 120)
Data znázorněte graficky. Vyzkoušejte následující čtyři modely:
$y = \beta_0 + \beta_1 x$, $y = \beta_0 + \beta_1 \sqrt{x}$, $y = \beta_0 + \beta_1 \log_{10} x$, $y = \beta_0 + \beta_1 1/x$. Vyberte ten model, který poskytuje nejvyšší index determinace. Určete regresní odhad týdenních nákladů pro stroj starý čtyři roky.

Řešení:
Nejprve data znázorníme graficky:



Datový soubor s proměnnými X a Y doplníme o proměnné SQRTX, LOGX a INVX. Hodnoty proměnné SQRTX resp. LOGX resp. INVX získáme tak, že do Dlouhého jména napíšeme =sqrt(x) resp. =Log10(x) resp. =1/x.

	1 X	2 Y	3 SQRTX	4 LOGX	5 INVX
1	1	35	1	0	1
2	1	52	1	0	1
3	3	81	1,732051	0,477121	0,333333
4	3	105	1,732051	0,477121	0,333333
5	5	100	2,236068	0,69897	0,2
6	6	125	2,44949	0,778151	0,166667
7	7	120	2,645751	0,845098	0,142857

Regresní analýzu provedeme tak, že roli nezávisle proměnné bude hrát proměnná X, pak SQRTX, LOGX a nakonec INVX.

Model s proměnnou X:

N=7	Výsledky regrese se závislou proměnnou : Y (stroje.sta) R= ,91004028 R2= ,82817331 Upravené R2= ,79380797 F(1,5)=24,099 p<,00444 Směrod. chyba odhadu : 15,487					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			39,44444	11,54341	3,417054	0,018898
X	0,910040	0,185379	13,14957	2,67862	4,909082	0,004439

Model s proměnnou SQRTX:

N=7	Výsledky regrese se závislou proměnnou : Y (stroje.sta) R= ,93923698 R2= ,88216611 Upravené R2= ,85859933 F(1,5)=37,433 p<,00169 Směrod. chyba odhadu : 12,825					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			-0,47736	15,29638	-0,031207	0,976312
SQRTX	0,939237	0,153515	48,55972	7,93690	6,118220	0,001691

Model s proměnnou LOGX:

N=7	Výsledky regrese se závislou proměnnou : Y (stroje.sta) R= ,95349135 R2= ,90914576 Upravené R2= ,89097491 F(1,5)=50,033 p<,00087 Směrod. chyba odhadu : 11,262					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			44,64571	7,49541	5,956407	0,001907
LOGX	0,953491	0,134799	93,23472	13,18100	7,073415	0,000874

Model s proměnnou INVX

N=7	Výsledky regrese se závislou proměnnou : Y (stroje.sta) R= ,94282234 R2= ,88891396 Upravené R2= ,86669676 F(1,5)=40,010 p<,00146 Směrod. chyba odhadu : 12,452					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			126,6192	7,67327	16,50134	0,000015
INVX	-0,942822	0,149054	-84,4832	13,35627	-6,32536	0,001456

Vidíme, že nejvyšší index determinace poskytuje model s proměnnou LOGX: $ID^2 = 90,9\%$. Má také nejmenší směrodatnou chybu odhadu.

Určíme regresní odhad týdenních nákladů pro stroj starý 4 roky v modelu s nezávisle proměnnou LOGX.Nejprve vypočteme log(4) = 0,602
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 0,602 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

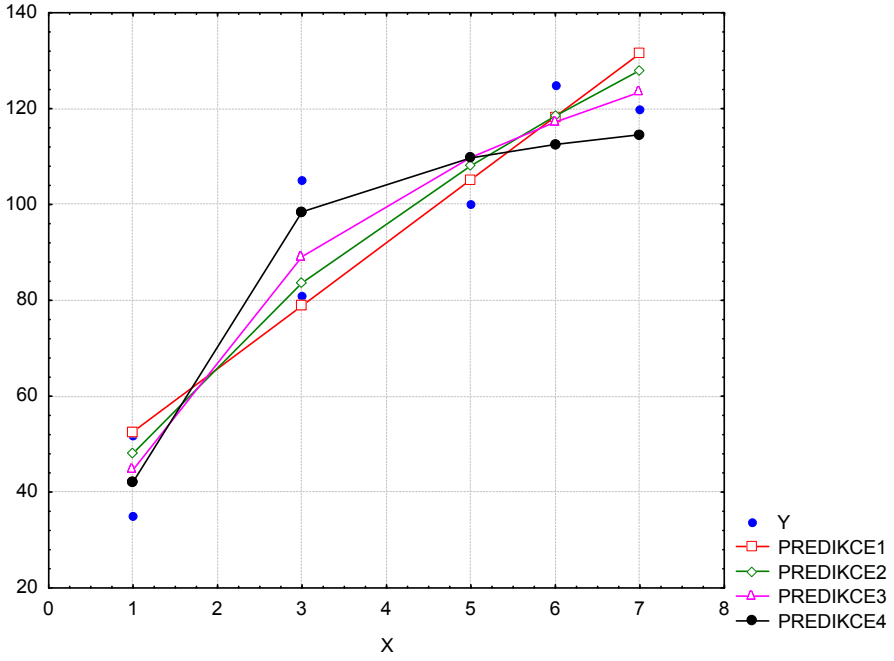
Proměnná	Předpovězené hodnoty (stroje.sta) proměnné: Y		
	B-váž.	Hodnota	B-váž. * Hodnot
LOGX	93,23472	0,602000	56,1273
Abs. člen			44,6457
Předpověď			100,7730
-95,0%LS			88,9277
+95,0%LS			112,6184

Bodový odhad je 100,77 Kč. Vidíme, že s pravděpodobností aspoň 0,95 budou týdenní náklady na údržbu stroje starého 4 roky činit minimálně 88,93 Kč a maximálně 112,62 Kč.

Nakonec znázorníme data se všemi čtyřmi regresními křivkami. K původnímu datovému souboru s proměnnými X,Y přidáme 4 nové proměnné PREDIKCE1, …, PREDIKCE4. Do Dlouhých jmen těchto proměnných napíšeme příslušné regresní rovnice, tj.
=39,44444+13,14957*x
=-0,4776+48,55972*sqrtx
=44,64571+93,23472*logx
=126,6192-84,4832*invx

	1 X	2 Y	3 SQRTX	4 LOGX	5 INVX	6 PREDIKCE1	7 PREDIKCE2	8 PREDIKCE3	9 PREDIKCE4
1	1	35	1	0	1	52,59401	48,08212	44,64571	42,136
2	1	52	1	0	1	52,59401	48,08212	44,64571	42,136
3	3	81	1,732051	0,477121	0,333333	78,89315	83,6303022	89,1299766	98,4581333
4	3	105	1,732051	0,477121	0,333333	78,89315	83,6303022	89,1299766	98,4581333
5	5	100	2,236068	0,69897	0,2	105,19229	108,105235	109,813983	109,72256
6	6	125	2,44949	0,778151	0,166667	118,34186	118,468936	117,196424	112,538667
7	7	120	2,645751	0,845098	0,142857	131,49143	127,999343	123,438189	114,550171

Obrázek vytvoříme pomocí vícenásobného bodového grafu.



Příklad 2.: Na podzim byla uskladněna zimní jablka. Po čase bylo vždy odebráno několik kusů a u každého byla posuzována chuť, tvrdost, kvalita slupky a celkový vzhled jablka. Vyšší počet bodů odpovídá lepší kvalitě ovoce. Doba, která uplynula od uskladnění, je nezávisle proměnná veličina X, počet bodů závisle proměnná veličina Y.

X	Y
0	5 6 4 5
2	9 7 8
4	9 8 10 10 8
6	8 5 7 4 6
8	3 1 2

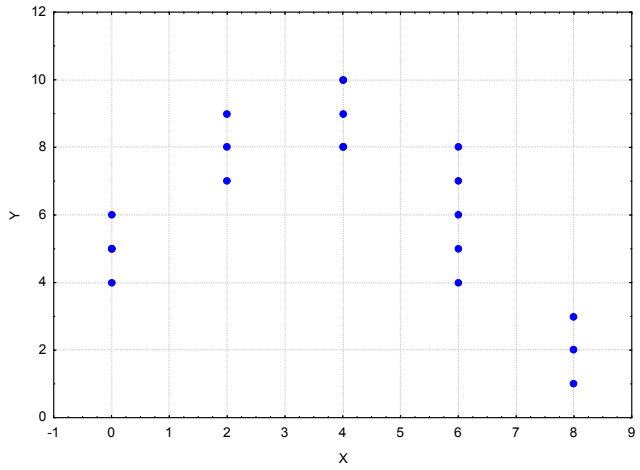
Na hladině významnosti 0,05 testujte hypotézu, že regresní přímka je vhodný model závislosti Y na X.

Řešení v systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 20 případy:

	1 X	2 Y
1	0	5
2	0	6
3	0	4
4	0	5
5	2	9
6	2	7
7	2	8
8	4	9
9	4	8
10	4	10
11	4	10
12	4	8
13	6	8
14	6	5
15	6	7
16	6	4
17	6	6
18	8	3
19	8	1
20	8	2

Data znázorníme graficky:



Je zřejmé, že přímka nebude vhodným regresním modelem.

Odhadneme parametry regresní přímky:

N=20	Výsledky regrese se závislou proměnnou : Y (jablka.sta) R= ,32440757 R2= ,10524027 Upravené R2= ,05553140 F(1,18)=2,1171 p<,16288 Směrod. chyba odhadu : 2,5200					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(18)	Úroveň p
Abs.člen			7,472222	1,011487	7,38737	0,000001
X	-0,324408	0,222955	-0,305556	0,209998	-1,45504	0,162877

Sestavíme tabulku ANOVA:
Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

Efekt	Analýza rozptylu (jablka.sta)				
	Součet čtverců	sv	Průměr čtverců	F	Úroveň p
Regres.	13,4444	1	13,44444	2,117132	0,162877
Rezid.	114,3056	18	6,35031		
Celk.	127,7500				

Vidíme, že $S_R = 13{,}4444$, $S_T = 127{,}75$

Provedeme jednofaktorovou analýzu rozptylu, abychom získali skupinový součet čtverců:
Statistiky – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné – Závislé – Y, Grupovací - X – OK – OK – Analýza rozptylu.

Proměnná	Analýza rozptylu (jablka.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
Y	107,7500	4	26,93750	20,00000	15	1,333333	20,20313	0,000007

Zde najdeme $S_A = 107{,}75$.

Vypočteme testovou statistiku $F = \frac{(107{,}75 - 13{,}4444)/(5-2)}{(127{,}75 - 107{,}75)/(20-5)} = \frac{31{,}4352}{1{,}3333} = 23{,}576$ a najdeme kritický obor $W = <F_{0{,}95}(3,15), \infty) = <4{,}1528, \infty)$. Jelikož $F \in W$, zamítáme na hladině významnosti 0,05 hypotézu, že přímka je vhodným regresním modelem závislosti kvality jablek na době uskladnění.