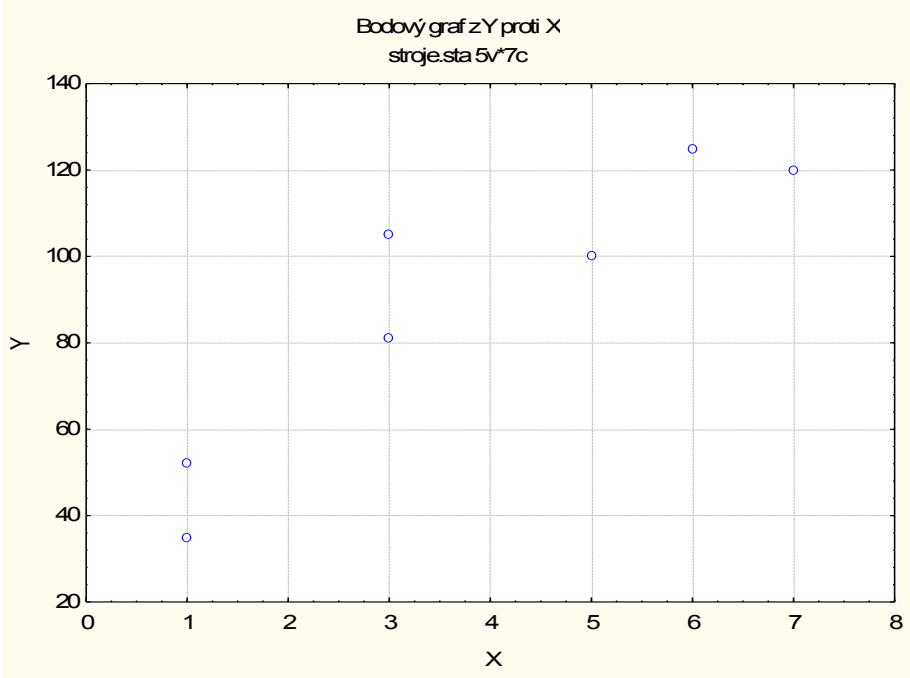
Cvičení 4.: Jednoduchá lineární regrese

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:

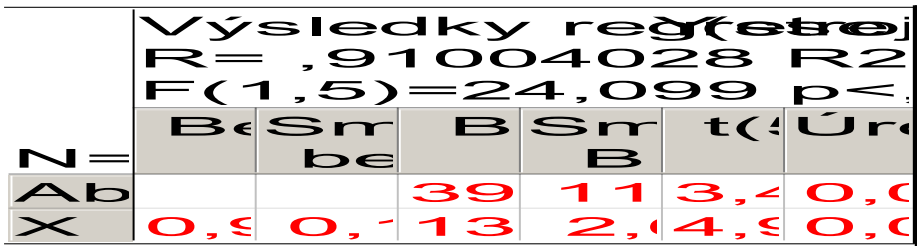


Datový soubor s proměnnými X a Y doplníme o proměnné SQRTX, LOGX a INVX. Hodnoty proměnné SQRTX resp. LOGX resp. INVX získáme tak, že do Dlouhého jména napíšeme =sqrt(x) resp. =Log10(x) resp. =1/x.

	1 X	2 Y	3 SQR	4 LOG	5 INV
1	1	35	1	0	1
2	1	52	1	0	1
3	3	81	1,73	0,47	0,33
4	3	105	1,73	0,47	0,33
5	5	100	2,23	0,69	0
6	6	125	2,44	0,77	0,16
7	7	120	2,64	0,84	0,14

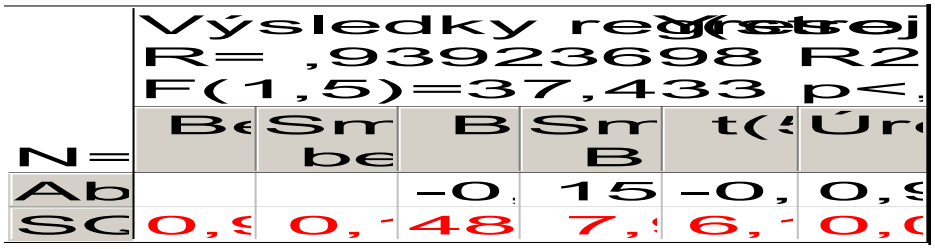
Regresní analýzu provedeme tak, že roli nezávisle proměnné bude hrát proměnná X, pak SQRTX, LOGX a nakonec INVX.

Model s proměnnou X:



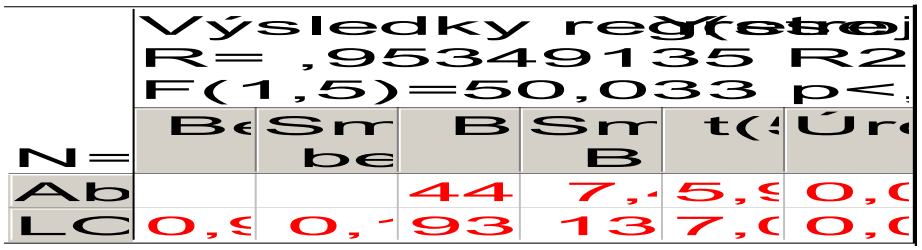
N =	Výsledky regrese R= ,91004028 R2 F(1,5)=24,099 p<					
	Be	Sm be	B	Sm B	t(	Úr
Ab			39	11	3,4	0,0
X	0,9	0,1	13	2,(	4,9	0,0

Model s proměnnou SQRTX:



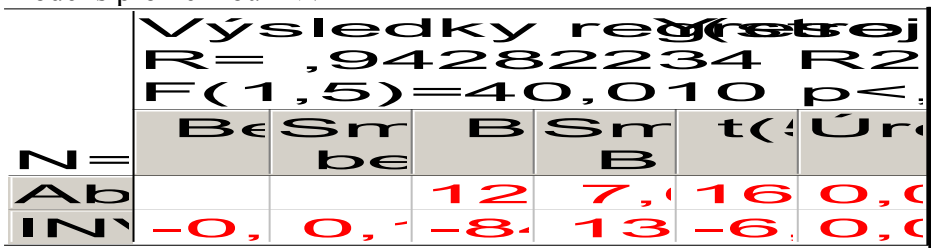
N =	Výsledky regrese R= ,93923698 R2 F(1,5)=37,433 p<					
	Be	Sm be	B	Sm B	t(	Úr
Ab			-0,	15	-0,	0,9
SC	0,9	0,1	48	7,9	6,1	0,0

Model s proměnnou LOGX:



N =	Výsledky regrese R= ,95349135 R2 F(1,5)=50,033 p<					
	Be	Sm be	B	Sm B	t(	Úr
Ab			44	7,4	5,9	0,0
LC	0,9	0,1	93	13	7,0	0,0

Model s proměnnou INVX



N =	Výsledky regrese R= ,94282234 R2 F(1,5)=40,010 p<					
	Be	Sm be	B	Sm B	t(	Úr
Ab			12	7,(	16	0,0
IN	-0,	0,1	-8,	13	-6,	0,0

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.

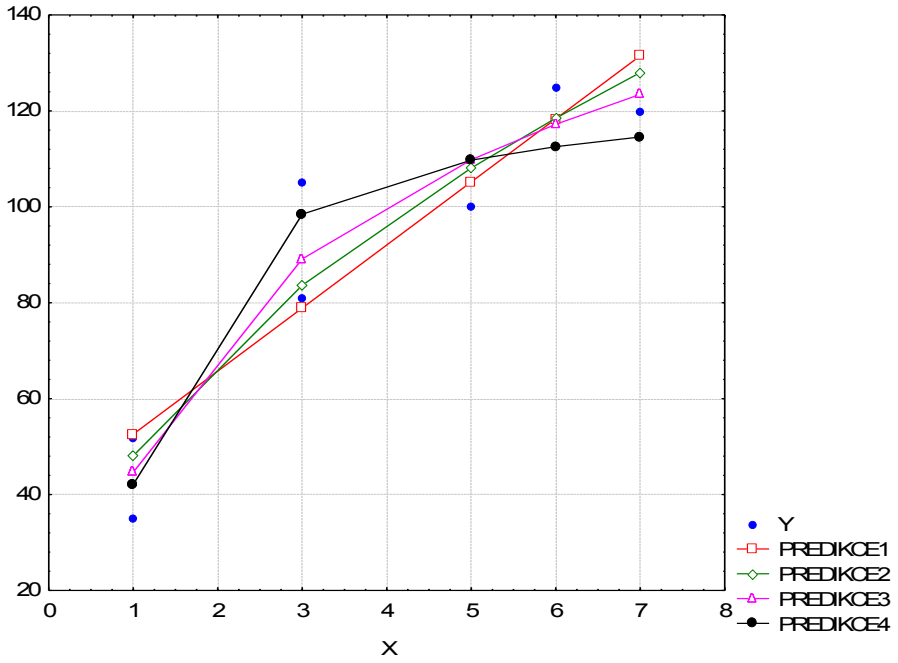
Proměnná	Předpovězené hodnoty (stroje.sta) proměnné: Y		
	B-váž.	Hodnota	B-váž. * Hodnot
LOGX	93,23472	0,602000	56,1273
Abs. člen			44,6457
Předpověď			100,7730
-95,0%LS			88,9277
+95,0%LS			112,6184

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 S	4 L	5 I	6 PR	7 PR	8 PR	9 PR
						52	48	44	4
						52	48	44	4
			1,	0,	0,	78	83	89	98
			1,	0,	0,	78	83	89	98
			2,	0,		10	10	10	10
			2,	0,	0,	11	11	11	11
			2,	0,	0,	13	12	12	11

Obrázek vytvoříme pomocí vícenásobného bodového grafu.



Příklad 2.: V regresním modelu paraboly, který znázorňuje závislost spotřeby benzínu na rychlosti automobilu Škoda 120 (datový soubor spotreba_benzinu.sta):

a) Určete 95 % intervaly spolehlivosti pro regresní parametry.

Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p dvě nové proměnné dm (pro dolní meze 95% intervalů spolehlivosti pro regresní parametry) a hm (pro horní meze 95% intervalů spolehlivosti pro regresní parametry). Do Dlouhého jména proměnné dm resp. hm napíšeme: =v3-v4*VStudent(0,975;5) resp. =v3+v4*VStudent(0,975;5)

	Výsledky regrese se závislou proměnnou : Y (spotreba) R= ,98403165 R2= ,96831829 Upravené R2= ,95564561 F(2,5)=76,410 p<,00018 Směrod. chyba odhadu : ,22973							
N=8	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p	dm =v3-v4*	hm =v3+v4
Abs.člen			9,751786	0,945689	10,31183	0,000148	7,320815	12,18276
X	-3,38045	0,602292	-0,150536	0,026821	-5,61264	0,002483	-0,21948	-0,08159
Xkv	4,22756	0,602292	0,001244	0,000177	7,01912	0,000905	0,000788	0,0017

Vidíme, že
$7{,}320815 < \beta_0 < 12{,}18276$ s pravděpodobností aspoň 0,95,
$-0{,}21948 < \beta_1 < -0{,}08159$ s pravděpodobností aspoň 0,95,
$0{,}000788 < \beta_2 < 0{,}0017$ s pravděpodobností aspoň 0,95

b) Na hladině významnosti 0,05 proveďte celkový F-test.
Testovou statistiku F-testu a odpovídající p-hodnotu najdeme v záhlaví výstupní tabulky regrese. Zde F = 76,41, p-hodnota < 0,00018, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku. (Výsledky F-testu jsou rovněž uvedeny v tabulce ANOVA.)

c) Na hladině významnosti 0,05 proveďte dílčí t-testy a vypočtěte relativní chyby odhadů regresních parametrů.
Výsledky dílčích t-testů jsou uvedeny ve výstupní tabulce regrese.
Testová statistika pro test hypotézy H_0: $\beta_0 = 0$ je 10,31183, p-hodnota je 0,000148. Hypotézu o nevýznamnosti parametru β_0 tedy zamítáme na hladině významnosti 0,05.
Testová statistika pro test hypotézy H_0: $\beta_1 = 0$ je -5,61264, p-hodnota je 0,002483. Hypotézu o nevýznamnosti parametru β_1 tedy zamítáme na hladině významnosti 0,05.
Testová statistika pro test hypotézy H_0: $\beta_2 = 0$ je 7,01912, p-hodnota je 0,000905. Hypotézu o nevýznamnosti parametru β_2 tedy zamítáme na hladině významnosti 0,05.
K upravené výstupní tabulce s mezemi intervalů spolehlivosti přidáme proměnnou chyba. Do jejího Dlouhého jména napíšeme
=100*abs(0,5*(hm-dm)/v3)

	Výsledky regrese se závislou proměnnou : Y (spotreba) R= ,98403165 R2= ,96831829 Upravené R2= ,95564561 F(2,5)=76,410 p<,00018 Směrod. chyba odhadu : ,22973								
N=8	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p	dm =v3-v4*	hm =v3+v4	chyba =100*a
Abs.člen			9,751786	0,945689	10,31183	0,000148	7,320815	12,18276	24,92847
X	-3,38045	0,602292	-0,150536	0,026821	-5,61264	0,002483	-0,21948	-0,08159	45,79987
Xkv	4,22756	0,602292	0,001244	0,000177	7,01912	0,000905	0,000788	0,0017	36,62259

Vidíme, že chyby odhadů jsou velké, v řádu desítek procent.

d) Určete regresní odhad spotřeby benzínu při rychlosti 80 km/h.
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi - Předpovědi závisle proměnné X: 80, Xkv 6400 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď: 5,6708

e) Vypočtěte střední absolutní procentuální chybu predikce (MAPE)

Ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua/předpoklady/předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat X, Y – OK. Ve vzniklé tabulce odstraníme proměnné 5 – 10, přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme
=100*abs(v4/v2)
Pak spočteme průměr této proměnné a zjistíme, že MAPE = 2,15%.