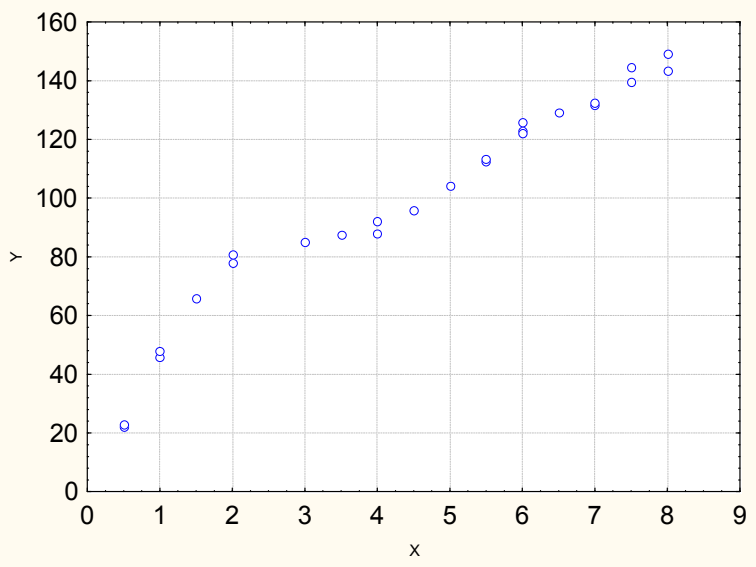
Téma č. 4.: Jednoduchá lineární regrese II

Příklad 1.: U 25 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 X	2 Y
1	0,5	22
2	0,5	23
3	1	46
4	1	48
5	1,5	66
6	2	78
7	2	81
8	3	85
9	3,5	87
10	4	88
11	4	92
12	4,5	96
13	5	104
14	5,5	112
15	5,5	113
16	6	123
17	6	122
18	6	126
19	6,5	129
20	7	132
21	7	133
22	7,5	140
23	7,5	145
24	8	143
25	8	149

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. Pro tento model proveďte reziduální analýzu. V případě, že rezidua vykazují autokorelaci, pokuste se o její eliminaci a v upraveném modelu 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	0,5	22	0,707107	-0,30103	2
2	0,5	23	0,707107	-0,30103	2
3	1	46	1	0	1
4	1	48	1	0	1
5	1,5	66	1,224745	0,176091	0,666667
6	2	78	1,414214	0,30103	0,5
7	2	81	1,414214	0,30103	0,5
8	3	85	1,732051	0,477121	0,333333
9	3,5	87	1,870829	0,544068	0,285714
10	4	88	2	0,60206	0,25
11	4	92	2	0,60206	0,25
12	4,5	96	2,12132	0,653213	0,222222
13	5	104	2,236068	0,69897	0,2
14	5,5	112	2,345208	0,740363	0,181818
15	5,5	113	2,345208	0,740363	0,181818
16	6	123	2,44949	0,778151	0,166667
17	6	122	2,44949	0,778151	0,166667
18	6	126	2,44949	0,778151	0,166667
19	6,5	129	2,54951	0,812913	0,153846
20	7	132	2,645751	0,845098	0,142857
21	7	133	2,645751	0,845098	0,142857
22	7,5	140	2,738613	0,875061	0,133333
23	7,5	145	2,738613	0,875061	0,133333
24	8	143	2,828427	0,90309	0,125
25	8	149	2,828427	0,90309	0,125

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

Model s proměnnou X:

	Regression Summary for Dependent Variable: Y (stroje.sta) R= ,97589348 R2= ,95236809 Adjusted R2= ,95029714 F(1,23)=459,87 p<,00000 Std.Error of estimate: 8,2722					
N=25	b*	Std.Err. of b*	b	Std.Err. of b	t(23)	p-value
Intercept			33,81446	3,474708	9,73160	0,000000
X	0,975893	0,045508	14,49670	0,676008	21,44457	0,000000

Model s proměnnou SQRTX:

	Regression Summary for Dependent Variable: Y (stroje.sta) R= ,98581506 R2= ,97183132 Adjusted R2= ,97060660 F(1,23)=793,51 p<0,0000 Std.Error of estimate: 6,3614					
N=25	b*	Std.Err. of b*	b	Std.Err. of b	t(23)	p-value
Intercept			-8,57352	4,036621	-2,12393	0,044646
SQRTX	0,985815	0,034996	53,48413	1,898667	28,16931	0,000000

Model s proměnnou LOGX:

	Regression Summary for Dependent Variable: Y (stroje.sta) R= ,97364958 R2= ,94799351 Adjusted R2= ,94573236 F(1,23)=419,25 p<,00000 Std.Error of estimate: 8,6437					
N=25	b*	Std.Err. of b*	b	Std.Err. of b	t(23)	p-value
Intercept			46,17199	3,119456	14,80130	0,000000
LOGX	0,973650	0,047552	97,53245	4,763337	20,47566	0,000000

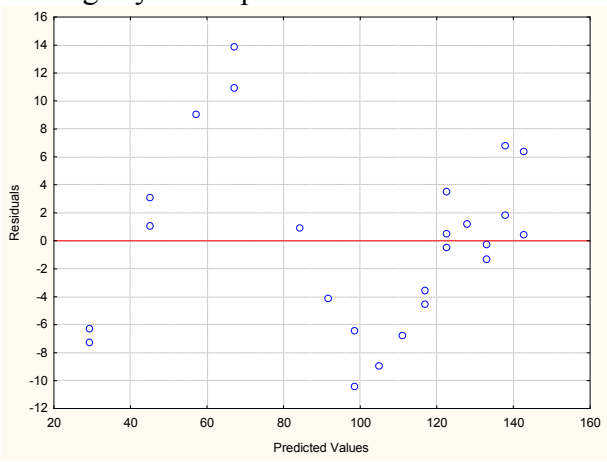
Model s proměnnou INVX:

N=25	Regression Summary for Dependent Variable: Y (stroje.sta) R= ,87243036 R2= ,76113474 Adjusted R2= ,75074929 F(1,23)=73,289 p<,00000 Std.Error of estimate: 18,525					
	b*	Std.Err. of b*	b	Std.Err. of b	t(23)	p-value
Intercept			126,2296	4,857212	25,98807	0,000000
INVX	-0,872430	0,101909	-60,9596	7,120723	-8,56088	0,000000

Vidíme, že nejvyšší index determinace poskytuje model s proměnnou SQRTX: $ID^2 = 97,2\%$.
Má také nejvyšší hodnotu testové statistiky F.
Prozkoumáme nezávislost reziduí pomocí Durbinovy – Watsonovy statistiky:
Statistiky – Vícenásobná regrese – proměnná Závislá: Y, nezávislá SQRTX – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika.

	Durbin-Watson d	Serial Corr.
Estimate	0,565691	0,697765

Hodnota Durbinovy Watsonovy statistiky je 0,5657, tedy rezidua vykazují silnou pozitivní autokorelaci. Kromě toho rezidua nejsou rozmístěna náhodně kolem 0, což je patrné z grafu závislosti reziduí na predikovaných hodnotách:
Reziduální analýza – Bodové grafy – Předpovědi vs. Rezidua



Model tedy musíme upravit:
Nejprve uložíme rezidua a predikované hodnoty:
Na záložce Reziduální analýza vybereme Uložit – Uložit rezidua a předpovědi – Vybrat vše – OK.
Pro proměnnou Rezidua z této tabulky použijeme cestu:
Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné Rezidua – ARIMA & autokorelační funkce – v Parametrech modelu ARIMA zvolíme p-Autoregresní 1 – OK (Zahájit odhady parametrů) – Souhrn: Odhady parametrů.

Paramet.	Input: Residuals (Spreadsheet2) Transformations: none Model:(1,0,0) MS Residual= 20,741					
	Param.	Asympt. Std.Err.	Asympt. t(24)	p	Lower 95% Conf	Upper 95% Conf
p(1)	0,697765	0,156004	4,472737	0,000159	0,375788	1,019741

Odhad koeficientu korelace dvou po sobě následujících reziduí je 0,6978 a je významně odlišný od 0 na hladině významnosti 0,05.

Pomocí volby Souhrn reziduí uložíme rezidua z autokorelace. K původnímu datovému souboru přidáme tři nové proměnné. Do první z nich okopírujeme rezidua z autokorelace, do druhé predikované hodnoty z původního modelu a do třetí proměnné (nazvané NOVE Y) uložíme součet reziduí a predikovaných hodnot.

	1 X	2 Y	3 SQRTX	4 LOGX	5 INVX	6 res. autokor.	7 predikce	8 NOVE Y
1	0,5	22	0,707106781	-0,301029996	2	-7,24548	29,25	22
2	0,5	23	0,707106781	-0,301029996	2	-1,18984	29,25	28,0556375
3	1	46	1	0	1	5,44726	44,91	50,357873
4	1	48	1	0	1	2,32925	44,91	47,2398677
5	1,5	66	1,22474487	0,176091259	0,666666667	6,91344	56,93	63,8443386
6	2	78	1,41421356	0,301029996	0,5	4,60744	67,06	71,6719038
7	2	81	1,41421356	0,301029996	0,5	6,30511	67,06	73,3695736
8	3	85	1,73205081	0,477121255	0,333333333	-8,78744	84,06	75,2762799
9	3,5	87	1,87082869	0,544068044	0,285714286	-4,77607	91,49	86,7100703
10	4	88	2	0,602059991	0,25	-7,51803	98,39	90,876719
11	4	92	2	0,602059991	0,25	0,85834	98,39	99,2530899
12	4,5	96	2,12132034	0,653212514	0,222222222	-4,42143	104,88	100,462032
13	5	104	2,23606798	0,698970004	0,2	-0,57787	111,02	110,442766
14	5,5	112	2,34520788	0,740362689	0,181818182	0,23576	116,86	117,09366
15	5,5	113	2,34520788	0,740362689	0,181818182	-0,35782	116,86	116,500074
16	6	123	2,44948974	0,77815125	0,166666667	3,00169	122,44	125,437007
17	6	122	2,44948974	0,77815125	0,166666667	-0,82933	122,44	121,605985
18	6	126	2,44948974	0,77815125	0,166666667	3,86843	122,44	126,303749
19	6,5	129	2,54950976	0,812913357	0,153846154	-1,27211	127,78	126,512691
20	7	132	2,64575131	0,84509804	0,142857143	-2,10253	132,93	130,829677
21	7	133	2,64575131	0,84509804	0,142857143	0,62081	132,93	133,553018
22	7,5	140	2,73861279	0,875061263	0,133333333	2,03454	137,90	139,933363
23	7,5	145	2,73861279	0,875061263	0,133333333	5,56122	137,90	143,460036
24	8	143	2,82842712	0,903089987	0,125	-4,34484	142,70	138,357612
25	8	149	2,82842712	0,903089987	0,125	6,13288	142,70	148,835332

Znovu provedeme regresní analýzu se závisle proměnnou NOVE Y a nezávisle proměnnou SQRTX:

	Regression Summary for Dependent Variable: NOVE Y (stroje3.st R= ,99239709 R2= ,98485197 Adjusted R2= ,98419337 F(1,23)=1495,3 p<0,0000 Std.Error of estimate: 4,6472					
N=25	b*	Std.Err. of b*	b	Std.Err. of b	t(23)	p-value
Intercept			-8,70033	2,948864	-2,95040	0,007177
SQRTX	0,992397	0,025663	53,63607	1,387029	38,66975	0,000000

Durbinova – Watsonova statistika v tomto novém modelu nabývá hodnoty 1,56 a již nesvědčí o existenci autokorelace reziduí.

Nyní určíme regresní odhad týdenních nákladů pro stroj starý 4 roky v tomto novém modelu se závisle proměnnou NOVE Y a nezávisle proměnnou SQRTX.
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi - Předpovědi závisle proměnné X: 2 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

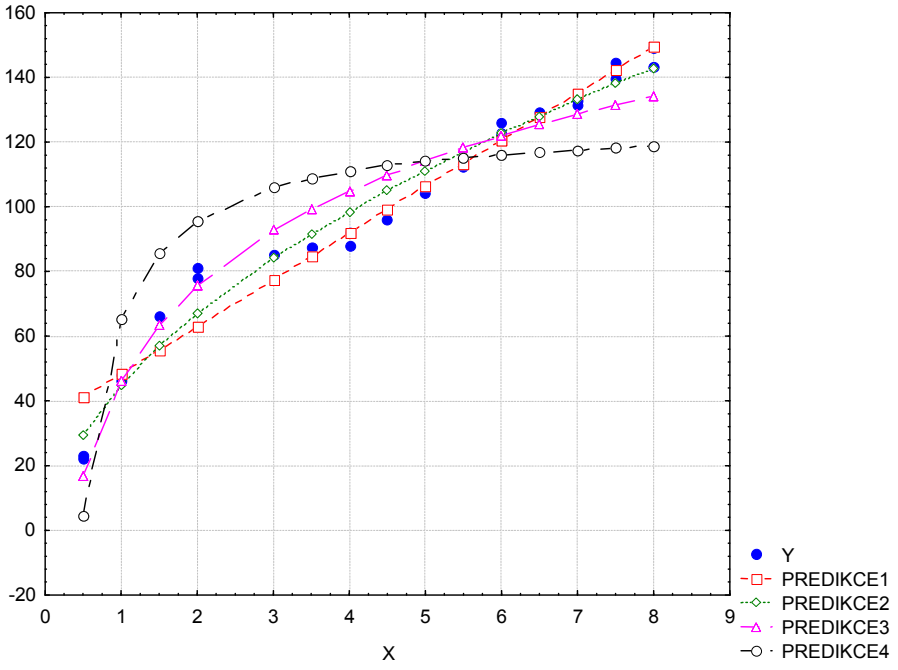
	Predicting Values for (stroje.sta variable: NOVE Y		
Variable	b-Weight	Value	b-Weight * Value
SQRTX	53,63607	2,000000	107,2721
Intercept			-8,7003
Predicted			98,5718
-95,0%CL			96,6484
+95,0%CL			100,4952

Bodový odhad je 98,57 Kč. Vidíme, že s pravděpodobností aspoň 0,95 budou týdenní náklady na údržbu stroje starého 4 roky činit minimálně 96,65 Kč a maximálně 100,50 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.
=33,8145 + 14,4967*X
=-8,7 + 53,6361*SQRTX
=46,172 + 97,5324*LOGX
=126,23 - 60,96*INVX

	1 X	2 Y	3 SQRTX	4 LOGX	5 INVX	6 PREDIKCE1	7 PREDIKCE2	8 PREDIKCE3	9 PREDIKCE4
1	0,5	22	0,707107	-0,30103	2	41,06285	29,22645	16,8118221	4,31
2	0,5	23	0,707107	-0,30103	2	41,06285	29,22645	16,8118221	4,31
3	1	46	1	0	1	48,3112	44,9361	46,172	65,27
4	1	48	1	0	1	48,3112	44,9361	46,172	65,27
5	1,5	66	1,224745	0,176091	0,666667	55,55955	56,9905384	63,3466031	85,59
6	2	78	1,414214	0,30103	0,5	62,8079	67,1529001	75,5321779	95,75
7	2	81	1,414214	0,30103	0,5	62,8079	67,1529001	75,5321779	95,75
8	3	85	1,732051	0,477121	0,333333	77,3046	84,2004503	92,7067811	105,91
9	3,5	87	1,870829	0,544068	0,285714	84,55295	91,6439549	99,2362621	108,812857
10	4	88	2	0,60206	0,25	91,8013	98,5722	104,892356	110,99
11	4	92	2	0,60206	0,25	91,8013	98,5722	104,892356	110,99
12	4,5	96	2,12132	0,653213	0,222222	99,04965	105,07935	109,881384	112,683333
13	5	104	2,236068	0,69897	0,2	106,298	111,233966	114,344222	114,038
14	5,5	112	2,345208	0,740363	0,181818	113,54635	117,087804	118,38135	115,146364
15	5,5	113	2,345208	0,740363	0,181818	113,54635	117,087804	118,38135	115,146364
16	6	123	2,44949	0,778151	0,166667	120,7947	122,681077	122,066959	116,07
17	6	122	2,44949	0,778151	0,166667	120,7947	122,681077	122,066959	116,07
18	6	126	2,44949	0,778151	0,166667	120,7947	122,681077	122,066959	116,07
19	6,5	129	2,54951	0,812913	0,153846	128,04305	128,04576	125,457391	116,851538
20	7	132	2,645751	0,845098	0,142857	135,2914	133,207782	128,59644	117,521429
21	7	133	2,645751	0,845098	0,142857	135,2914	133,207782	128,59644	117,521429
22	7,5	140	2,738613	0,875061	0,133333	142,53975	138,188509	131,518825	118,102
23	7,5	145	2,738613	0,875061	0,133333	142,53975	138,188509	131,518825	118,102
24	8	143	2,828427	0,90309	0,125	149,7881	143,0058	134,252534	118,61
25	8	149	2,828427	0,90309	0,125	149,7881	143,0058	134,252534	118,61

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.

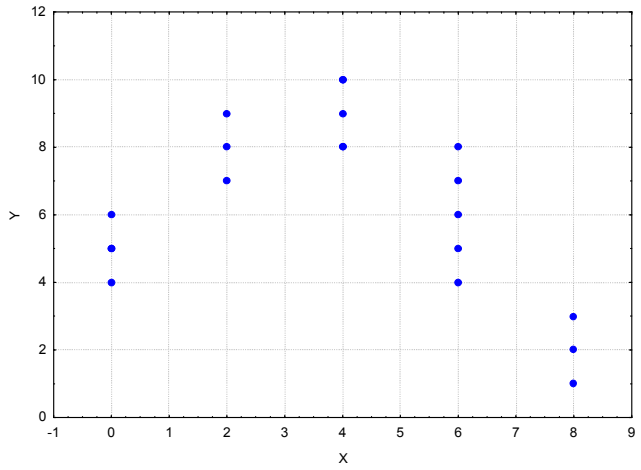
X	Y
0	5 6 4 5
2	9 7 8
4	9 8 10 10 8
6	8 5 7 4 6
8	3 1 2

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.

Test adekvátnosti modelu můžeme též provést pomocí Obecných regresních modelů:

Statistiky – Pokročilé lineární/nelineární modely – Obecné regresní modely – Jednorozměrná regrese - OK – na záložce Možnosti zaškrtneme Kvalita proložení – OK – Závislá Y, Spoj. nezáv. prom. X – OK – Více výsledků – Celkové R – ve stromové struktuře vlevo vybereme Test kvality modelu.

	Test of Lack of Fit (zimni_jablka.sta)										
Dependnt Variable	SS Residual	df Residual	MS Residual	SS Pure Err	df Pure Err	MS Pure Err	SS Lack of Fit	df Lack of Fit	MS Lack of Fit	F	p
Y	114,3056	18	6,350309	20,00000	15	1,333333	94,30556	3	31,43519	23,57639	0,000006

Hodnota testové statistiky je 23,576 a odpovídající p-hodnota je blízká 0. Na hladině významnosti 0,05 tedy zamítáme hypotézu, že přímka je vhodným modelem k popisu závislosti kvality jablek na době skladování.

Použijeme-li model $y = \beta_0 + \beta_1 x + \beta_2 x^2$, nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že tento model je adekvátní, neboť odpovídající p-hodnota je 0,4619:

	Test of Lack of Fit (zimni_jablka.sta)										
Dependnt Variable	SS Residual	df Residual	MS Residual	SS Pure Err	df Pure Err	MS Pure Err	SS Lack of Fit	df Lack of Fit	MS Lack of Fit	F	p
Y	22,16943	17	1,304084	20,00000	15	1,333333	2,169434	2	1,084717	0,813538	0,461919

Odhadnuté parametry:

	Regression Summary for Dependent Variable: Y (zimni_jablka.sta R= ,90909975 R2= ,82646235 Adjusted R2= ,80604616 F(2,17)=40,481 p<,00000 Std.Error of estimate: 1,1420					
N=20	b*	Std.Err. of b*	b	Std.Err. of b	t(17)	p-value
Intercept			5,038438	0,542163	9,29322	0,000000
X	2,32875	0,331422	2,193419	0,312162	7,02653	0,000002
Xkv	-2,78576	0,331422	-0,325953	0,038779	-8,40547	0,000000