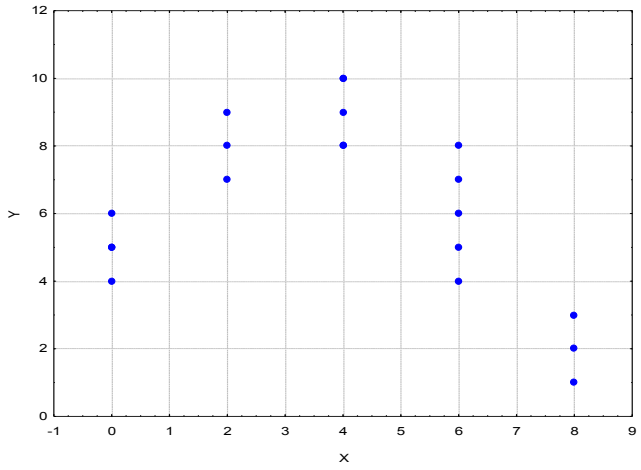
Cvičení 5.: Pokročilé metody v jednoduché lineární regresi

Příklad 1.: 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:
Načteme datový soubor zimní_jablka.sta se dvěma proměnnými X a Y a 20 případy.
Data znázorníme graficky:



Je zřejmé, že přímka nebude vhodným regresním modelem.

Odhadneme parametry regresní přímky:

	Výsledky regrese se : R= ,32440757 R2= ,1 F(1,18)=2,1171 p<,1€					
N=:	Be	Sm be	B	Sm B	t(1	Úrc
Abs			7,4	1,C	7,:	0,C
X	−0,:	0,2	−0,:	0,2	−1,	0,1

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

Efe	Analýza rozptylu (j				
	Sou čtve	s	Prů čtve	F	Úro
Reg	13,	-	13,4	2,1	0,16
Rez	114	1	6,3		
Cel	127				

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.

Pr	Analýza rozptylu Označ. efekty js							
	S e1	e	P e1	S cl	cl	P cl	F	p
Y	1(	-,2€	2(	-	1,2(			0,

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

Vypočteme testovou statistiku $F = \dfrac{(107{,}75 - 3{,}4444)/(3-2)}{(127{,}75 - 107{,}75)/(20-5)} = \dfrac{31{,}4352}{1{,}3333} = 23{,}576$ a najdeme kritický obor $W = \langle F_{0,95}(3,15), \infty) = \langle 4{,}1528, \infty)$. 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.

Dependnt Variable	Test of Lack of Fit (zimni_jablka.sta)										
	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:

Dependnt Variable	Test of Lack of Fit (zimni_jablka.sta)										
	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:

N=20	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					
	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

Výsledný model má tvar: $y = 5{,}0384 + 2{,}1934x - 0{,}3260x^2$.
Vidíme rovněž, že poměr determinace (uvedený ve výstupní tabulce regrese pod označením R2) vzrostl z 10,52% na 82,64%.

Příklad 2.: Jsou známy údaje o počtu obyvatel USA v letech 1815 až 1975 (v miliónech osob):

1815	1825	1835	1845	1855	1865	1875	1885	1895	1905	1915	1925	1935	1945	1955	1965	1975
8,3	11	14,7	19,7	26,7	35,2	44,4	55,9	68,9	83,2	98,8	114,2	127,1	140,1	164	190,9	214,3

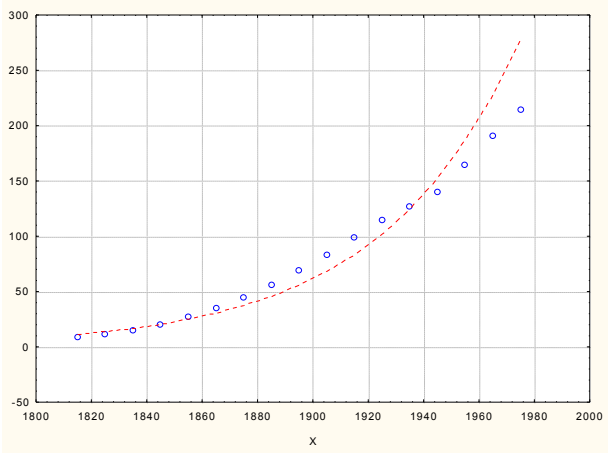
Předpokládáme, že růst populace se řídí exponenciální regresní funkcí $y = e^{\beta_0 + \beta_1 x}$, kde y je počet jedinců a x je čas, x = 1815, 1825, …, 1975.
Odhadněte parametry exponenciální regresní funkce. Znázorněte data s proloženou regresní funkcí. Pomocí D-W statistiky testujte hypotézu, že mezi rezidui neexistuje pozitivní autokorelace.
Návod: Data jsou uložena v souboru populace_USA.sta. V datovému souboru přidáme novou proměnnou lnY, do jejíhož Dlouhého jména napíšeme Log(Y). Provedeme regresní analýzu se závisle proměnnou LnY a nezávisle proměnnou X.

Výsledky regrese se závislou proměnnou : lny (populace_USA.sta)
R= ,98522411 R2= ,97066655 Upravené R2= ,96871099
F(1,15)=496,36 p<,00000 Směrod. chyba odhadu : ,18230

N=17	b*	Sm.chyba z b*	b	Sm.chyba z b	t(15)	p-hodn.
Abs.člen			-34,0828	1,710803	-19,9221	0,000000
X	0,985224	0,044222	0,0201	0,000902	22,2792	0,000000

Výsledný model má tedy tvar: $y = e^{-34{,}0828+0{,}201.x}$
Dílčí t-testy vedou k zamítnutí hypotéz o nevýznamnosti regresních parametrů β_0, β_1, obě p-hodnoty jsou blízké 0. Testová statistika celkového F-testu nabývá hodnotu 496,36, odpovídající p-hodnota je také velmi blízká 0. Exponenciální model vysvětluje variabilitu počtu osob v USA v letech 1815 – 1975 z 97%.

Grafické znázornění dat s proloženou regresní funkcí:



D-W statistika nabývá hodnoty 0,1532. Kritické hodnoty pro $\alpha = 0{,}05$, n = 15, p = 2 jsou: $d_L = 0{,}95$, $d_U = 1{,}54$. Testová statistika je menší než d_L, tedy jsme na hladině významnosti 0,05 prokázali existenci pozitivní autokorelace reziduí.

Nepovinný úkol: Postupem popsaným v přednášce odstraňte problém autokorelovaných reziduí.