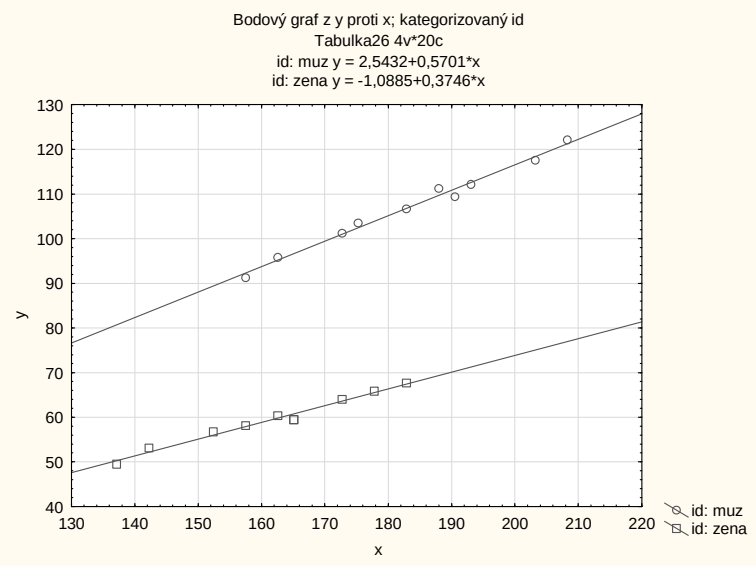
Cvičení 10.: Pokročilé metody v jednoduché lineární regresi

Příklad 1.: Máme k dispozici údaje o výšce (proměnná X) a hmotnosti (proměnná Y) 10 mužů a 10 žen. Údaje jsou uloženy v souboru hmotnost_vyska.sta. Předpokládáme, že závislost hmotnosti na výšce lze pro muže i ženy modelovat pomocí regresní přímky. Na hladině významnosti 0,05 testujte následující hypotézy:
a) rozptyly náhodných odchylek v 1. a 2. modelu jsou shodné;
b) regresní přímky jsou totožné;
c) regresní přímky mají shodné směrnice.

Řešení v systému STATISTICA:
Otevřeme datový soubor hmotnost_vyska.sta.
Znázorníme data s proloženými regresními přímkami.
Grafy – Bodové grafy – Proměnné x,y – OK – Kategorizovaný – Kategorie X – Zapnuto – Změnit proměnnou – id – OK – Rozložení Přes sebe – OK.



Z obrázku je vidět, že úseky se budu lišit a směrnice zřejmě také.

Ad a) Provedeme test hypotézy o shodě rozptylů náhodných odchylek v daných dvou modelech.
Statistiky – Vícenásobná regrese – Select cases – Zapnout filtr – zadáme id = 1 – OK – Proměnné y, x – OK – OK – Detailní výsledky – ANOVA.
Analogicky pro 2. model zadáme id = 0.

Efekt	Analýza rozptylu (hmotnost_vyska.sta) Zhrnout podmínku: id=1				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	271,9827	1	271,9827	359,8123	0,000000
Rezid.	6,0472	8	0,7559		
Celk.	278,0299				

Efekt	Analýza rozptylu (hmotnost_vyska.sta) Zhrnout podmínku: id=0				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	800,2528	1	800,2528	669,9265	0,000000
Rezid.	9,5563	8	1,1945		
Celk.	809,8092				

Vypočteme testovou statistiku $T_0 = \frac{\frac{S_E}{n-2}}{\frac{S_E^*}{n^*-2}} = \frac{\frac{6{,}0472}{8}}{\frac{9{,}5563}{8}} = 0{,}6328$.

Kritický obor:

$$W = \langle 0, F_{\alpha/2}(n-2, n^*-2)\rangle \cup \langle F_{1-\alpha/2}(n-2, n^*-2), \infty) = \langle 0, F_{0{,}025}(8,8)\rangle \cup \langle F_{0{,}975}(8,8), \infty) =$$

$$= \langle 0;0{,}2256\rangle \cup \langle 4{,}4333; \infty)$$

Testová statistika nepatří do kritického oboru, hypotézu o homogenitě rozptylů nezamítáme na hladině významnosti 0,05.

Ad b) Testujeme hypotézu o totožnosti dvou regresních přímek.

Testová statistika má tvar $T_0 = \frac{(S_{EE^*} - S_E - S_E^*)/2}{(S_E + S_E^*)/(n + n^* - 4)}$. Reziduální součty čtverců S_E a S_E^* již známe, S_E = 6,0472 a S_E^* = 9,5563. Stanovíme reziduální součet čtverců S_{EE}^* ve sdruženém modelu.

Statistiky – Vícenásobná regrese – OK – Proměnné – Závislá y, Nezávislé x – OK – OK – Detailní výsledky – ANOVA.

Efekt	Analýza rozptylu (hmotnost_vyska.sta)				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	7931,19	1	7931,194	31,59286	0,000025
Rezid.	4518,79	18	251,044		
Celk.	12449,98				

$$T_0 = \frac{(4518{,}79 - 6{,}0472 - 9{,}5563)/2}{(6{,}0472 + 9{,}5563)/16} = 2308{,}8084$$

Kritický obor:

$$W = \langle F_{1-\alpha}(2, n + n^* - 4), \infty) = \langle F_{0{,}95}(2,16), \infty) = \langle 3{,}6337; \infty)$$

Testová statistika patří do kritického oboru, hypotézu o totožnosti regresních přímek zamítáme na hladině významnosti 0,05.

Ad c) Provedeme test rovnoběžnosti dvou regresních přímek.
K datovému souboru přidáme novou proměnnou id*x, která vznikne jako součin proměnných id a x.
Statistiky – Vícenásobná regrese – OK – Proměnné – Závislá y, Nezávislé x, id id*x – OK – OK – Výpočet: výsledky regrese..

N=20	Výsledky regrese se závislou proměnnou : y (hmotnost_vyska.sta) R= ,99937316 R2= ,99874670 Upravené R2= ,99851171 F(3,16)=4250,1 p<0,0000 Směrod. chyba odhadu : ,98753					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(16)	p-hodn.
Abs.člen			2,54316	3,663263	0,69423	0,497493
x	0,420912	0,014694	0,57013	0,019903	28,64589	0,000000
id	-0,072778	0,103452	-3,63161	5,162228	-0,70350	0,491857
id*x	-0,637639	0,097802	-0,19552	0,029989	-6,51968	0,000007

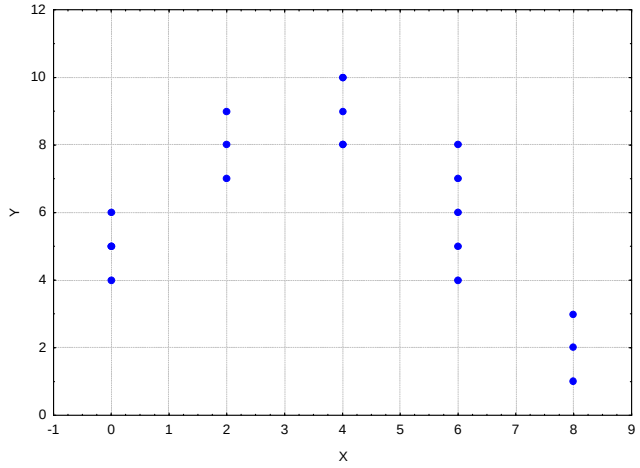
Testovou statistiku najdeme na řádku id*y, ve sloupci t(16): t_0 = -6,5197. Odpovídající p-hodnota je velmi blízká 0, tedy na hladině významnosti 0,05 zamítáme hypotézu, že směrnice daných dvou regresních přímek jsou totožné.

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



Je zřejmé, že přímka nebude vhodným regresním modelem.

Test adekvátnosti modelu provedeme 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.st 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	-0,32595	0,038779	-8,40547	0,000000

Výsledný model má tvar: $y = 5{,}0384 + 2{,}1934x - 0{,}3260x^2$.

Příklad 3.: 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.

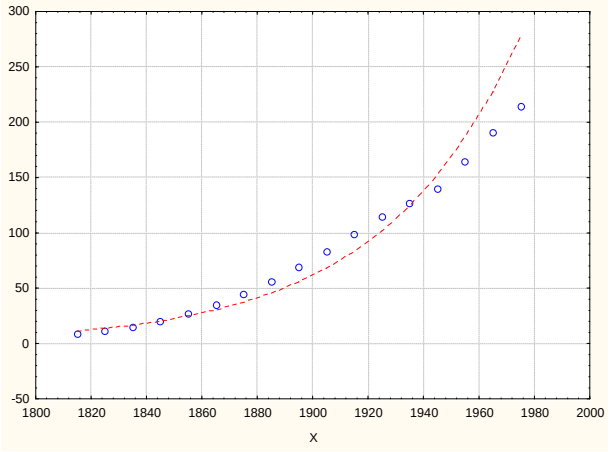
Návod: Data jsou uložena v souboru populace_USA.sta. V datovému souboru přidáme novou proměnnou LnY, do jejíhož Dlouhého jména napíšeme Log(Y). Provedeme regresní analýzu se závisle proměnnou LnY a nezávisle proměnnou X.

N=17	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					
	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.

Příklad k samostatnému řešení: Ředitel státní správy nebyl spokojen s prací jistého oddělení. Nařídil proto, aby po dobu půl roku nejméně jednou měsíčně byla práce pracovníků tohoto oddělení kontrolována a hodnocena. Po půl roce obdržel výsledky hodnocení a chtěl vědět, zda se práce zlepšila. V tabulce jsou uvedeny průměrné hodnoty bodového hodnocení (škála 1 až 10, 1 nejlepší, 10 nejhorší) za příslušné měsíce:

body	8,0	7,8	7,3	6,4	6,0	5,6
měsíc	leden	leden	leden	únor	únor	únor
body	5,4	4,8	5,7	5,0	4,8	4,7
měsíc	březen	březen	březen	duben	květen	červen

Proveďte test adekvátnosti přímkového modelu [p = 0,043] a kvadratického modelu [p = 0,6]. Pomocí kvadratického modelu najděte 95% empirický interval spolehlivosti pro predikci bodového hodnocení pro měsíc červen a pomocí statistického softwaru nakreslete graf 95% pásu spolehlivosti. [predikce = 4,896, dolní mez = 4,128, horní mez = 5,664]