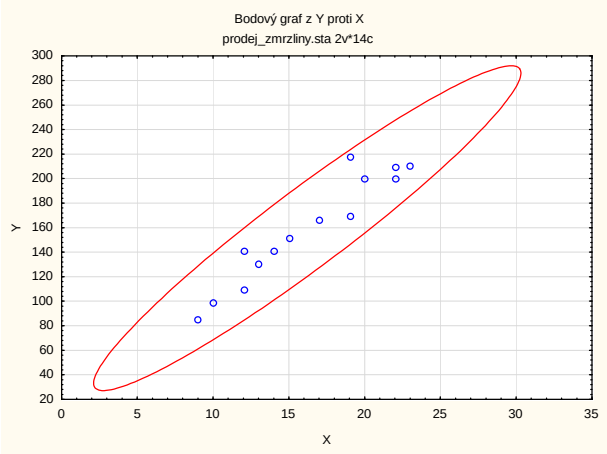
Cvičení 9.: Jednoduchá lineární regrese

Vzorový příklad: Po dobu 14 dnů byl u stánku se zmrzlinou zjišťován prodej kopečků zmrzliny (veličina Y – počet kopečků prodaných za den) v závislosti na průměrné denní teplotě (veličina Y – ve stupních Celsia).

X	20	19	12	10	9	22	23	19	22	12	13	14	15	17
Y	200	218	141	99	85	210	211	170	200	110	131	141	152	166

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtěte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

Načteme datový soubor prodej_zmrzliny.sta se dvěma proměnnými X a Y a 14 případy: Zobrazíme dvourozměrný tečkový diagram s proloženou elipsou 95% konstantní hustoty pravděpodobnosti, s jehož pomocí posoudíme dvourozměrnou normalitu dat: Grafy – Bodové grafy – vypneme Typ proložení – Proměnné X, Y - OK . Na záložce Detaily vybereme Elipsa Normální – OK. Ve vzniklém dvourozměrném tečkovém diagramu změníme rozsah zobrazených hodnot na vodorovné a svislé ose, abychom viděli celou elipsu



Ze vzhledu diagramu je patrné, že předpoklad dvourozměrné normality je oprávněný a že mezi teplotou a prodejem zmrzliny existuje vcelku silná přímá lineární závislost.

Testování hypotézy o nezávislosti: Statistika – Základní statistiky /Tabulky - Korelační matice – OK – 2 seznamy proměnných X, Y, OK. Na záložce Možnosti zaškrtneme Zobrazit detailní tabulku výsledků – Souhrn.

Prom. X & prom. Y	Korelace (prodej_zmrzliny.sta) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD) Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	16,2143	4,69334									
Y	159,5714	44,09032	0,954717	0,911484	11,11612	0,000000	14	14,14842	8,968820	-0,002646	0,101628

Ve výstupní tabulce najdeme hodnotu výběrového korelačního koeficientu R_{12} (r = 0,954717, tzn., že mezi X a Y existuje velmi silná přímá lineární závislost), realizaci testové statistiky t = 11,11612 a p-hodnotu pro test hypotézy o nezávislosti (p je velmi blízké 0, H_0 tedy zamítáme na hladině významnosti 0,05).

b) Předpokládejte, že závislost prodeje zmrzlina na teplotě lze vystihnout regresní přímkou. Vypočtěte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnná X - OK – OK – Výpočet: Výsledky regrese.

N=14	Výsledky regrese se závislou proměnnou : Y (prodej_zmrzliny.sta) R= ,95471650 R2= ,91148360 Upravené R2= ,90410723 F(1,12)=123,57 p<,00000 Směrod. chyba odhadu : 13,653					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(12)	p-hodn.
Abs.člen			14,14842	13,58155	1,04174	0,318067
X	0,954717	0,085886	8,96882	0,80683	11,11612	0,000000

Ve výstupní tabulce najdeme koeficient b_0 ve sloupci B na řádku označeném Abs. člen, koeficient b_1 ve sloupci B na řádku označeném X. Rovnice regresní přímky:
$y = 14{,}14842 + 8{,}96882\ x$.
Znamená to, že při nulové teplotě by se prodalo 14,15 kopečků zmrzliny a při zvýšení teploty o jeden stupeň by se prodej zvedl o 9 kopečků.

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

Efekt	Analýza rozptylu (prodej_zmrzliny.sta)				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	23034,49	1	23034,49	123,5681	0,000000
Rezid.	2236,94	12	186,41		
Celk.	25271,43				

Odhad rozptylu najdeme na řádku Rezid., ve sloupci Průměr čtverců, tedy $s^2 = 186{,}41$.
Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2. V našem případě $ID^2 = 0{,}9115$, tedy variabilita prodeje zmrzliny je z 91,15 % vysvětlena teplotou.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry.

Ve výstupní tabulce výsledků regrese přidáme za proměnnou p-hodn. 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;12) resp. =v3+v4*VStudent(0,975;12)

N=14	Výsledky regrese se závislou proměnnou : Y (prodej_zmrzliny.sta) R= ,95471650 R2= ,91148360 Upravené R2= ,90410723 F(1,12)=123,57 p<,00000 Směrod. chyba odhadu : 13,653							
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(12)	p-hodn.	dm =v3-v4*V	hm =v3+v4*V
Abs.člen			14,14842	13,58155	1,04174	0,318067	-15,443231	43,7400634
X	0,954717	0,085886	8,96882	0,80683	11,11612	0,000000	7,2108882	10,7267521

Vidíme, že $-15{,}44 < \beta_0 < 43{,}74$ s pravděpodobností aspoň 0,95 a $7{,}21 < \beta_1 < 10{,}73$ s pravděpodobností aspoň 0,95.

e) 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 = 123,57, p-hodnota < 0,0000, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku. (Výsledky F-testu jsou rovněž uvedeny v tabulce ANOVA.)

f) Na hladině významnosti 0,05 proveďte dílčí t-testy
Výsledky dílčích t-testů jsou uvedeny ve výstupní tabulce regrese.

	Výsledky regrese se závislou proměnnou : Y (prodej_zmrzliny.sta) R= ,95471650 R2= ,91148360 Upravené R2= ,90410723 F(1,12)=123,57 p<,00000 Směrod. chyba odhadu : 13,653					
N=14	b*	Sm.chyba z b*	b	Sm.chyba z b	t(12)	p-hodn.
Abs.člen			14,14842	13,58155	1,04174	0,318067
X	0,954717	0,085886	8,96882	0,80683	11,11612	0,000000

Testová statistika pro test hypotézy H_0: $\beta_0 = 0$ je 1,04174, p-hodnota je 0,318067. Hypotézu o nevýznamnosti úseku regresní přímky tedy nezamítáme na hladině významnosti 0,05. Testová statistika pro test hypotézy H_0: $\beta_1 = 0$ je 11,11612, p-hodnota je 0,000000. Hypotézu o nevýznamnosti směrnice regresní přímky tedy zamítáme na hladině významnosti 0,05.

g) Vypočtěte regresní odhad počtu prodaných kopečků zmrzliny při teplotě 16°C.

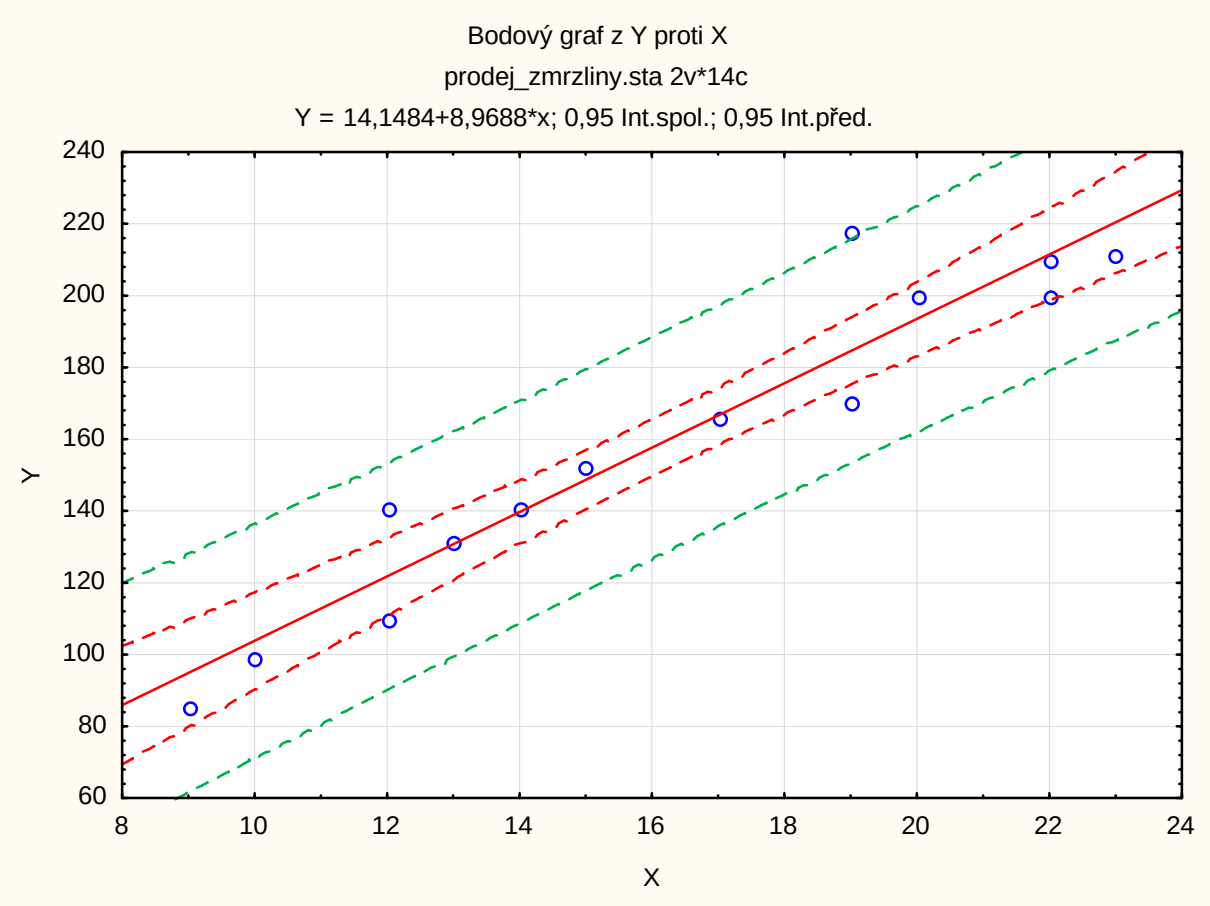
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 16 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

	Předpovězené hodnoty (prodej_zmrzliny.sta) proměnné: Y		
Proměnná	b-váha	Hodnota	b-váha * Hodnot
X	8,968820	16,00000	143,5011
Abs. člen			14,1484
Předpověď			157,6495
-95,0%LS			149,6902
+95,0%LS			165,6089

Při teplotě 16°C je predikovaná hodnota prodeje 157,65 kopečků.

h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou a 95% pásem spolehlivosti a 95% predikčním pásem.

Grafy – Bodové grafy – ponecháme Typ proložení: Lineární – Proměnné X, Y – OK – zapneme Regresní pásy – Spolehlivost - OK. Ve vytvořeném grafu 2x klikneme na jeho pozadí, z nabídky Spojnice vybereme Regresní pásy – Přidat nový pár pásů - zvolíme Typ Predikční – změníme barvu z červené na modrou - OK.



i) Vypočtěte střední absolutní procentuální chybu predikce (MAPE)

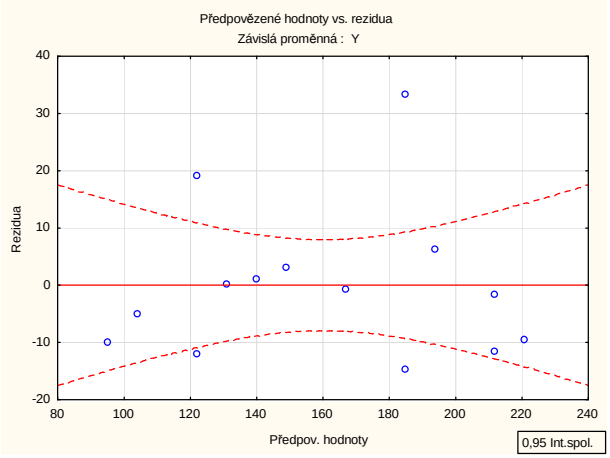
Ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua / předpoklady / předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat vše – OK. Ve vzniklé tabulce odstraníme proměnné 5 – 10, přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme
=100*abs(v4/v2)
Pak spočteme průměr této proměnné a zjistíme, že MAPE = 5,9 %.

j) Proveďte analýzu reziduí.
Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:
Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	Durbin-Watsonovo d (prodej_zmrzliny.sta) a sériové korelace reziduí	
	Durbin-Watson.d	Sériové korelace
Odhad	0,835657	0,572812

Hodnota této statistiky je velmi vzdálená od 2, svědčí o tom, že rezidua jsou kladně korelovaná.

Posouzení homoskedasticity reziduí
Reziduální analýza – Bodové grafy – Předpovědi vs. rezidua



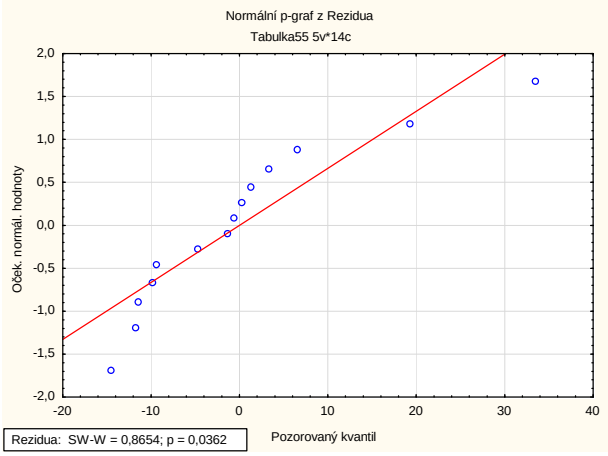
Rezidua jsou kolem nuly rozmístěna náhodně.

Testování nulovosti střední hodnoty reziduí:
Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

Proměnná	Test průměrů vůči referenční konstantě (hodnotě) (Tabulka55)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000002	13,11762	14	3,505832	0,00	-0,000000	13	1,000000

Na hladině významnosti 0,05 nezamítáme hypotézu, že střední hodnota reziduí je 0.

Posouzení normality reziduí:
Přepneme se na datovou tabulku, v níž jsou uložena rezidua. Pomocí normálního pravděpodobnostního grafu, který vykreslíme společně s výsledky S-W testu normality, získáme graf:
Na záložce Pravděpodobnostní grafy zvolíme Normální pravděpodobnostní graf reziduí:

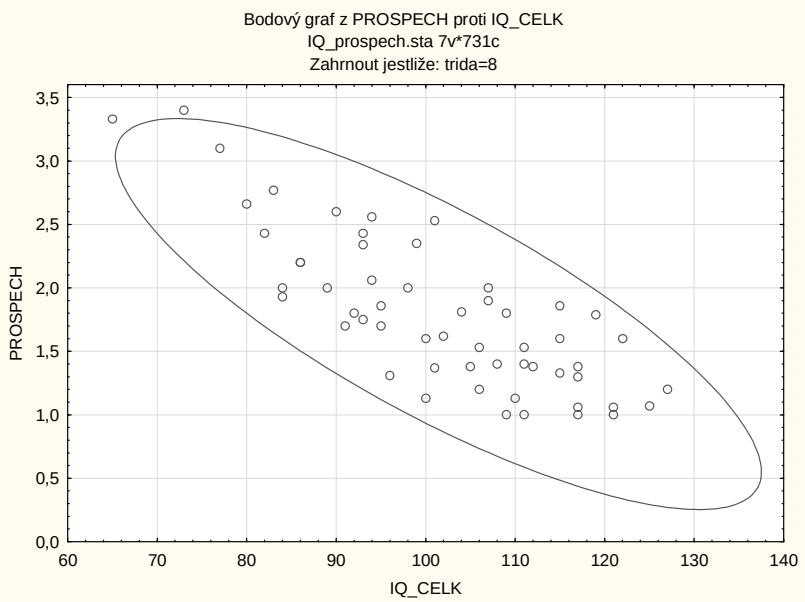


Rezidua se odchylují od ideální přímky, p-hodnota S-W testu je 0,0362, tedy rezidua se neřídí normálním rozložení.
V neprospěch jednoduchého regresního modelu hovoří nízká hodnota D-W statistiky a porušení normality reziduí. Lze to vysvětlit tím, že na prodej zmrzliny mají vliv i jiné faktory než jenom průměrná denní teplota.

Příklad k samostatnému řešení: V rámci psychologického výzkumu byly u 731 dětí ze základních škol zjišťovány následující údaje:
Pohlaví (1 – chlapec, 2 – dívka) – proměnná SEX
IQ celkové – proměnná IQ_CELK
Třída (1. až 9.) – proměnná TRIDA
Vzdělání matky (1 – základní, 2 – SŠ, 3 – VŠ) – proměnná VM
Vzdělání otce (1 – základní, 2 – SŠ, 3 – VŠ) – proměnná VO
Sídlo (1 – město, 2 – venkov) – proměnná SIDLO
Prospěch (průměrný prospěch na pololetním vysvědčení) – Proměnná PROSPECH
Údaje jsou uloženy v souboru IQ_prospech.sta.

Pro žáky z 8. třídy pomocí lineární regrese s nezávisle proměnnou IQ_CELK vysvětlete hodnoty proměnné PROSPECH.

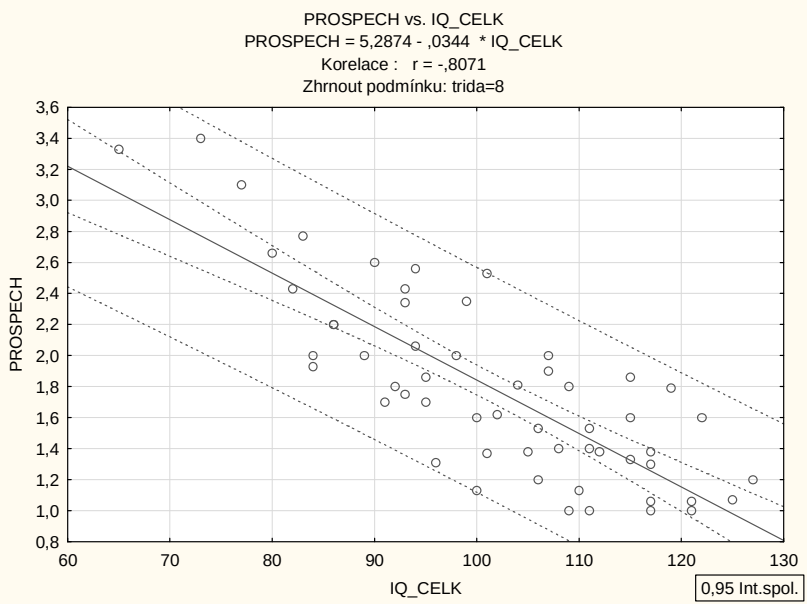
a) Dvourozměrnou normalitu dat orientačně posuďte dvourozměrným tečkovým diagramem s 95% elipsou konstantní hustoty pravděpodobnosti.



b) Vypočtěte odhady regresních parametrů, napište rovnici regresní přímky a interpretujte její parametry.

N=56	Výsledky regrese se závislou proměnnou : PROSPECH (IQ_prospech.sta) R= ,80710847 R2= ,65142408 Upravené R2= ,64496897 F(1,54)=100,92 p<,00000 Směrod. chyba odhadu : ,35806 Zhrnout podmínku: trida=8					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(54)	p-hodn.
Abs.člen			5,287439	0,351073	15,0608	0,000000
IQ_CELK	-0,807108	0,080344	-0,034447	0,003429	-10,0457	0,000000

c) Do dvourozměrného tečkového diagramu zakreslete regresní přímku s 95% pásem spolehlivosti a 95% predikčním pásem.



d) Najděte odhad rozptylu, proveďte celkový F-test a rovněž dílčí t-testy o významnosti regresních parametrů. (F-test je významný, oba dílčí t-testy rovněž, odhad rozptylu je 0,1282)

e) Najděte 95% intervaly spolehlivosti pro regresní parametry.
$4{,}5836 < \beta_0 < 5{,}9913$ s pravděpodobností aspoň 0,95,
$-0{,}0413 < \beta_1 < -0{,}0276$ s pravděpodobností aspoň 0,95.

f) Vypočtěte index determinace a interpretujte ho. Vypočtěte rovněž střední absolutní procentuální chybu predikce (MAPE) (ID^2 = 65 %, MAPE = 17,8 %).

g) Proveďte analýzu reziduí.