

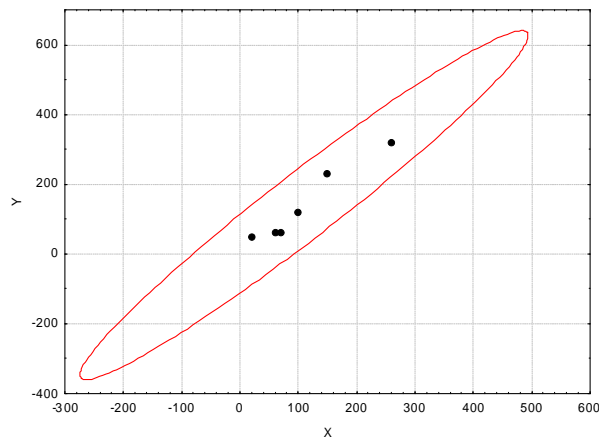
Cvičení 11.: Jednoduchá lineární regrese

Vzorový příklad: U šesti obchodníků byla zjišťována poptávka po určitém druhu zboží loni (veličina X - v kusech) a letos (veličina Y - v kusech).

číslo. obchodníka	1	2	3	4	5	6
poptávka loni (X)	20	60	70	100	150	260
poptávka letos (Y)	50	60	60	120	230	320

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtete 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 obchodnici.sta se dvěma proměnnými X a Y a 6 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 Details vybereme Elipsa Normální – OK. Ve vzniklém dvourozměrném tečkovém diagramu změním 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 loňskou a letošní poptávkou 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.

Korelace (obchodnici.sta)											
Označ. korelace jsou významné na hlad. $p < ,05000$											
(Celé případy vynechány u ChD)											
Prom. X & prom. Y	Průměr	Sm.Odch.	$r(X,Y)$	r^2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	110,0000	85,3229									
Y	140,0000	111,1755	0,971977	0,944739	8,269474	0,001167	6	0,686813	1,266484	5,566343	0,745955

Ve výstupní tabulce najdeme hodnotu výběrového korelačního koeficientu R_{12} ($r = 0,971977$, tzn. že mezi X a Y existuje velmi silná přímá lineární závislost), realizaci testové statistiky $t = 8,269474$ a p-hodnotu pro test hypotézy o nezávislosti ($p = 0,001167$, H_0 tedy zamítáme na hladině významnosti 0,05).

b) Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Vypočtete 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.

Výsledky regrese se závislou proměnnou : Y (obchodnici.sta) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219						
N=6	b*	Sm.chyba z b*	b	Sm.chyba z b	t(4)	p-hodn.
Abs.člen			0,686813	20,64236	0,033272	0,975052
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167

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:

$$y = 0,686813 + 1,266484 x.$$

Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

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

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

Analýza rozptylu (obchodnici.sta)					
Efekt	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	58384,89	1	58384,89	68,38420	0,001167
Rezid.	3415,11	4	853,78		
Celk.	61800,00				

Odhad rozptylu najdeme na řádku Rezid., ve sloupci Průměr čtverců, tedy $s^2 = 853,78$.

Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2. V našem případě $ID^2 = 0,9447$, tedy variabilita letošní poptávky je z 94,5% vysvětlena regresní přímkou.

d) Najděte 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: $=v3-v4*VStudent(0,975;4)$ resp. $=v3+v4*VStudent(0,975;4)$

Výsledky regrese se závislou proměnnou : Y (obchodnici.sta) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219								
N=6	b*	Sm.chyba z b*	b	Sm.chyba z b	t(4)	p-hodn.	dm =v3-v4*VSt	hm =v3+v4*VSt
Abs.člen			0,686813	20,64236	0,033272	0,975052	-56,625557	57,9991833
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167	0,84126639	1,69170064

Vidíme, že $-56,63 < \beta_0 < 58$ s pravděpodobností aspoň 0,95 a $0,841 < \beta_1 < 1,692$ 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 $F = 68,384$, p-hodnota $< 0,00117$, 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 (obchodníci.sta) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219						
N=6	b*	Sm.chyba z b*	b	Sm.chyba z b	t(4)	p-hodn.
Abs.člen			0,686813	20,64236	0,033272	0,975052
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167

Testová statistika pro test hypotézy $H_0: \beta_0 = 0$ je 0,033272, p-hodnota je 0,975052. Hypotézu o nevýznamnosti úseku regresní přímky tedy nezamítáme na hladině významnosti 0,05.

Testová statistika pro test hypotézy $H_0: \beta_1 = 0$ je 8,269474, p-hodnota je 0,001167. Hypotézu o nevýznamnosti směrnice regresní přímky tedy zamítáme na hladině významnosti 0,05.

g) Vypočítejte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

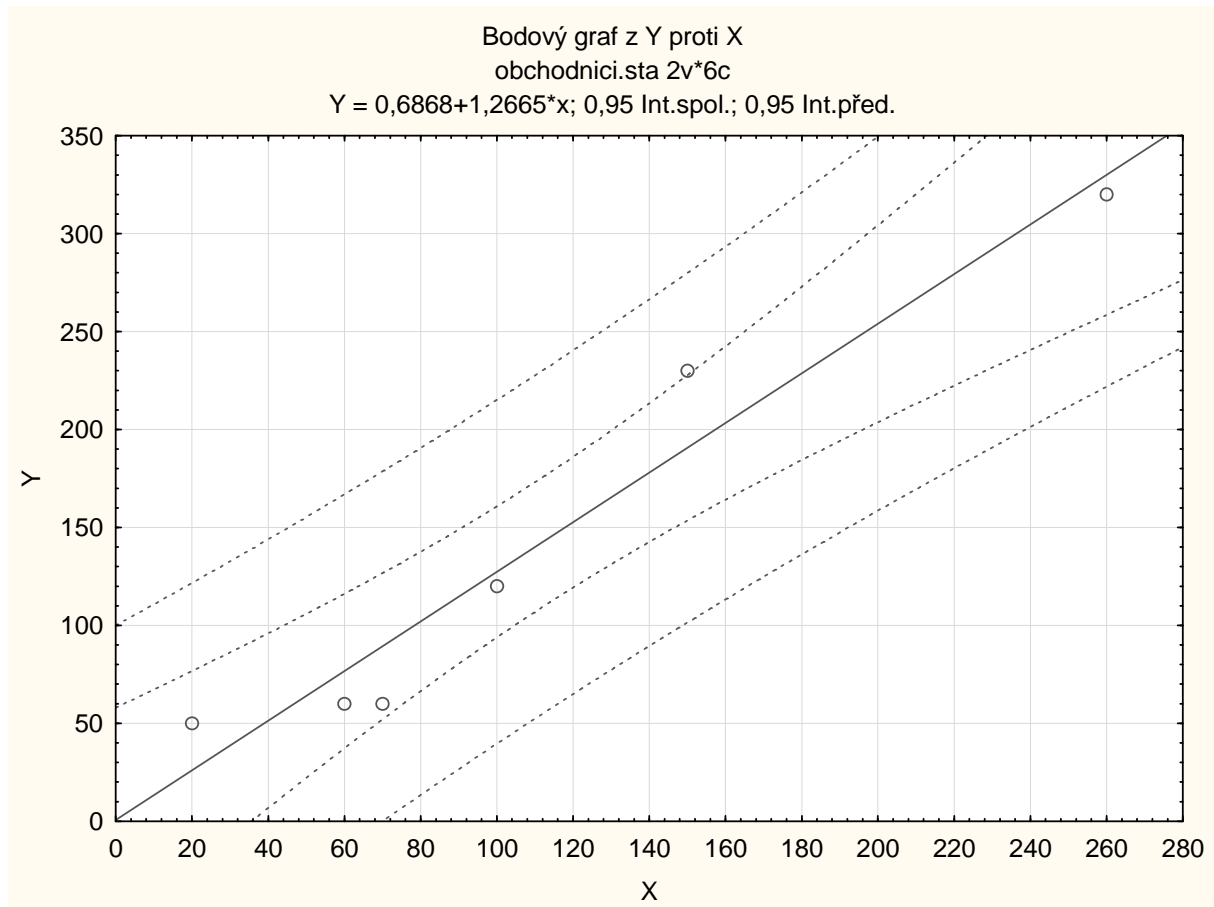
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 110 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

Předpovězené hodnoty (obchodníci.sta) proměnné: Y			
Proměnná	b-váha	Hodnota	b-váha * Hodnot
X	1,266484	110,0000	139,3132
Abs. člen			0,6868
Předpověď			140,0000
-95,0%LS			106,8803
+95,0%LS			173,1197

Při loňské poptávce 110 kusů je predikovaná hodnota letošní poptávky 140 kusů.

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) Vypočtete 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

$$= 100 * \text{abs}(v4/v2)$$

Pak spočteme průměr této proměnné a zjistíme, že $\text{MAPE} = 25,17\%$.

j) Proved'te analýzu reziduí.

Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:

Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x – OK – na záložce

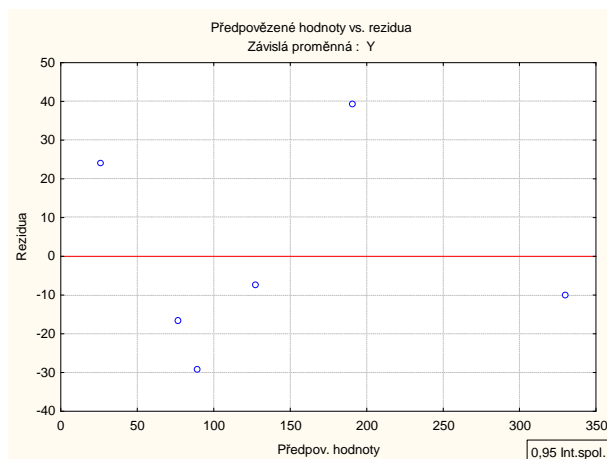
Rezidua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	Durbin-Watson.d	Sériové korelace
Odhad	2,022847	-0,113505

Hodnota této statistiky je blízká 2, svědčí o tom, že rezidua jsou nekorelovaná.

Posouzení homoskedasticity reziduí

Reziduální analýza – Bodové grafy – Předpovědi vs. rezidua



Rezidua jsou kolem nuly rozmístěna náhodně.

Testování nulovosti střední hodnoty reziduí:

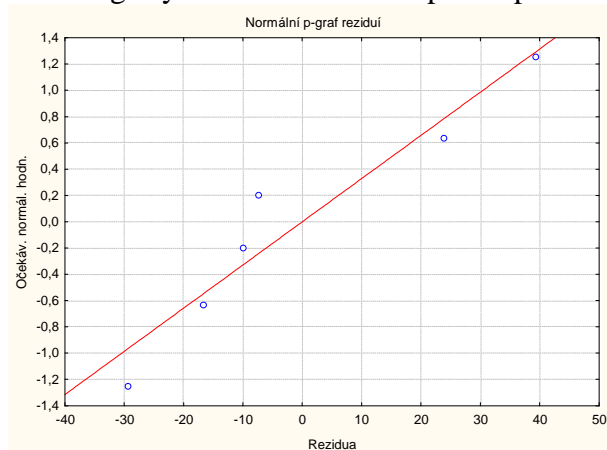
Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000003	26,13469	6	10,66944	0,00	-0,000000	5	1,000000

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

Posouzení normality reziduí:

Na záložce Pravděpodobnostní grafy zvolíme Normální pravděpodobnostní graf reziduí:



Rezidua se řadí kolem ideální přímky, lze tedy soudit, že se řídí normálním rozložením.

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

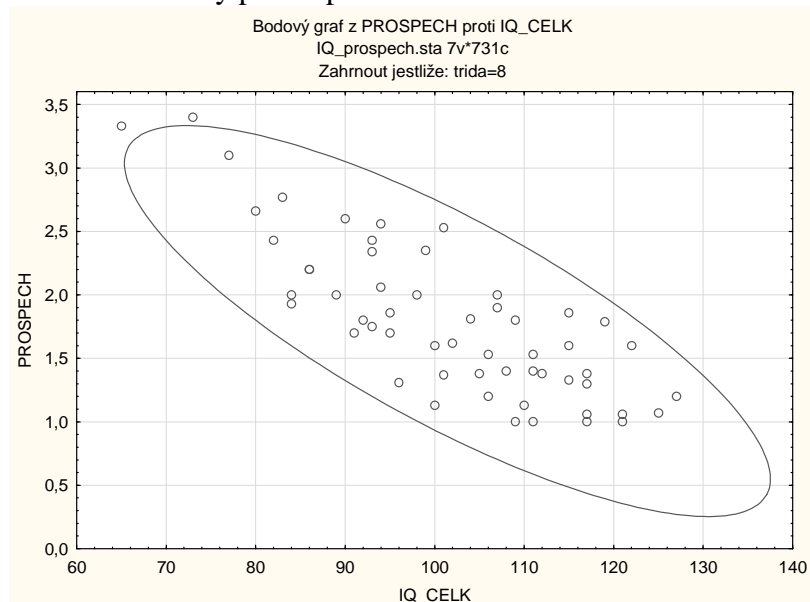
Sídlo (1 – město, 2 – venkov) – proměnná SIDLO

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.

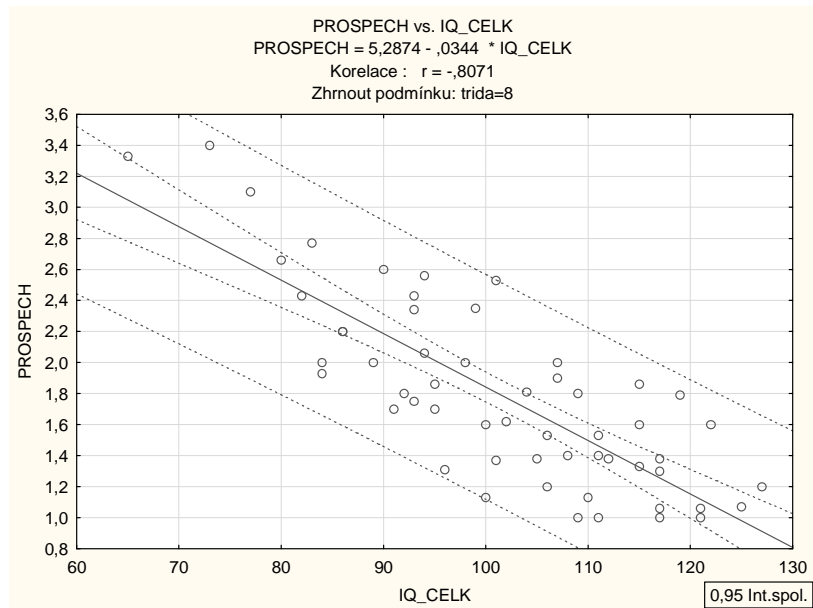
a) Dvourozměrnou normalitu dat orientačně posuďte dvourozměrným tečkovým diagramem s 95% elipsou konstantní hustoty pravděpodobnosti.



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

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						
N=56	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.