Cvičení 12: Jednoduchá regresní analýza I

Příklad 1.: V dílně pracuje 15 dělníků, u nichž byl zjištěn počet směn odpracovaných za měsíc (proměnná X) a počet zhotovených výrobků (proměnná Y).

X: 20 21 18 17 20 18 19 21 20 14 16 19 21 15 15
Y: 92 93 83 80 91 85 82 98 90 60 73 86 96 64 81

Předpokládáme, že data pocházejí z dvourozměrného normálního rozložení (orientační ověření viz příklad v přednášce 10) a že přímka je vhodným modelem závislosti veličiny Y na veličině X.
Máte k dispozici výsledky regresní analýzy, které poskytl systém STATISTICA.

N=15	Výsledky regrese se závislou proměnnou : Y (smeny a vyrobky.sta) R= ,92718009 R2= ,85966293 Upravené R2= ,84886777 F(1,13)=79,634 p<,00000 Směrod. chyba odhadu : 4,2834					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(13)	p-hodn.
Abs.člen			5,010135	8,875949	0,564462	0,582049
X	0,927180	0,103900	4,302365	0,482123	8,923795	0,000001

Efekt	Analýza rozptylu (smeny a vyrobky.sta)				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	1461,083	1	1461,083	79,63411	0,000001
Rezid.	238,517	13	18,347		
Celk.	1699,600				

a) Sestavte regresní matici.
Výsledek:

1	20
1	21
1	18
1	17
1	20
1	18
1	19
1	21
1	20
1	14
1	16
1	19
1	21
1	15
1	15

b) Napište rovnici regresní přímky.
Výsledek: y = 5,0101 + 4,3024 x

c) Jaký je regresní odhad počtu zhotovených výrobků pro dělníka, který odpracoval za měsíc 18 směn?

Výsledek: 82,45

d) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.
Výsledek: $s^2 = 18{,}347$, $ID^2 = 0{,}8597$. Znamená to, že variabilita hodnot závisle proměnné veličiny je z 85,97% vysvětlena regresní přímkou.

e) Najděte 95% intervaly spolehlivosti pro regresní parametry.
Výsledek:
$-14{,}1654 < \beta_0 < 24{,}1456$ s pravděpodobností aspoň 0,95.
$3{,}2596 < \beta_1 < 5{,}3452$ s pravděpodobností aspoň 0,95.

f) Na hladině významnosti 0,05 proveďte celkový F-test.
Výsledek: Na hladině významnosti 0,05 zamítáme hypotézu, že dostačující je model konstanty.

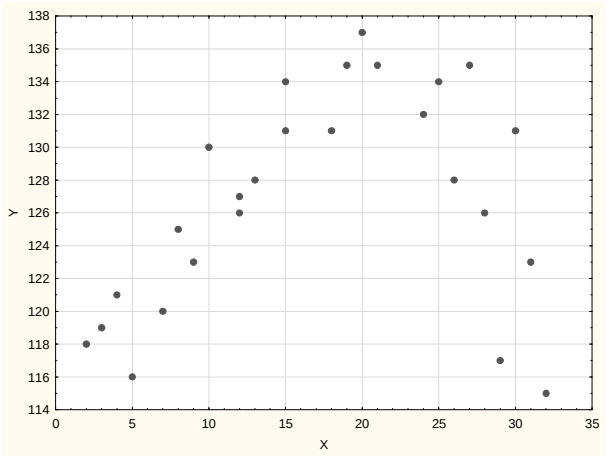
g) Na hladině významnosti 0,05 proveďte dílčí t-testy.
Výsledek:
Hypotézu o nevýznamnosti regresního parametru β_0 nezamítáme na hladině významnosti 0,05.
Hypotézu o nevýznamnosti regresního parametru β_1 zamítáme na hladině významnosti 0,05.

Upozornění:V případě modelu regresní přímky je dílčí t-test pro parametr β_1 ekvivalentní s celkovým F-testem.

Příklad 2.: U 26 dělníků byla zjištěna délka praxe (veličina X, v letech) a počet zhotovených výrobků za směnu (veličina Y):

2	4	15	3	28	10	7	20	9	15	29	19	12	31	18	13	5	25	8	27	21	32	12	24	30	26
118	121	134	119	126	130	120	137	123	131	117	135	127	123	131	128	116	134	125	135	135	115	126	132	131	128

Na základě dvourozměrného tečkového diagramu lze soudit, že vhodným modelem závislosti počtu výrobků na počtu let praxe bude regresní parabola.

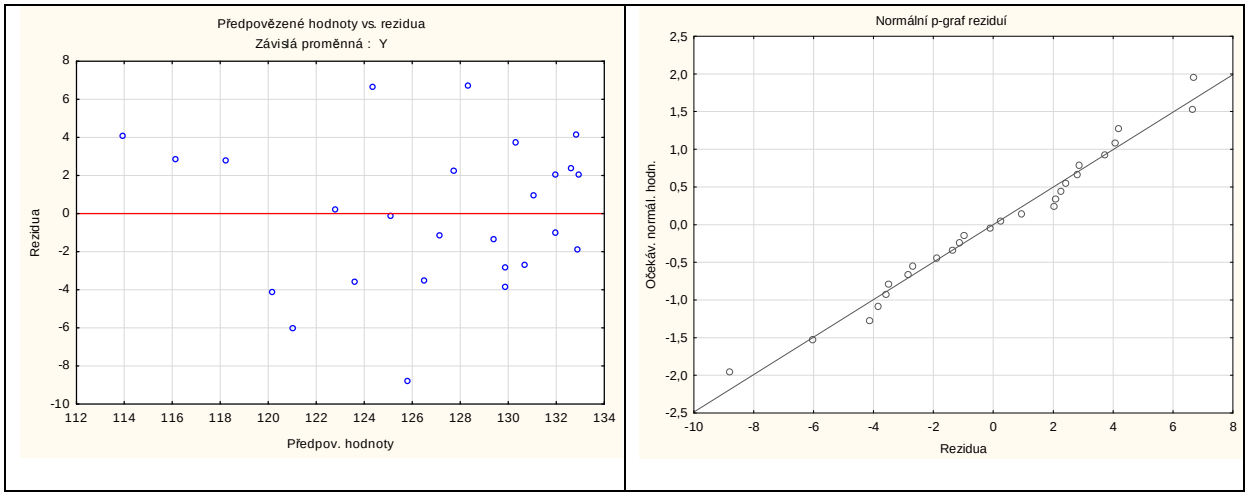


Máte k dispozici výsledky regresní analýzy, které poskytl systém STATISTICA.

N=26	Výsledky regrese se závislou proměnnou : Y (praxe_vyrobky.sta) R= ,81747863 R2= ,66827130 Upravené R2= ,63942533 F(2,23)=23,167 p<,00000 Směrod. chyba odhadu : 3,9940					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(23)	p-hodn.
Abs.člen			109,1303	2,758631	39,55958	0,000000
X	3,64589	0,542332	2,5390	0,377677	6,72261	0,000001
Xkv	-3,42712	0,542332	-0,0677	0,010717	-6,31924	0,000002

Efekt	Analýza rozptylu (praxe_vyrobky.sta)				
	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	739,134	2	369,5669	23,16688	0,000003
Rezid.	366,905	23	15,9524		
Celk.	1106,038				

	Durbin-Watsonovo d (praxe_vyrobky.sta) a sériové korelace reziduí	
	Durbin-Watson.d	Sériové korelace
Odhad	1,989379	-0,074106



a) Napište regresní rovnici vyjadřující závislost počtu zhotovených výrobků za směnu na délce praxe.
Výsledek: $y = 109{,}1303 + 2{,}539x - 0{,}0677x^2$

b) Odhadněte, kolik výrobků za směnu vyrobí dělník, jehož doba praxe je 10 let.
Výsledek: 128

c) Z kolika procent je variabilita počtu zhotovených výrobků za směnu vysvětlena uvedeným regresním modelem paraboly?
Výsledek: Model vysvětluje variabilitu veličiny Y z 66,8%.

d) Je na hladině významnosti 0,05 dostačující model konstanty? Rozhodnutí zdůvodněte.
Výsledek: Není, protože p-hodnota celkového F-testu je velmi blízká 0.

e) Najděte odhad rozptylu.

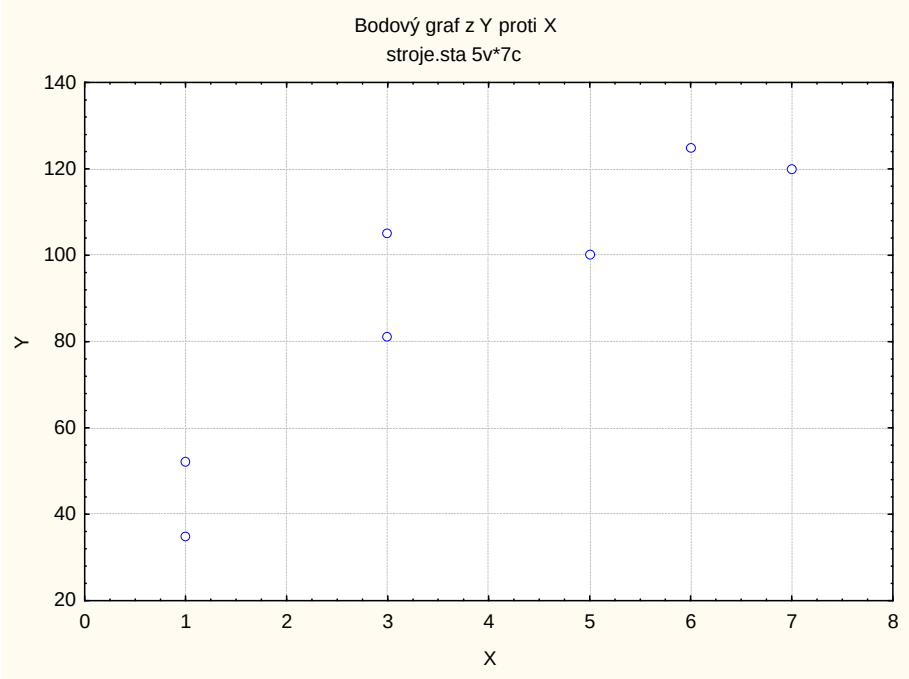
Výsledek: $s^2 = 15{,}95$

f) Na hladině významnosti 0,05 proveďte dílčí t-testy.
Výsledek:
Hypotézu o nevýznamnosti regresního parametru β_0 zamítáme na hladině významnosti 0,05.
Hypotézu o nevýznamnosti regresního parametru β_1 zamítáme na hladině významnosti 0,05.
Hypotézu o nevýznamnosti regresního parametru β_2 zamítáme na hladině významnosti 0,05.

g) Lze považovat rezidua za nezávislá, homoskedastická a normálně rozložená?
Výsledek: Ano.

Příklad 3.: U sedmi 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:
(1,35), (1,52), (3,81), (3,105), (5,100), (6,125), (7, 120)
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. Určete regresní odhad týdenních nákladů pro stroj starý čtyři roky.

Návod na řešení pomocí systému STATISTICA:
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	1	35	1	0	1
2	1	52	1	0	1
3	3	81	1,732051	0,477121	0,333333
4	3	105	1,732051	0,477121	0,333333
5	5	100	2,236068	0,69897	0,2
6	6	125	2,44949	0,778151	0,166667
7	7	120	2,645751	0,845098	0,142857

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

N=7	Výsledky regrese se závislou proměnnou : Y (stroje.sta) R= ,91004028 R2= ,82817331 Upravené R2= ,79380797 F(1,5)=24,099 p<,00444 Směrod. chyba odhadu : 15,487					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			39,44444	11,54341	3,417054	0,018898
X	0,910040	0,185379	13,14957	2,67862	4,909082	0,004439

Model s proměnnou SQRTX:

N=7	Výsledky regrese se závislou proměnnou : Y (stroje.sta) R= ,93923698 R2= ,88216611 Upravené R2= ,85859933 F(1,5)=37,433 p<,00169 Směrod. chyba odhadu : 12,825					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			-0,47736	15,29638	-0,031207	0,976312
SQRTX	0,939237	0,153515	48,55972	7,93690	6,118220	0,001691

Model s proměnnou LOGX:

N=7	Výsledky regrese se závislou proměnnou : Y (stroje.sta) R= ,95349135 R2= ,90914576 Upravené R2= ,89097491 F(1,5)=50,033 p<,00087 Směrod. chyba odhadu : 11,262					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			44,64571	7,49541	5,956407	0,001907
LOGX	0,953491	0,134799	93,23472	13,18100	7,073415	0,000874

Model s proměnnou INVX

N=7	Výsledky regrese se závislou proměnnou : Y (stroje.sta) R= ,94282234 R2= ,88891396 Upravené R2= ,86669676 F(1,5)=40,010 p<,00146 Směrod. chyba odhadu : 12,452					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			126,6192	7,67327	16,50134	0,000015
INVX	-0,942822	0,149054	-84,4832	13,35627	-6,32536	0,001456

Vidíme, že nejvyšší index determinace poskytuje model s proměnnou LOGX: ID^2 = 90,9 %.
Má také nejmenší směrodatnou chybu odhadu.

Určíme regresní odhad týdenních nákladů pro stroj starý 4 roky v modelu s nezávisle proměnnou LOGX. Nejprve vypočteme log(4) = 0,602
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 0,602 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

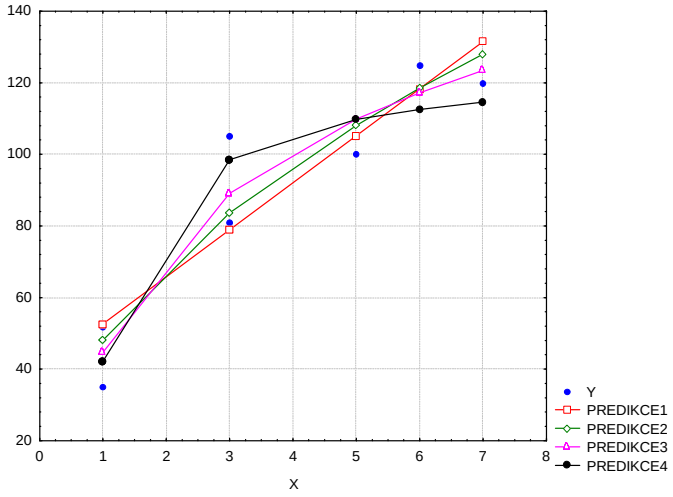
Proměnná	Předpovězené hodnoty (stroje.sta) proměnné: Y B-váž.	Hodnota	B-váž. * Hodnot
LOGX	93,23472	0,602000	56,1273
Abs. člen			44,6457
Předpověď			100,7730
-95,0%LS			88,9277
+95,0%LS			112,6184

Bodový odhad je 100,77 Kč. Vidíme, že s pravděpodobností aspoň 0,95 budou týdenní náklady na údržbu stroje starého 4 roky činit minimálně 88,93 Kč a maximálně 112,62 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.
=39,44444+13,14957*x
=-0,4776+48,55972*sqrtx
=44,64571+93,23472*logx
=126,6192-84,4832*invx

	1 X	2 Y	3 SQRTX	4 LOGX	5 INVX	6 PREDIKCE1	7 PREDIKCE2	8 PREDIKCE3	9 PREDIKCE4
1	1	35	1	0	1	52,59401	48,08212	44,64571	42,136
2	1	52	1	0	1	52,59401	48,08212	44,64571	42,136
3	3	81	1,732051	0,477121	0,333333	78,89315	83,6303022	89,1299766	98,4581333
4	3	105	1,732051	0,477121	0,333333	78,89315	83,6303022	89,1299766	98,4581333
5	5	100	2,236068	0,69897	0,2	105,19229	108,105235	109,813983	109,72256
6	6	125	2,44949	0,778151	0,166667	118,34186	118,468936	117,196424	112,538667
7	7	120	2,645751	0,845098	0,142857	131,49143	127,999343	123,438189	114,550171

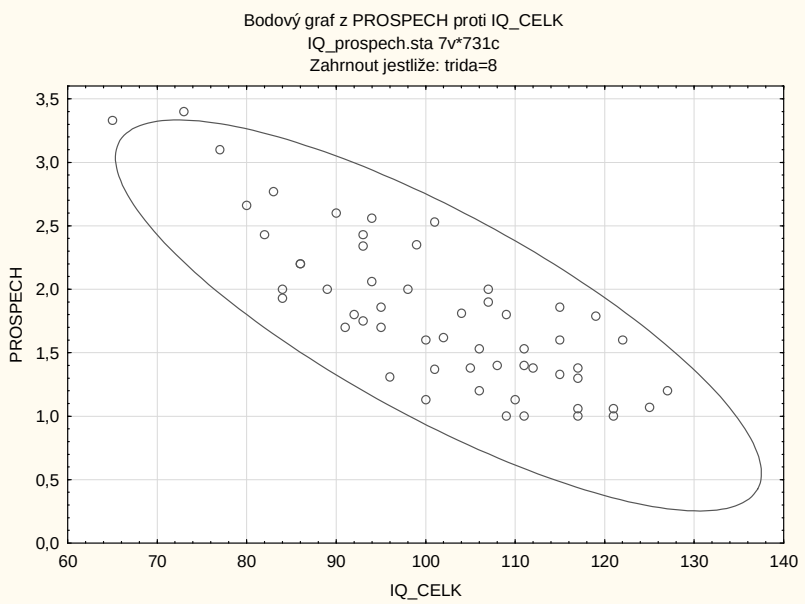
Obrázek vytvoříme pomocí vícenásobného bodového grafu.



Příklad 4. (k samostatnému řešení pomocí STATISTIKY): 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.
Následující úkoly provádějte pro žáky z 8. třídy.

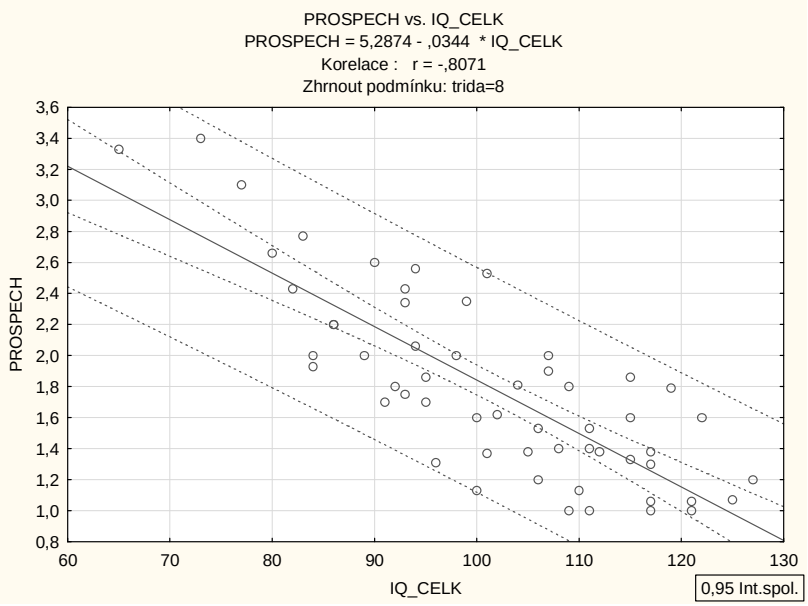
a) Dvourozměrnou normalitu dat orientačně posuďte dvourozměrným tečkovým diagramem s 95% elipsou konstantní hustoty pravděpodobnosti.



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.