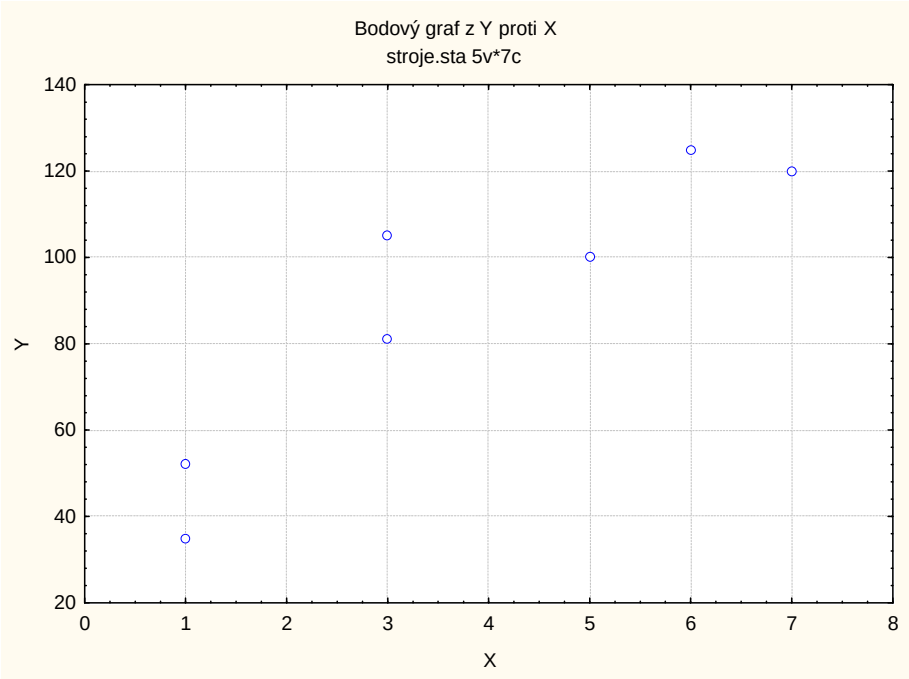
Cvičení 10.: Jednoduchá lineární regrese

Vzorový příklad: 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.

Řešení:
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.

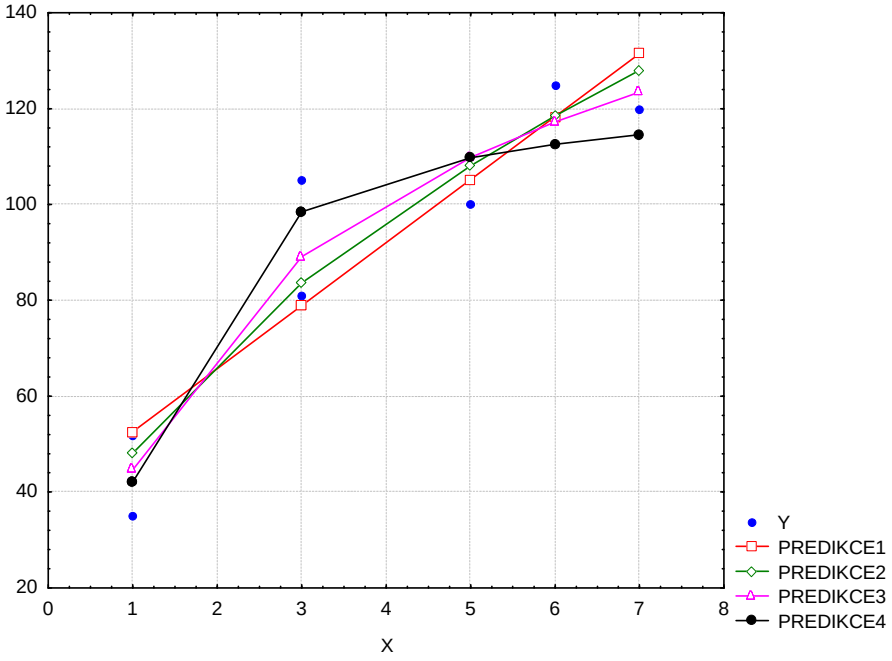
Proměnná	Předpovězené hodnoty (stroje.sta) proměnné: Y		
	B-váž.	Hodnota	B-váž. * Hodnot
LOGX	93,23472	0,602000	56,1273
Abs. člen			44,6457
Předpověď			100,7730
-95,0%LS			88,9277
+95,0%LS			112,6184

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.

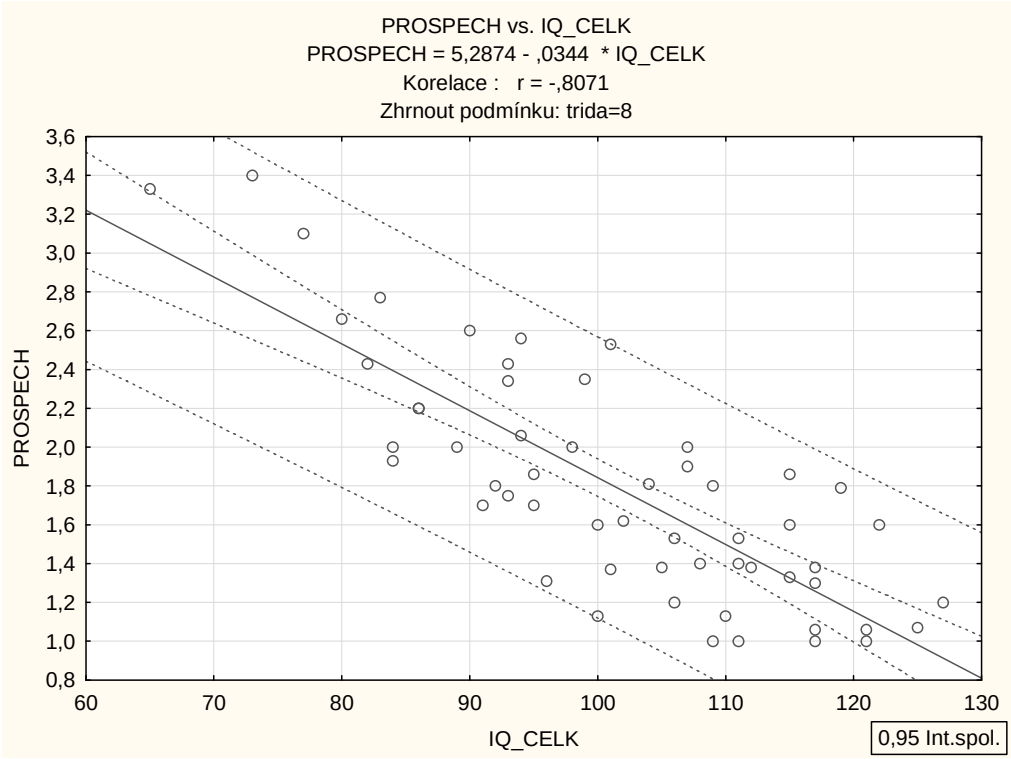


Příklady k samostatnému řešení

Příklad 1.: 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.
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.
e) Najděte 95% intervaly spolehlivosti pro regresní parametry a zjistěte relativní chyby odhadů regresních parametrů. (Pro β_0 je relativní chyba odhadu 13,3 %, pro β_1 20 %.)
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.

Příklad 2.: V r. 2010 bylo u studentů MU provedeno dotazníkové šetření, které se týkalo údajů o mobilech. Na dotazník odpovědělo 67 respondentů.

Dotazník:
1. Jaká je značka Vašeho mobilního telefonu?
a) LG
b) Nokia
c) Samsung
d) Sony Ericsson
e) jiná
f) nemám mobilní telefon

2. Jaký operační systém Váš mobilní telefon používá?
a) Android
b) Symbian
c) Windows Mobile
d) jiný
e) žádný

3. Na pětibodové stupnici vyjádřete spokojenost s funkcemi svého mobilu
a) naprostá spokojenost
b) spokojenost
c) neutrální postoj
d) nespokojenost
e) naprostá nespokojenost

4. Máte v úmyslu si do konce roku 2010 pořídit jiný mobil?
a) ano
b) ne

5. Jaká je barva Vašeho mobilu?
a) bílá
b) černá
c) stříbrná
d) jiná

6. Uveďte hmotnost svého mobilu (v g)

7. Uveďte výšku svého mobilu (v mm)

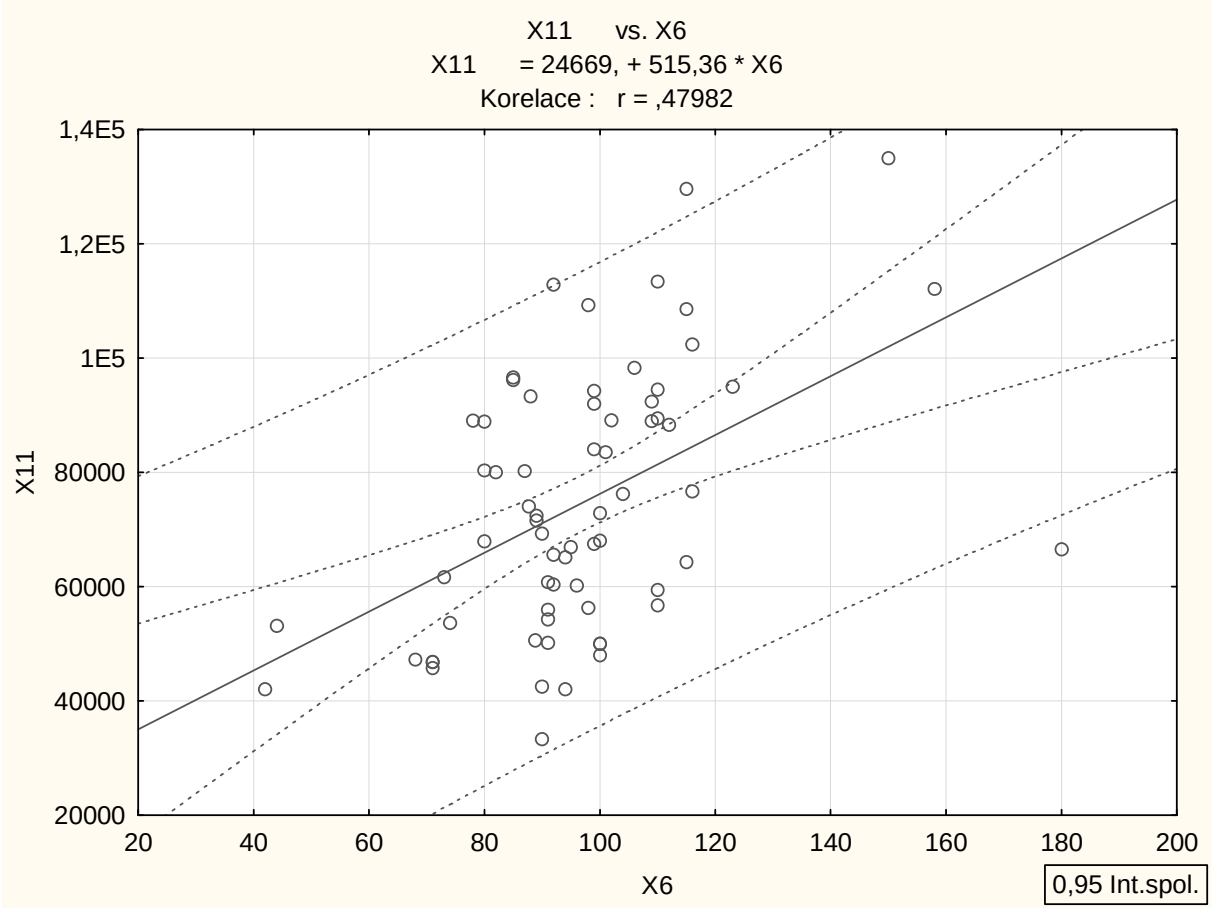
8. Uveďte šířku svého mobilu (v mm)

9. Uveďte hloubku svého mobilu (v mm)

10. Jaké je Vaše pohlaví?
a) muž
b) žena

Získaná data jsou uložena v souboru dotazník_mobily.sta.

Modelujte závislost objemu mobilu na jeho hmotnosti pomocí regresní přímky.
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.
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.
e) Najděte 95% intervaly spolehlivosti pro regresní parametry a zjistěte relativní chyby odhadů regresních parametrů. (Pro β_0 je relativní chyba odhadu 93,6 %, pro β_1 45,3 %.)
f) Vypočtěte index determinace a interpretujte ho. Vypočtěte rovněž střední absolutní procentuální chybu predikce (MAPE). (ID^2 = 23 %, MAPE = 24,2 %)
g) Proveďte analýzu reziduí.