Téma 4: Korelace a regrese

Vzorový příklad: Pro následující datové soubory proveďte korelační, resp. regresní analýzu.

Postup ve STATISTICE:

1. Načtěte soubor **znamky.sta**. Vypočtěte Spearmanův korelační koeficient známek z matematiky a angličtiny pro všechny studenty, pak zvlášť pro muže a zvlášť pro ženy. Získané výsledky interpretujte.
   (Spearmanův korelační koeficient měří těsnost lineární závislosti dvou ordinálních proměnných x, y a počítá se podle vzorce:

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2 ,$$

   kde R_i je pořadí x_i - tj. počet těch hodnot $x_1, ..., x_n$, které jsou $\leq x_i$ a Q_i je pořadí y_i.)
   Hodnoty Spearmanova korelačního koeficientu (stejně tak hodnoty dále uvedeného Pearsonova korelačního koeficientu) interpretujeme podle následující tabulky:

Absolutní hodnota korelačního koeficientu	Interpretace hodnoty
0	lineární nezávislost
(0, 0,1)	velmi nízký stupeň závislosti
[0,1, 0,3)	nízký stupeň závislosti
[0,30, 0,50)	mírný stupeň závislosti
[0,50, 0,70)	význačný stupeň závislosti
[0,70, 0,90)	vysoký stupeň závislosti
(0,90, 1)	velmi vysoký stupeň závislosti
1	úplná lineární závislost

   Návod: Po načtení souboru zvolíme Statistics – Nonparametrics – Correlations – OK – Variables First variable list X, Second variable list Y – OK – Spearman R. Počítáme-li r_S pro muže, vybereme v tabulce Nonparametric Correlation tlačítko Select Cases – Specific, select by Z=1.

Řešení:

Pro všechny

Pair of Variables	Spearman Rank Order Correlations (zna MD pairwise deleted Marked correlations are significant at p < Valid N	Spearman R	t(N-2)	p-level
X & Y	20	0,688442	4,027090	0,000791

Pro muže (if Z=1)

Pair of Variables	Spearman Rank Order Correlations (zna MD pairwise deleted Marked correlations are significant at p < Valid N	Spearman R	t(N-2)	p-level
X & Y	10	0,373544	1,138990	0,287662

Pro ženy (if Z=0)

Pair of Variables	Spearman Rank Order Correlations (zna MD pairwise deleted Marked correlations are significant at p <			
	Valid N	Spearman R	t(N-2)	p-level
X & Y	10	0,860314	4,773446	0,001402

Komentář: Ve skupině všech studentů je Spearmanův koeficient korelace roven 0,6884, což svědčí o význačné těsnosti pořadové závislosti. U mužů nabývá tento koeficient hodnoty pouze 0,3735, tedy mezi známkami z matematiky a angličtiny existuje u mužů pouze mírná pořadová závislost. Naproti tomu u žen je sledovaná pořadová závislost vysoká, protože Spearmanův koeficient je 0,8603.

2. Vysvětlení významu Pearsonova korelačního koeficientu: Načtěte soubor **korkoef.sta**, který obsahuje proměnné X,Y1,Y2,Y3,Y4, X4. Vypočtěte Pearsonovy korelační koeficienty dvojic proměnných (X,Y1), (X,Y2), (X,Y3), (X4,Y4) a pro každou z uvedených dvojic proměnných nakreslete dvourozměrný tečkový diagram. Pro které dvojice proměnných se hodí Pearsonův korelační koeficient jako vhodná míra těsnosti lineární závislosti?
Návod: Statistics – Basis Statistics/Tables – Correlation matrices – OK – One variable list X, Y1 – OK – Summary: Correlation matrix – Návrat do Product-Moment and Partial Correlations – Advanced/plot – 2D Scatterplots – OK – First X, Second Y1 – OK. Analogicky pro ostatní dvojice proměnných.

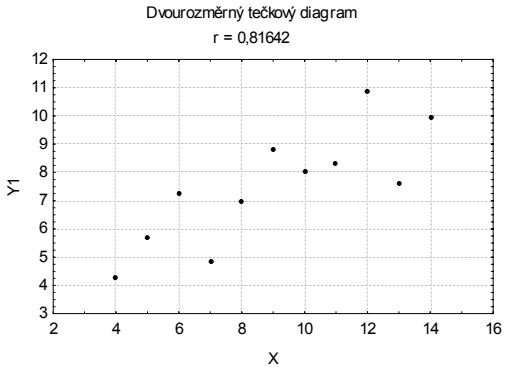
Řešení:

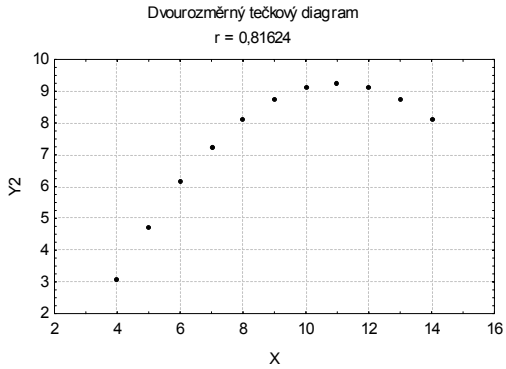
Variable	Correlations (korkoe X	Y1
X	1,000000	0,816421
Y1	0,816421	1,000000

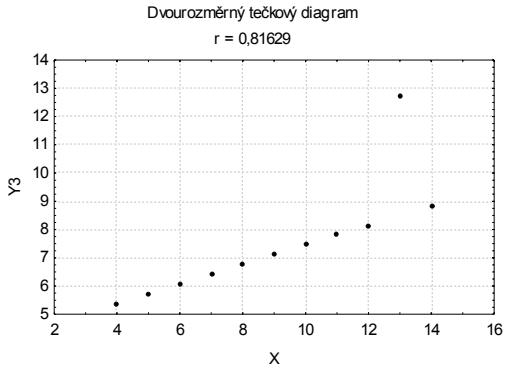
Variable	Correlations (korkoe X	Y2
X	1,000000	0,816237
Y2	0,816237	1,000000

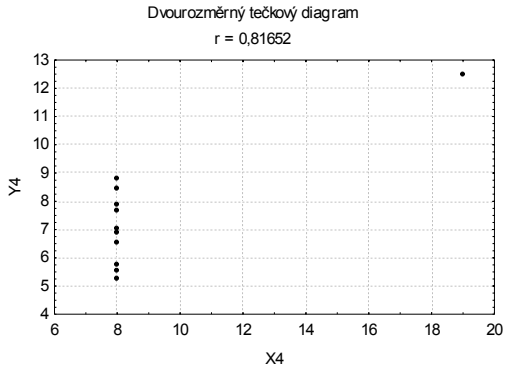
Variable	Correlations (korkoe X	Y3
X	1,000000	0,816287
Y3	0,816287	1,000000

Variable	Correlations (korkoe X4	Y4
X4	1,000000	0,816521
Y4	0,816521	1,000000









Komentář: Ve všech čtyřech případech nabývá koeficient korelace hodnoty 0,816, což by svědčilo o vysokém stupni těsnosti lineárního vztahu mezi sledovanými dvojicemi veličin. Při pohledu na dvourozměrné tečkové diagramy je však zřejmé, že pouze v prvním případě je použití Pearsonova korelačního koeficientu oprávněné.

3. Načtěte do STATISTIKY soubor **ocel.sta**. Vypočtěte kovarianci a Pearsonův koeficient korelace meze plasticity a meze pevnosti. Porovnejte s výsledky ve skriptech Popisná statistika (str. 30).
   Návod: Po načtení souboru zvolíme Statistics - Multiple Regression - Variables Independent X, Dependent Y – OK – OK – Residuals/assumption-prediction – Descriptive statistics – Covariances. Pro získání korelačního koeficientu zvolíme Correlation místo Covariances.
   Vysvětlení: Kovariance vyjde ve STATISTICE jinak než ve skriptech, protože ve STATISTICE se ve vzorci pro výpočet kovariance nepoužívá 1/n, ale 1/(n-1).

Řešení:

	Correlations (ocel)	
Variable	X	Y
X	1,000000	0,934548
Y	0,934548	1,000000

	Covariances (ocel)	
Variable	X	Y
X	1070,240	1002,471
Y	1002,471	1075,125

Komentář: Kovariance meze plasticity a meze pevnosti vyšla 1002,471, tedy mezi těmito dvěma znaky existuje určitý stupeň přímé lineární závislosti. Koeficient korelace meze plasticity a meze pevnosti nabývá hodnoty 0,9345, což svědčí o velmi vysokém stupni přímé lineární závislosti obou znaků (viz tabulku v úkolu 1).

4. Určete koeficienty regresní přímky meze pevnosti na mez plasticity a stanovte index determinace. Určete regresní odhad meze pevnosti, je-li mez plasticity 110. Nakreslete regresní přímku do dvourozměrného tečkového diagramu.
   Návod: V tabulce Multiple Regression zvolíme Variables Independent X, Dependent Y – OK – Summary:Regression results. Ve výstupní tabulce najdeme koeficient b_0 ve sloupci B na řádku označeném Intercept, koeficient b_1 ve sloupci B na řádku označeném X, index determinace pod označením R2.
   Pro výpočet predikované hodnoty zvolíme Residuals/assumption/prediction Predict dependent variable X:110 - OK. Ve výstupní tabulce je hledaná hodnota označena jako Predictd.
   Nakreslení regresní přímky: Návrat do Multiple Regression – Residuals / assumption / prediction – Perform residuals analysis – Scatterplots – Bivariate correlation – X, Y – OK.

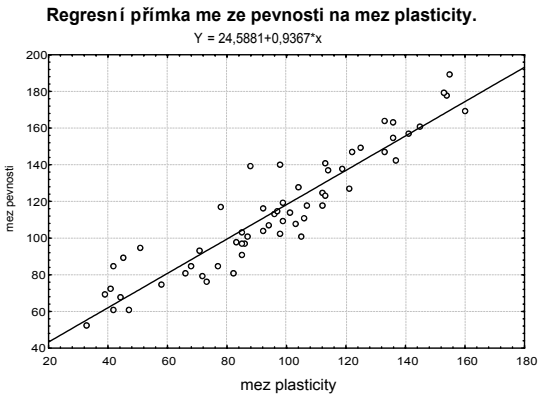
Jiný způsob: Do dvourozměrného tečkového diagramu nakreslíme regresní přímku tak, že v tabulce 2D Scatterplots zvolíme Fit Linear, OK.

Řešení:

Statistic	Summary Value
Multiple R	0,9345
Multiple R2	0,8734
Adjusted R2	0,8712
F(1,58)	400,0641
p	0,0000
Std.Err. of Estimate	11,7677

Variable	Predicting Values for (ocel) variable: Y		
	B-Weight	Value	B-Weight * Value
X	0,936679	110,0000	103,0346
Intercept			24,5881
Predicted			127,6228
-95,0%CL			124,3063
+95,0%CL			130,9392

N=60	Regression Summary for Dependent Variable: Y (ocel) R= ,93454811 R2= ,87338017 Adjusted R2= ,87119707 F(1,58)=400,06 p<0,0000 Std.Error of estimate: 11,768					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(58)	p-level
Intercept			24,58814	4,740272	5,18707	0,000003
X	0,934548	0,046724	0,93668	0,046830	20,00160	0,000000



Komentář: Regresní přímka meze pevnosti na mez plasticity má rovnici:
Y = 24,58814 + 0,93668 X.
Index determinace nabývá hodnoty 0,8734, tedy variabilita meze pevnosti je z 87,34% vysvětlena regresní přímkou.
Je-li mez plasticity 110, je predikovaná hodnota meze pevnosti rovna 127,62.
Na dvourozměrném tečkovém diagramu je vidět, že regresní přímka je vhodná pro modelování závislosti meze pevnosti na mezi plasticity – tečky jsou rozmístěny vcelku rovnoměrně kolem regresní přímky.

5. 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: Datový soubor s proměnnými X a Y doplňte o proměnné SQRTX, LOGX a INVX. Hodnoty proměnné SQRTX získáte tak, že do Long Name napíšete =sqrt(x). (Analogicky

pro ostatní proměnné.) Regresní analýzu provedete tak, že roli nezávisle proměnné bude hrát proměnná X, pak SQRTX, LOGX a nakonec INVX.

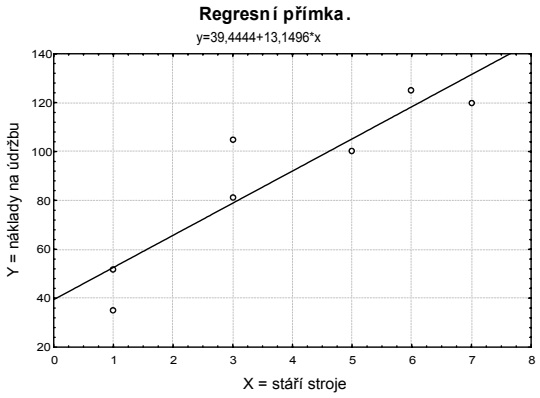
Řešení:

Model s proměnnou X

Statistic	Summary Value
Multiple R	0,91004
Multiple R2	0,82817
Adjusted R2	0,79381
F(1,5)	24,09909
p	0,00444
Std.Err. of Estimate	15,48711

Variable	Predicting Values for (stroje) variable: Y		
	B-Weight	Value	B-Weight * Value
X	13,14957	4,000000	52,5983
Intercept			39,4444
Predicted			92,0427
-95,0%CL			76,8676
+95,0%CL			107,2179

N=7	Regression Summary for Dependent Variable: Y (stroje) R= ,91004028 R2= ,82817331 Adjusted R2= ,79380797 F(1,5)=24,099 p<,00444 Std.Error of estimate: 15,487					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(5)	p-level
Intercept			39,44444	11,54341	3,417054	0,018898
X	0,910040	0,185379	13,14957	2,67862	4,909082	0,004439

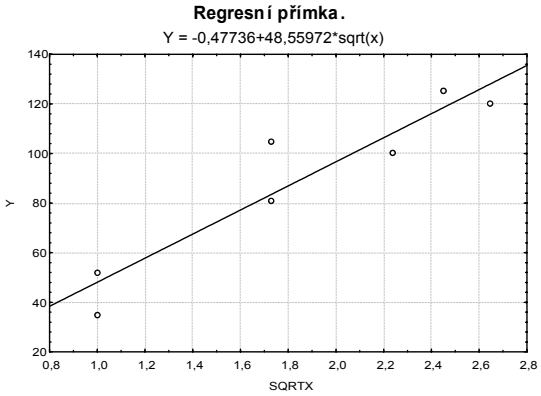


Model s odmocninou

Statistic	Summary Value
Multiple R	0,93924
Multiple R2	0,88217
Adjusted R2	0,85860
F(1,5)	37,43261
p	0,00169
Std.Err. of Estimate	12,82508

Variable	Predicting Values for (stroje) variable: Y		
	B-Weight	Value	B-Weight * Value
SQRTX	48,55972	2,000000	97,1194
Intercept			-0,4774
Predicted			96,6421
-95,0%CL			83,6962
+95,0%CL			109,5880

N=7	Regression Summary for Dependent Variable: Y (stroje) R= ,93923698 R2= ,88216611 Adjusted R2= ,85859933 F(1,5)=37,433 p<,00169 Std.Error of estimate: 12,825					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(5)	p-level
Intercept			-0,47736	15,29638	-0,031207	0,976312
SQRTX	0,939237	0,153515	48,55972	7,93690	6,118220	0,001691

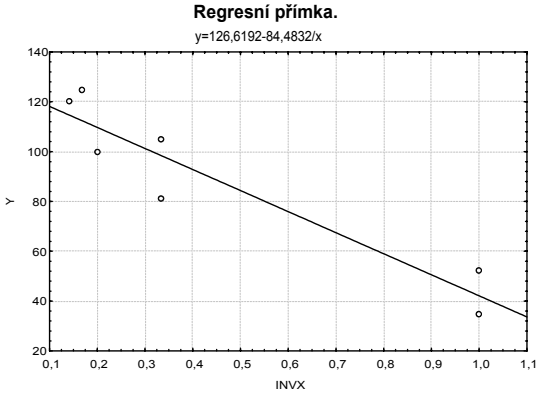


Model s převrácenou hodnotou

Statistic	Summary Value
Multiple R	0,94282
Multiple R2	0,88891
Adjusted R2	0,86670
F(1,5)	40,01016
p	0,00146
Std.Err. of Estimate	12,45245

Variable	Predicting Values for (stroje) variable: Y		
	B-Weight	Value	B-Weight * Value
INVX	-84,4832	0,250000	-21,1208
Intercept			126,6192
Predicted			105,4984
-95,0%CL			91,5231
+95,0%CL			119,4738

N=7	Regression Summary for Dependent Variable: Y (stroje) R= ,94282234 R2= ,88891396 Adjusted R2= ,86669676 F(1,5)=40,010 p<,00146 Std.Error of estimate: 12,452					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(5)	p-level
Intercept			126,6192	7,67327	16,50134	0,000015
INVX	-0,942822	0,149054	-84,4832	13,35627	-6,32536	0,001456

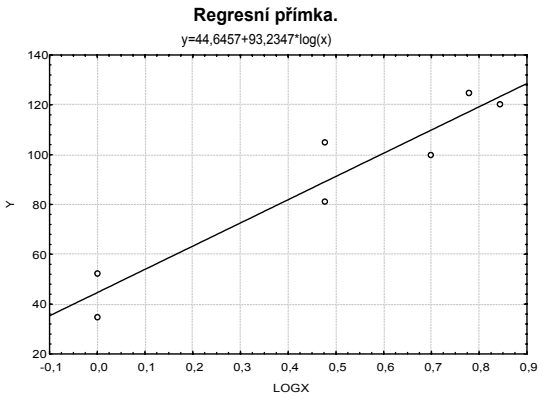


Model s logaritmem

Statistic	Summary Value
Multiple R	0,95349
Multiple R2	0,90915
Adjusted R2	0,89097
F(1,5)	50,03321
p	0,00087
Std.Err. of Estimate	11,26153

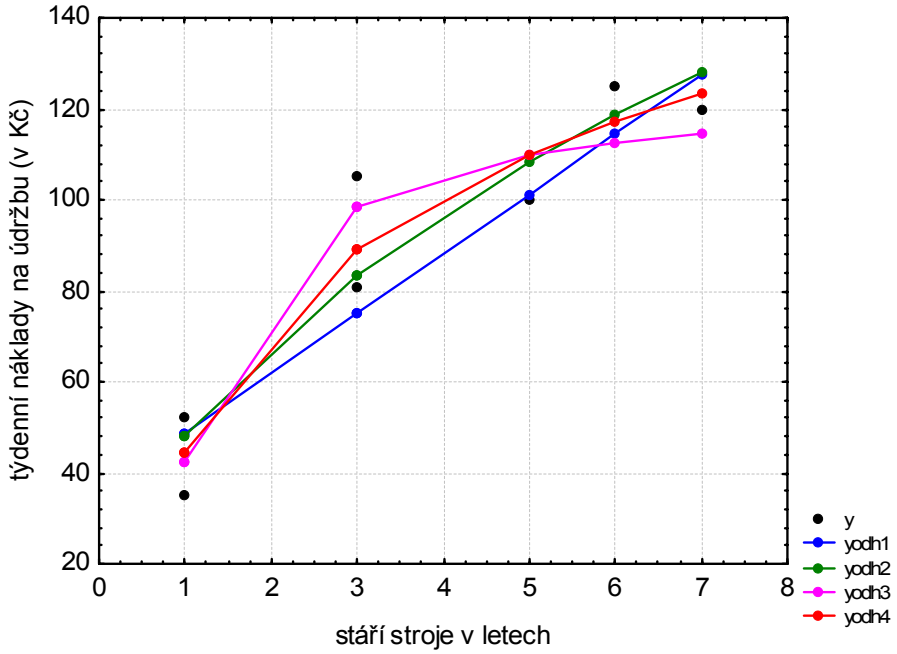
Variable	Predicting Values for (stroje) variable: Y		
	B-Weight	Value	B-Weight * Value
LOGX	93,23472	0,602060	56,1329
Intercept			44,6457
Predicted			100,7786
-95,0%CL			88,9325
+95,0%CL			112,6247

N=7	Regression Summary for Dependent Variable: Y (stroje) R= ,95349135 R2= ,90914576 Adjusted R2= ,89097491 F(1,5)=50,033 p<,00087 Std.Error of estimate: 11,262					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(5)	p-level
Intercept			44,64571	7,49541	5,956407	0,001907
LOGX	0,953491	0,134799	93,23472	13,18100	7,073415	0,000874



Nejvyšší hodnotu indexu determinace vykazuje model s logaritmem.

Výsledky všech čtyř modelů:



Komentář: Abychom získali graf s výsledky všech čtyř modelů, musíme datový soubor ve STATISTICE uspořádat podle hodnot proměnné X:

	1 x	2 y
1	1	35
2	1	52
3	3	81
4	3	105
5	5	100
6	6	125
7	7	120

K tomuto datovému souboru přidáme další čtyři proměnné yodh1, yodh2, yodh3 a yodh4. Do Long Name těchto proměnných postupně napíšeme =35,44+13,15*x, =-0,48+48,56*sqrt(x), =126,62-84,48/x, =44,65+93,23*log10(x). Dostaneme soubor:

	1 x	2 y	3 yodh1	4 yodh2	5 yodh3	6 yodh4
1	1	35	48,59	48,08	42,14	44,65
2	1	52	48,59	48,08	42,14	44,65
3	3	81	74,89	83,62839	98,46	89,13201
4	3	105	74,89	83,62839	98,46	89,13201
5	5	100	101,19	108,1035	109,724	109,815
6	6	125	114,34	118,4672	112,54	117,197
7	7	120	127,49	127,9977	114,5514	123,4385

a pomocí vícenásobného bodového grafu vytvoříme výše uvedený obrázek.