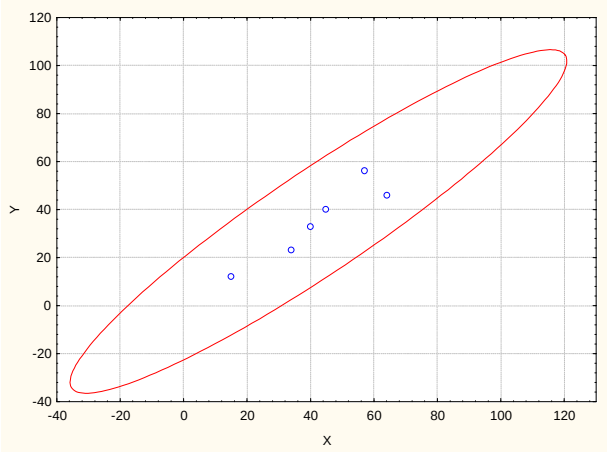
Cvičení 10: Korelační analýza

Příklad 1.: Zjišťovalo se, kolik mg kyseliny mléčné je ve 100 ml krve matek prvorodiček (veličina X) a u jejich novorozenců (veličina Y) těsně po porodu. Byly získány tyto výsledky:

Číslo matky	1	2	3	4	5	6
x_i	40	64	34	15	57	45
y_i	33	46	23	12	56	40

Nakreslete dvourozměrný tečkový diagram, vypočtěte výběrový korelační koeficient, sestrojte 95% interval spolehlivosti pro korelační koeficient a na hladině významnosti 0,05 testujte hypotézu o nezávislosti výsledků obou měření.
Návod: Otevřeme datový soubor kyselina_mlecna.sta. Obvyklým způsobem zobrazíme dvourozměrný tečkový diagram, s jehož pomocí posoudíme dvourozměrnou normalitu dat. Grafy – Bodové grafy – vypneme lineární proložení - Proměnné X, Y – OK – Detaily - Elipsa normální – OK. Ve vzniklém grafu upravíme měřítka na vodorovné a svislé ose:



Testování hypotézy o nezávislosti:
První možnost – pomocí testové statistiky T: Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměn. – X, Y – OK – na záložce Možnosti vybereme Zobrazit detailní tabulku výsledků – Výpočet.

Prom. X & prom. Y	Korelace (Tabulka3) 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	42,50000	17,39828									
Y	35,00000	15,89969	0,934832	0,873912	5,265339	0,006232	6	-1,30823	0,854311	6,696994	1,022943

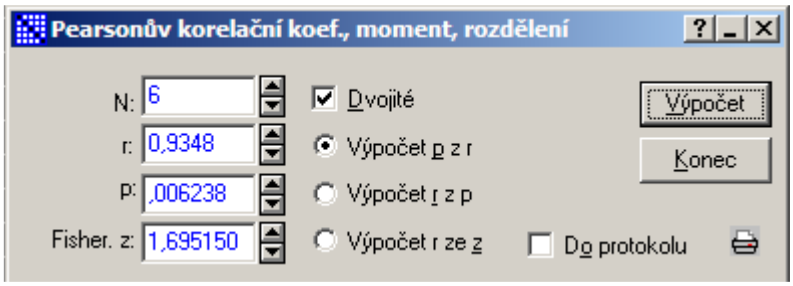
Ve výstupní tabulce je mj. hodnotu výběrového korelačního koeficientu R_{12} (r=0,9348, tzn. že mezi X a Y existuje silná přímá lineární závislost), hodnota testové statistiky (t = 5,2653) a p-hodnotu pro test hypotézy o nezávislosti (p=0,006232), H_0 tedy zamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že mezi oběma koncentracemi existuje závislost.

Druhá možnost – pomocí intervalu spolehlivosti pro ρ: Statistiky – Analýza síly testu – Odhad intervalu – Jedna korelace, t-test – OK – Pozorované R: -0,9348, N: 6, zaškrtneme Fisherovo Z (původ.) – Vypočítat.

	Odhad intervalu Jedna korelace, t-test
	Hodnota
Pozorovaný korel. koef. R	0,9348
Korelace dle nulové hypotézy (Ró0)	0,0000
Oboustranná p-hodnota	0,0033
Velikost vz. ve skup. (N)	6,0000
Interval spolehlivosti	0,9500
Meze spolehlivosti (Fisher. Z původní):	
Ró:	
Dolní mez	0,5106
Horní mez	0,9930

95% interval spolehlivosti pro ρ má tedy meze 0,5106 a 0,9930, nepokrývá hodnotu 0 a tudíž hypotézu o nezávislosti veličin X, Y zamítáme na hladině významnosti 0,05.

Třetí možnost – pomocí pravděpodobnostního kalkulátoru: Pokud známe výběrový koeficient korelace a rozsah výběru, můžeme test nezávislosti veličin X, Y provést pomocí Pravděpodobnostního kalkulátoru**.**
Statistiky – Pravděpodobnostní kalkulátor – Korelace – zadáme n a r, zaškrtneme Výpočet p z r – Výpočet.



Příklad k samostatnému řešení: V náhodném výběru 10 dvoučlenných domácností byl zjišťován měsíční příjem (veličina X, v tisících Kč) a vydání za potraviny (veličina Y, v tisících Kč).

x_i	15	21	34	35	39	42	58	64	75	90
y_i	3	4,5	6,5	6	7	8	9	8	9,5	10,5

Vypočtěte výběrový koeficient korelace. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti veličin X, Y. Sestrojte 95% asymptotický interval spolehlivosti pro ρ. (Data jsou uložena v souboru prijem_vydani.sta).
Výsledek: r_{12} = 0,9405, H_0 zamítáme na hladině významnosti 0,05, s pravděpodobností aspoň 0,95 platí: 0,7623 < ρ < 0,9862

Příklad 2.: V psychologickém výzkumu bylo vyšetřeno 426 hochů a 430 dívek. Ve skupině hochů činil výběrový koeficient korelace mezi verbální a performační složkou IQ 0,6033, ve skupině dívek činil 0,5833. Za předpokladu dvourozměrné normality dat testujte na hladině významnosti 0,05 hypotézu, že korelační koeficienty se neliší.

Výpočet pomocí systému STATISTICA:

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma korelačními koeficienty. Do políčka r1 napíšeme 0,6033, do políčka N1 napíšeme 426, do políčka r2 napíšeme 0,5833, do políčka N2 napíšeme 430 - Výpočet. Dostaneme p-hodnotu 0,6528, tedy nezamítáme nulovou hypotézu o shodě dvou koeficientů korelace na asymptotické hladině významnosti 0,05.

Příklad k samostatnému řešení: Načtěte datový soubor IQ.sta. Za předpokladu dvourozměrné normality dat (orientačně ověřte pomocí dvourozměrného tečkového diagramu) testujte na hladině významnosti 0,1 hypotézu, že korelační koeficienty mezi verbální a performační složkou IQ jsou stejné u dětí z města a venkova.
Výsledek: p = 0,0784, tedy s rizikem omylu nejvýše 10 % jsme prokázali, že korelační koeficienty se liší.

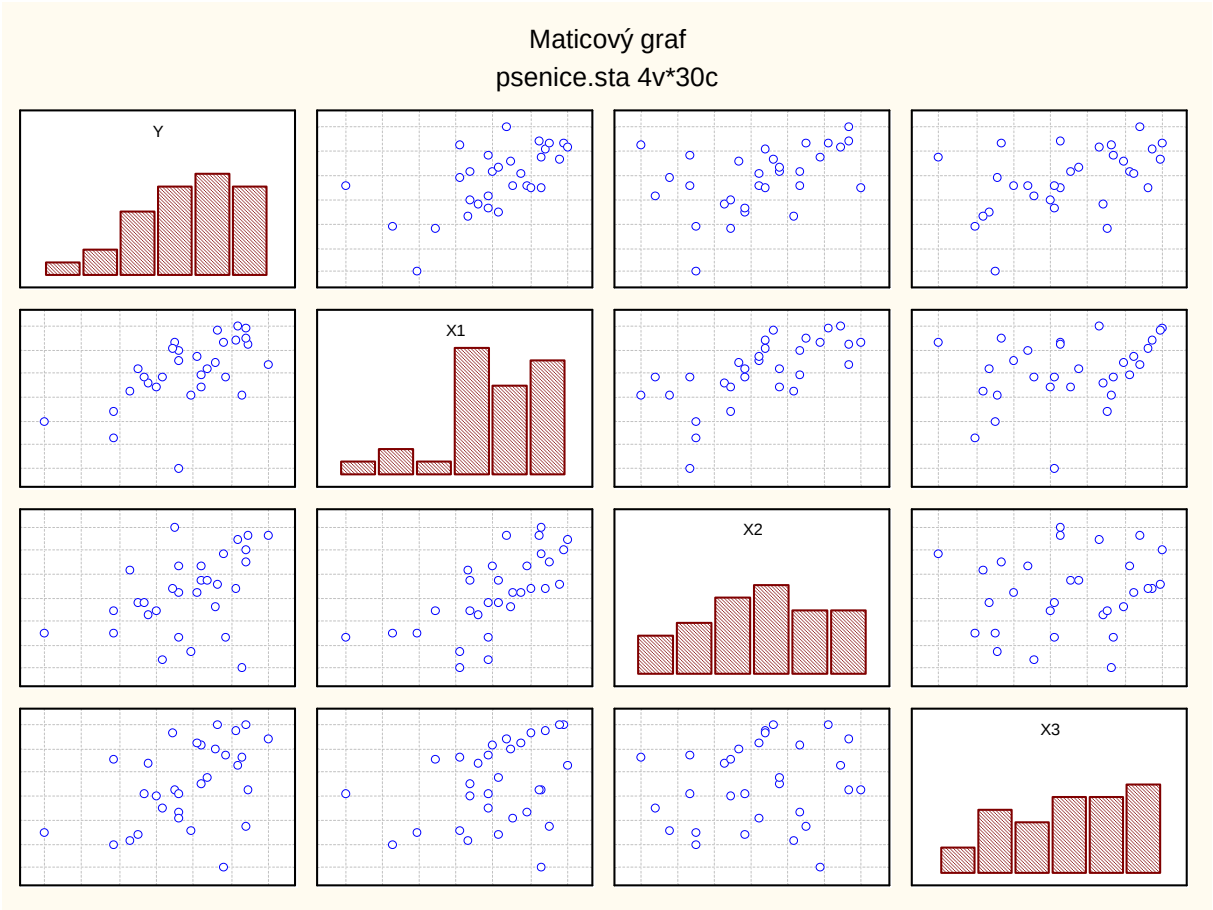
Příklad na mnohonásobnou a parciální korelaci (příklad je převzat ze skript Michálek Jaroslav, Osecký Pavel, Pešek Josef, Rod Jan, Vondráček Jiří: Biometrika, SNTL Praha 1982)
Během 30 let od roku 1913 do roku 1942 byly na 20 vybraných farmách ve Švédsku v oblasti Kalmar sledovány následující čtyři náhodné veličiny:
Y … průměrný výnos pšenice z podzimní setby (v kg/ha)
X_1 … průměrná teplota vzduchu během předchozí zimy (říjen – březen) v oblasti Kalmar (ve °C)
X_2 … průměrná teplota vzduchu během vegetačního období (duben – září) v oblasti Kalmar (ve °C)
X_3 … celkové srážky během vegetačního období, počítané jako průměr ze tří různých meteorologických stanic (v mm)
Data jsou uložena v souboru psenice.sta.

Úkol 1.: Měli bychom předpokládat, že náhodný vektor $(Y, X_1, X_2, X_3)'$ se řídí čtyřrozměrným normálním rozložením, tedy naše data jsou realizacemi náhodného výběru rozsahu 30 z tohoto normálního rozložení. Přesvědčíme se alespoň o jednorozměrné normalitě sledovaných proměnných.
Normalitu proměnných Y, X_1, X_2, X_3 posuďte Andersonovým - Darlingovým testem s hladinou významnosti 0,05.
Řešení:
A-D test je dostupný v modulu Rozdělení & simulace, Proložení dat rozděleními. Dostaneme tyto výsledky:

	AD stat.	AD p-hodn.
Y	0,322445	0,920115
X1	0,493906	0,751739
X2	0,185603	0,993464
X3	0,405985	0,841766

Na hladině významnosti 0,05 nelze ani v jednom případě zamítnout hypotézu o normalitě.

Úkol 2.: Pomocí dvourozměrných tečkových diagramů znázorněte závislost mezi všemi dvojicemi náhodných veličin..
Řešení:
Grafy – Maticové grafy – Proměnné – Vybrat vše – OK



Úkol 3.: Vypočtěte výběrové korelační koeficienty pro všechny dvojice náhodných veličin a na hladině významnosti 0,05 testujte hypotézy o nezávislosti. Najděte 95% intervaly spolehlivosti pro všech šest korelačních koeficientů.

Řešení:

Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměnných – Proměnné 1-4 – OK – na záložce Možnosti zaškrtneme Zobrazit r, úrovně p, počty N a zaškrtneme Zobrazit dlouhá jména proměnných – Výpočet

Proměnná	Korelace (psenice) Označ. korelace jsou významné na hlad. p < ,05000 N=30 (Celé případy vynechány u ChD) Y	X1	X2	X3
Y: výnos	1,0000	,5962	,4188	,4542
	p= ---	p=,001	p=,021	p=,012
X1: zimni teploty	,5962	1,0000	,6703	,3205
	p=,001	p= ---	p=,000	p=,084
X2: letni teploty	,4188	,6703	1,0000	,1370
	p=,021	p=,000	p= ---	p=,471
X3: srazky	,4542	,3205	,1370	1,0000
	p=,012	p=,084	p=,471	p= ---

Vidíme, že korelační koeficient mezi:

a) výnosem a zimní teplotou je 0,5962, p-hodnota je 0,001, tedy na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin Y a X_1;

b) výnosem a letní teplotou je 0,4188, p-hodnota je 0,021, tedy na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin Y a X_2;
c) výnosem a srážkami je 0,4542, p-hodnota je 0,012, tedy na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin Y a X_3;
d) zimní teplotou a letní teplotou je 0,6703, p-hodnota je 0,000, tedy na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin X_1 a X_2;
e) zimní teplotou a srážkami je 0,3205, p-hodnota je 0,084, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti veličin X_1 a X_3;
f) letní teplotou a srážkami je 0,137, p-hodnota je 0,471, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti veličin X_2 a X_3.

Meze intervalů spolehlivosti pro koeficienty korelace získáme pomocí modulu Analýza síly testu.
Např. koeficient korelace mezi výnosem a zimní teplotou se s pravděpodobností přibližně 0,95 nachází v intervalu (0,3; 0,79).

Úkol 4.: Vypočtěte všechny výběrové parciální korelační koeficienty mezi Y a ostatními proměnnými a na hladině významnosti 0,05 testujte hypotézy o jejich nevýznamnosti.
Řešení:
Postup ukážeme na výpočtu $r_{Y,X_1.X_2}$, tj. při zkoumání závislosti výnosu na zimních teplotách při vyloučení vlivu letních teplot a na výpočtu $r_{Y,X_2.X_1}$, tj. při zkoumání závislosti výnosu na letních teplotách při vyloučení vlivu zimních teplot.

Statistiky – Základní statistiky/tabulky – Korelační matice – OK – na záložce Možnosti zaškrtneme Zobrazit r, úrovně p, počty N a zaškrtneme Zobrazit dlouhá jména proměnných, na záložce Detaily zvolíme Parciální korelace – 1. seznam proměnných Y, X1, druhý seznam proměnných X2 – OK

	Parciální korelace (psenice.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=30 (Celé případy vynechány u ChD)	
Proměnná	Y	X1
Y: výnos	1,0000	,4682
	p= ---	p=,010
X1: zimni teploty	,4682	1,0000
	p=,010	p= ---

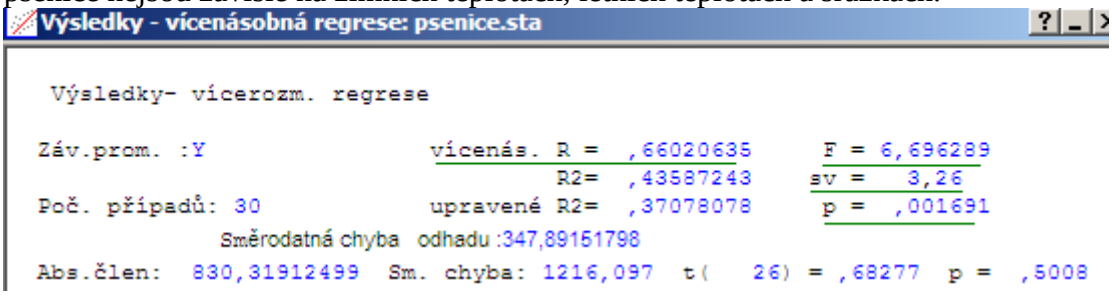
Vidíme, že výběrový parciální korelační koeficient $r_{Y,X_1.X_{21}}$ je 0,4682, p-hodnota je 0,01, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti $\rho_{Y,X_1.X_2}$.
Analogicky 1. seznam proměnných Y, X2, druhý seznam proměnných X1 – OK

	Parciální korelace (psenice.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=30 (Celé případy vynechány u ChD)	
Proměnná	Y	X2
Y: výnos	1,0000	,0322
	p= ---	p=,868
X2: letni teploty	,0322	1,0000
	p=,868	p= ---

V tomto případě výběrový parciální korelační koeficient $r_{Y,X_2.X_1}$ je 0,0322, p-hodnota je 0,868, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nevýznamnosti $\rho_{Y,X_2.X_1}$.

Interpretace: Výběrový korelační koeficient $r_{Y,X_1} = 0{,}5962$, což je podstatně více než $r_{Y,X_2} = 0{,}4188$. Mohlo by to znamenat, že vliv zimních teplot na výnosy pšenice je vyšší než vliv letních teplot. Pokud zkoumáme závislost Y na X_1 při vyloučení vlivu X_2, dostaneme výběrový parciální korelační koeficient 0,4682, což je poněkud nižší než 0,5962. Ovšem když zkoumáme závislost Y na X_2 při vyloučení vlivu X_1, dostaneme výběrový parciální korelační koeficient 0,0322, což je zcela nevýznamná korelace.
Stejným způsobem vypočteme a prozkoumáme další parciální korelační koeficienty. Pro kontrolu: $r_{Y,X_1.X_3} = 0{,}534$, p = 0,033, $r_{Y,X_2.X_3} = 0{,}4041$, p = 0,03, $r_{Y,X_3.X_1} = 0{,}346$, p = 0,066, $r_{Y,X_3.X_2} = 0{,}4412$, p = 0,017, $r_{Y,X_1.(X_2,X_3)} = 0{,}388$, p = 0,041, $r_{Y,X_2.(X_1,X_3)} = 0{,}0756$, p = 0,702, $r_{Y,X_3.(X_1,X_2)} = 0{,}3519$, p = 0,066.
Z těchto výsledků vyplývá, že na výnosy mají silný vliv zimní teploty a srážky, zatímco vliv letních teplot je způsoben silnou korelací mezi zimními a letními teplotami.

Úkol 5.: Vypočtěte výběrový koeficient mnohonásobné korelace mezi výnosy a ostatními proměnnými a na hladině významnosti 0,05 testujte hypotézu o jeho nevýznamnosti.
Řešení:
Statistiky – Vícenásobná regrese – Proměnné – Závislá proměnná Y, seznam nezáv. proměnných X1, X2, X3 – OK – OK.
Koeficient $r_{Y,(X_1,X_2,X_3)}$ najdeme v záhlaví výstupní tabulky pod označením Vícenás. R = 0,6602.
Hodnota testové statistiky pro test nevýznamnosti koeficientu mnohonásobné korelace $\rho_{Y,(X_1,X_2,X_3)}$ je 6,6963, počet stupňů volnosti čitatele je 3, jmenovatele 26, odpovídající p-hodnota je 0,001691, tedy na hladině významnosti 0,05 zamítáme hypotézu, že výnosy pšenice nejsou závislé na zimních teplotách, letních teplotách a srážkách.



Upozornění: Povšimněte si, že všechny výběrové párové korelační koeficienty veličiny Y s ostatními proměnnými jsou v absolutní hodnotě menší než výběrový koeficient mnohonásobné korelace: $r_{Y,X_1} = 0{,}5962$, $r_{Y,X_2} = 0{,}4188$, $r_{Y,X_3} = 0{,}4542$, zatímco $r_{Y,(X_1,X_3,X_3)} = 0{,}6602$.