Téma č. 6.: Jednoduchá, mnohonásobná a parciální korelace

Příklad: Výnosy pšenice (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)

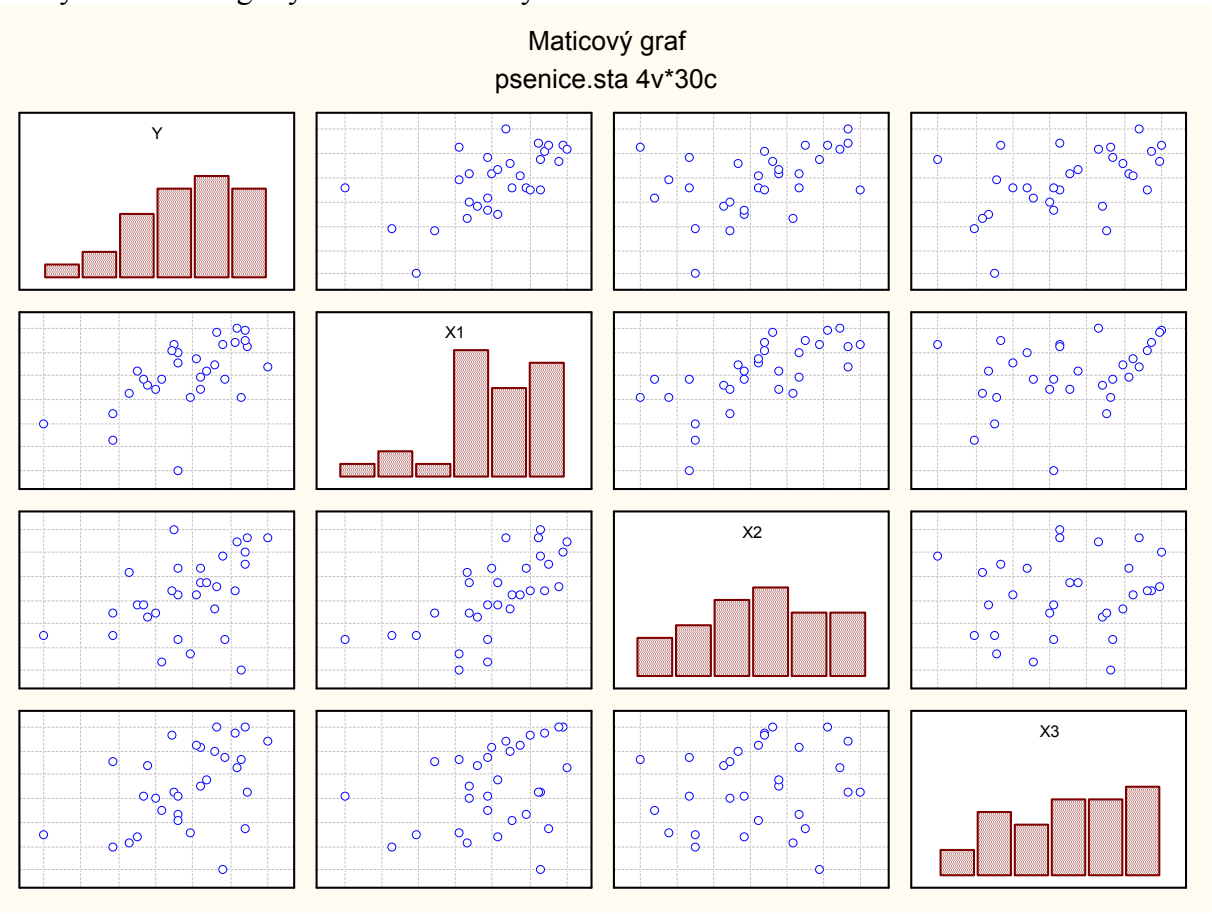
	1 Y	2 X1	3 X2	4 X3
1	1990	2,7	12,8	230
2	1950	3,1	13,7	268
3	1630	1,9	12	188
4	1720	1,3	11,7	315
5	1560	1	12,7	180
6	1680	1,6	12	261
7	1980	2,3	12,2	216
8	2180	1,7	12,8	346
9	2370	3,1	13,1	131
10	1790	1,1	11,8	256
11	2400	1,6	11,2	327
12	1410	0,1	11,8	320
13	2570	3,7	13,2	382
14	2180	1,1	12,5	279
15	2150	2,5	12,2	351
16	2530	0,8	10,5	324
17	2100	0,8	10,9	196
18	2330	3,6	12,4	381
19	1850	1,6	10,7	237
20	2230	1,9	12,5	289
21	2310	2,2	11,9	338
22	2600	3	13,5	267
23	2480	3,2	12,3	372
24	1940	2,8	12,3	367
25	2770	2,1	13,5	358
26	2570	3,3	12,9	202
27	2510	3,8	13,4	311
28	1420	-1,1	11,3	172
29	810	-0,4	11,3	194
30	1990	-2,4	11,2	261

Budeme 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.

Úkol 1.: Pomocí dvourozměrných tečkových diagramů znázorněte závislost mezi všemi dvojicemi náhodných veličin. Vypočtěte výběrové korelační koeficienty pro všechny dvojice náhodných veličin a na hladině významnosti 0,05 testujte hypotézy o nezávislosti.

Řešení:

Grafy – Maticové grafy – Proměnné – Vybrat vše – OK



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.

Úkol 2.: 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= ---

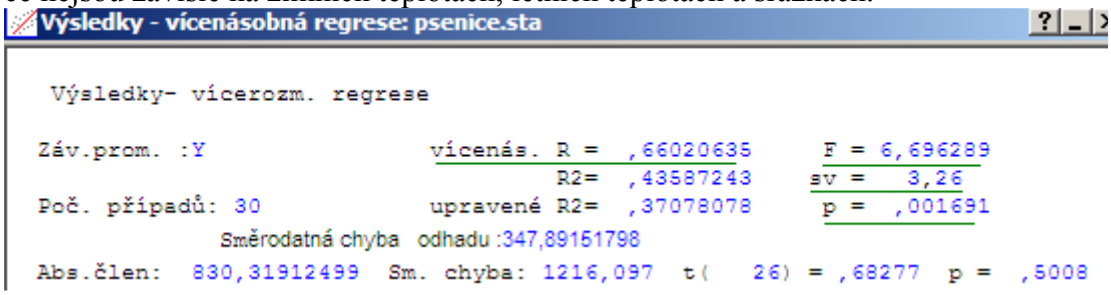
Vidíme, že výběrový parciální korelační koeficient $r_{Y,X_1.X_{21}}$ je 0,4682, p-hodnota je 0,01, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti $\rho_{Y,X_1.X_2}$.
Analogicky 1. seznam proměnných Y, X2, druhý seznam proměnných X1 – OK

	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 3.: 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 $r_{Y,(X_1,X_3,X_3)} = 0{,}6602$.

Příklad k samostatnému řešení
U sedmi vybraných domácností byly zjištěny tyto údaje: Y – výdaje za potraviny a nápoje za jeden měsíc (v tisících Kč), X_1 – počet členů, X_2 – celkový čistý příjem (v tisících Kč).

Y	4	3	4	1	6	4	5
X_1	4	2	4	1	5	3	4
X_2	20	18	22	13	25	22	23

a) Pomocí výběrového koeficientu mnohonásobné korelace posuďte na hladině významnosti 0,05, zda výdaje domácnosti závisí na počtu členů a celkovém čistém příjmu. Orientačně ověřte normalitu dat.
b) Vypočtěte výběrové korelační koeficienty $r_{YX_1}, r_{YX_2}, r_{X_1X_2}$ a na hladině významnosti 0,05 testujte hypotézu o nezávislosti každé dvojice proměnných. Dále vypočtěte parciální korelační koeficienty $r_{Y,X_1.X_2}, r_{Y,X_2.X_1}$, interpretujte je a testujte jejich významnost pro $\alpha = 0{,}05$.

Návod: Vytvořte datový soubor o třech proměnných Y, X_1, X_2 a sedmi případech. Normalitu proměnných Y, X_1, X_2 posuďte např. N-P plotem a S-W testem.

ad a) $r_{Y.\,\mathbf{X}} = 0{,}98268275$), testová statistika F nabývá hodnoty 56,25025, odpovídající p-hodnota je 0,001179, tedy na hladině významnosti 0,05 zamítáme hypotézu, že výdaje domácnosti nezávisí na počtu členů a příjmech.

ad b) $r_{YX_1} = 0{,}942837$ (čím má domácnost více členů, tím jsou vyšší výdaje za potraviny a nápoje), p-hodnota = 0,001455, $r_{YX_2} = 0{,}976274$ (čím jsou vyšší příjmy domácnosti, tím jsou vyšší výdaje za potraviny a nápoje), p-hodnota = 02000164, $r_{X_1X_2} = 0{,}921055$ (čím více členů domácnost má, tím jsou vyšší příjmy), p-hodnota = 0,003222. Ve všech třech případech je na hladině významnosti 0,05 prokázána závislost odpovídajících dvojic proměnných.
$r_{Y,X_1.X_2} = 0{,}517453$, tedy při eliminaci vlivu příjmu existuje mezi výdaji za potraviny a nápoje a počtem členů domácnosti středně silná přímá lineární závislost. Tato závislost však není prokazatelná na hladině významnosti 0,05, protože p-hodnota pro test hypotézy o nulovosti parciálního korelačního koeficientu je 0,293097. Výběrový parciální korelační koeficient $r_{Y,X_2.X_1}$ je 0,831168, tedy při eliminaci vlivu počtu členů domácnosti existuje mezi výdaji za potraviny a nápoje a příjmy domácnosti dosti silná přímá lineární závislost. Tato závislost je prokazatelná na hladině významnosti 0,05, protože p-hodnota pro test hypotézy o nulovosti parciálního korelačního koeficientu je 0, 040350.