

Cvičení 3.: Mnohonásobná a parciální korelace

Příklad 1.: 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)

X₁ ... průměrná teplota vzduchu během předchozí zimy (říjen – březen) v oblasti Kalmar (ve °C)

X₂ ... průměrná teplota vzduchu během vegetačního období (duben – září) v oblasti Kalmar (ve °C)

X₃ ... celkové srážky během vegetačního období, počítané jako průměr ze tří různých meteorologických stanic (v mm)

Data jsou uložena v souboru pšenice.sta.

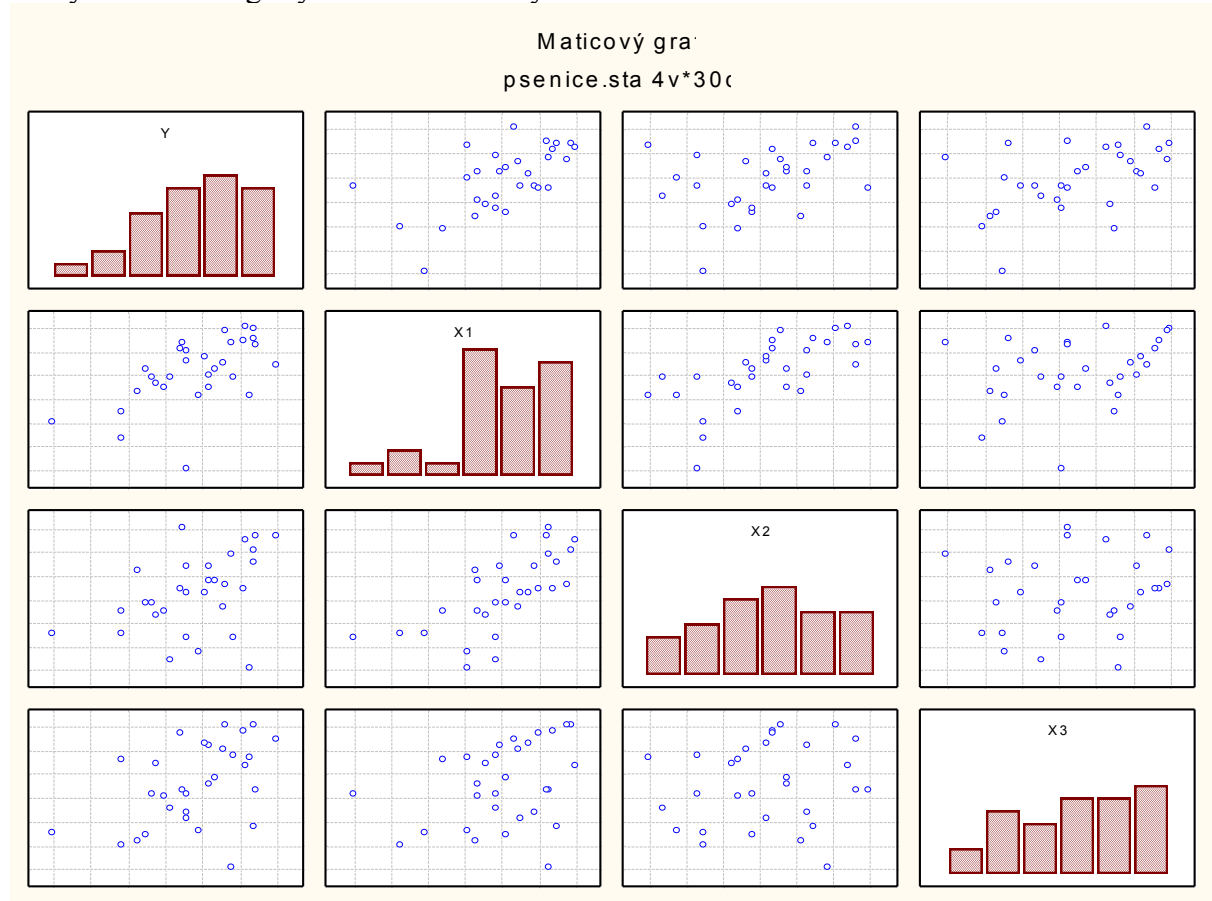
	1 Y	2 X ₁	3 X ₂	4 X ₃
1	1990	2,7	12,8	230
2	1950	3,1	13,7	268
3	1630	1,9	12	188
4	1720	1,3	11,7	315
5	1560	1	12,7	180
6	1680	1,6	12	261
7	1980	2,3	12,2	216
8	2180	1,7	12,8	346
9	2370	3,1	13,1	131
10	1790	1,1	11,8	256
11	2400	1,6	11,2	327
12	1410	0,1	11,8	320
13	2570	3,7	13,2	382
14	2180	1,1	12,5	279
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 znázorníte závislost mezi všemi dvojicemi náhodných veličin. Vypočítejte 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:

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, úroveň p, počty N a zaškrtneme Zobrazit dlouhá jména proměnných – Výpočet

Proměnná	Korelace (pšenice)			
	Y	X1	X2	X3
Y: výnos	1,0000	,5962	,4188	,4542
	p= --	p=,001	p=,021	p=,012
X1: zimní teploty	,5962	1,0000	,6703	,3205
	p=,001	p= --	p=,000	p=,084
X2: letní 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 X₁;
- 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 X₂;
- 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 X₃;
- 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 X₁ a X₂;
- e) zimní teplotou a srážkami je 0,3205, p-hodnota je 0,084, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti veličin X₁ a X₃;
- f) letní teplotou a srážkami je 0,137, p-hodnota je 0,471, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti veličin X₂ a X₃.

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 2.: Vypočítejte 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, úroveň 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 v y nechány u ChD)		
Proměnná	Y	X1
Y: výnos	1,0000	,4682
	p= --	p=,010
X1: zimní teploty	,4682	1,0000
	p=,010	p= --

Vidíme, že výběrový parciální korelační koeficient r_{Y,X_1,X_2} je 0,4682, p-hodnota je 0,01, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti ρ_{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 v y nechány u ChD)		
Proměnná	Y	X2
Y: výnos	1,0000	,0322
	p= --	p=,868
X2: letní 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 ρ_{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: $r_{Y, X_1, X_3} = 0,534$, $p = 0,033$, $r_{Y, X_2, X_3} = 0,4041$, $p = 0,03$, $r_{Y, X_3, X_1} = 0,346$, $p = 0,066$, $r_{Y, X_3, X_2} = 0,4412$, $p = 0,017$, $r_{Y, X_1, \langle X_2, X_3 \rangle} = 0,388$, $p = 0,041$, $r_{Y, X_2, \langle X_1, X_3 \rangle} = 0,0756$, $p = 0,702$, $r_{Y, X_3, \langle X_1, X_2 \rangle} = 0,3519$, $p = 0,066$.

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.: Vypočítejte 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 $r_{Y, \langle X_1, X_2, X_3 \rangle}$ najdeme v záhlaví výstupní tabulky pod označením Vícenás. R = 0,6602.

Hodnota testové statistiky pro test nevýznamnosti koeficientu mnohonásobné korelace $\rho_{Y, \langle X_1, X_2, X_3 \rangle}$ 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.

Výsledky - vícenásobná regrese: pšenice.sta			
Výsledky- vícerozm. regrese			
Záv.prom. :Y	vícenás. R = ,66020635	F = 6,696289	
	R2= ,43587243	sv = 3,26	
Poč. případů: 30	upravené R2= ,37078078	p = ,001691	
	Směrodatná chyba odhadu :347,89151798		
Abs. člen: 830,31912499	Sm. chyba: 1216,097	t(26) = ,68277	p = ,5008

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 $r_{Y, \langle X_1, X_2, X_3 \rangle} = 0,6602$.

Příklad 2.: U 19 vzorků potravinářské pšenice byl zjišťován obsah zinku v zrně (proměnná Y), v kořenech (proměnná X_1), v otrubách (X_2) a ve stonku a listech (X_3).

Y	X_1	X_2	X_3
175	164	198	162
169	160	198	159
175	158	211	164
181	162	211	162
539	520	567	523
526	502	540	491
344	339	355	334
475	460	500	446
820	683	813	695
841	731	832	714
828	710	846	697
775	716	818	709
622	543	635	563
661	577	712	580
579	505	596	531
936	790	946	814
903	806	946	834
927	793	912	824
889	820	919	807

a) Normalitu proměnných Y , X_1 , X_2 , X_3 posuďte pomocí K-S testu s hladinou významnosti 0,05.

b) Závislost mezi dvojicemi proměnných (Y, X_1) , (Y, X_2) , (Y, X_3) znázorněte dvourozměrnými tečkovými diagramy.

c) Vypočítejte výběrovou korelační matici všech čtyř proměnných a pro $\alpha = 0,05$ otestujte významnost jednotlivých korelačních koeficientů.

d) Vypočítejte výběrové parciální korelační koeficienty $r_{Y, X_1 \cdot \langle X_2, X_3 \rangle}$, $r_{Y, X_2 \cdot \langle X_1, X_3 \rangle}$, $r_{Y, X_3 \cdot \langle X_1, X_2 \rangle}$ a porovnejte je s výběrovými párovými korelačními koeficienty r_{YX_1} , r_{YX_2} , r_{YX_3} . Na hladině významnosti $\alpha = 0,05$ testujte hypotézy o nevýznamnosti parciálních korelačních koeficientů $\rho_{Y, X_1 \cdot \langle X_2, X_3 \rangle}$, $\rho_{Y, X_2 \cdot \langle X_1, X_3 \rangle}$, $\rho_{Y, X_3 \cdot \langle X_1, X_2 \rangle}$.

Řešení: Načteme datový soubor zinek.sta.

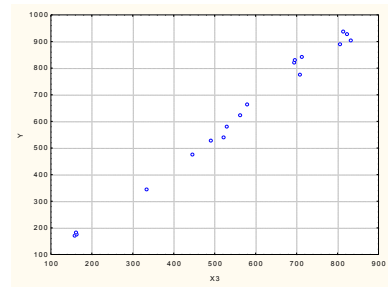
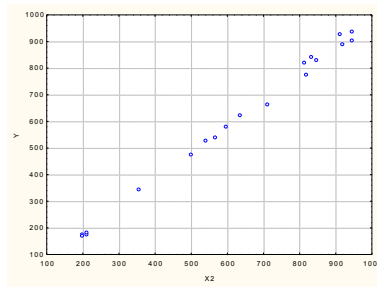
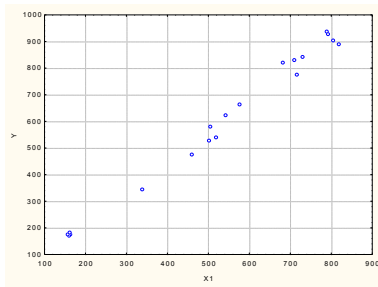
ad a) Výsledky K-S testu normality

proměnná	testová statistika	p-hodnota
Y	0,15792	> 0,2
X_1	0,15613	> 0,2
X_2	0,18177	< 0,1
X_3	0,16420	< 0,2

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

ad b)

Dvourozměrné tečkové diagramy dvojic (Y, X_1) , (Y, X_2) , (Y, X_3) svědčí o existenci dosti silné přímé lineární závislosti.



ad c) Výběrová korelační matice proměnných Y, X₁, X₂, X₃ spolu s odpovídajícími p-hodnotami:

Proměnná	Y	X1	X2	X3
Y	1,0000	,9947	,9981	,9959
	p= --	p=,000	p=0,00	p=0,00
X1	,9947	1,0000	,9954	,9980
	p=,000	p= --	p=,000	p=0,00
X2	,9981	,9954	1,0000	,9962
	p=0,00	p=,000	p= --	p=0,00
X3	,9959	,9980	,9962	1,0000
	p=0,00	p=0,00	p=0,00	p= --

Na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti jednotlivých korelačních koeficientů.

ad d)

Výběrový koeficient parciální korelace $r_{Y,X_1 \cdot \langle X_2, X_3 \rangle}$

Proměnná	Y	X1
Y	1,0000	-,0390
	p= --	p=,882
X1	-,0390	1,0000
	p=,882	p= --

Výběrový koeficient korelace r_{YX_1} je 0,9947, zatímco $r_{Y,X_1 \cdot \langle X_2, X_3 \rangle}$ je -0,039.

Pokud eliminujeme vliv proměnných X₂, X₃, tak mezi proměnnými Y a X₁ existuje velmi slabá nepřímá lineární závislost, která není na hladině 0,05 významná.

Výběrový koeficient parciální korelace $r_{Y,X_2 \cdot \langle X_1, X_3 \rangle}$

Proměnná	Y	X2
Y	1,0000	,7515
	p= ---	p=,001
X2	,7515	1,0000
	p=,001	p= ---

Výběrový koeficient korelace r_{YX_2} je 0,9981, zatímco $r_{Y,X_2 \cdot \langle X_1, X_3 \rangle}$ poklesl na 0,7515.

Pokud eliminujeme vliv proměnných X₁, X₃, tak mezi proměnnými Y a X₂ existuje silná přímá lineární závislost, která je na hladině 0,05 významná.

Výběrový koeficient parciální korelace $r_{Y,X_3 \cdot \langle X_1, X_2 \rangle}$

Proměnná	Y	X3
Y	1,0000	,2230
	p= ---	p=,390
X3	,2230	1,0000
	p=,390	p= ---

Výběrový koeficient korelace r_{YX_3} je 0,99589, zatímco $r_{Y,X_3 \cdot (X_1, X_2)}$ je pouze 0,223.

Pokud eliminujeme vliv proměnných X_1, X_2 , tak mezi proměnnými Y a X_3 existuje slabá přímá lineární závislost, která není na hladině 0,05 významná.

Vidíme, že existují značné rozdíly mezi párovými a parciálními výběrovými korelačními koeficienty.