Téma č. 6.: Mnohonásobná a parciální korelace, mnohonásobná regrese

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)
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

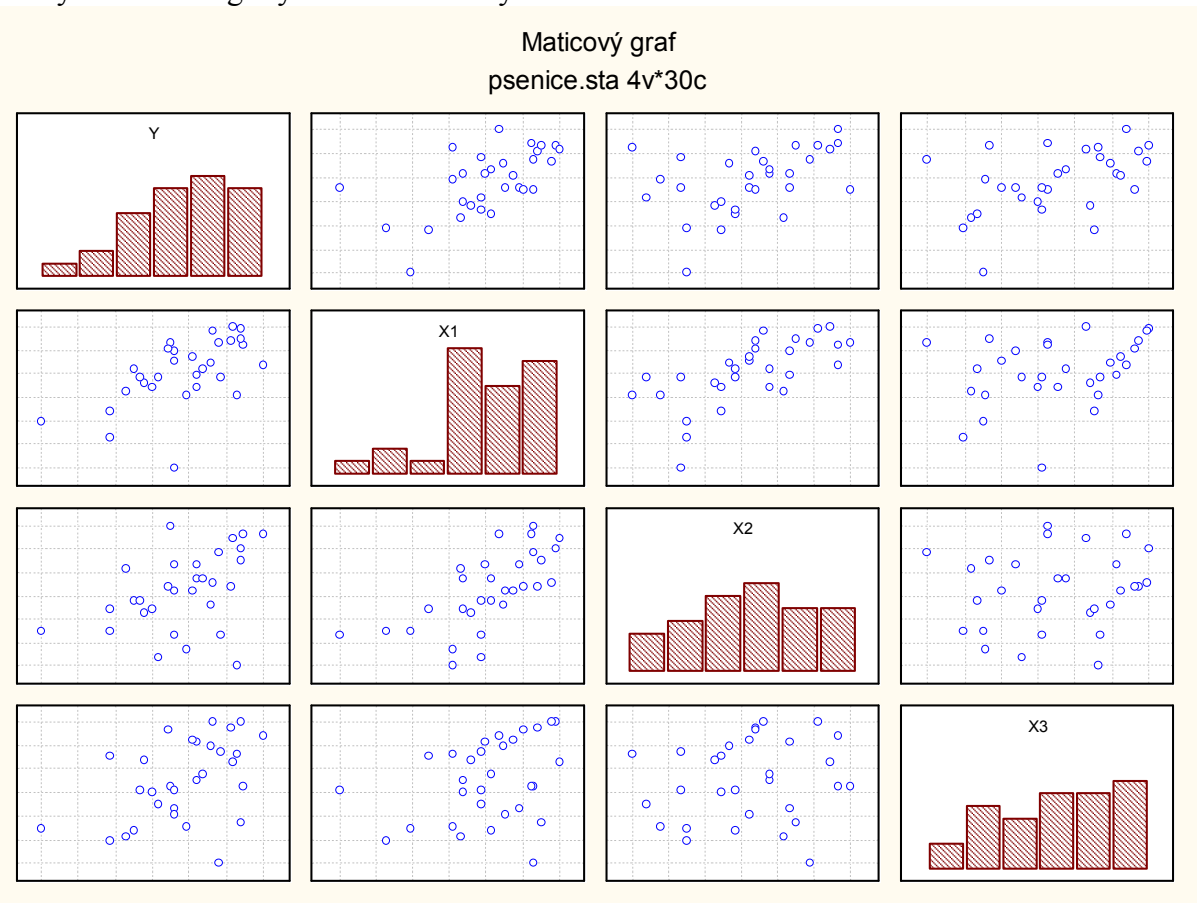
Budeme předpokládat, že náhodný vektor $(Y, X_1, X_2, X_3)'$ se řídí čtyřrozměrným normálním rozložením, tedy naše data jsou realizacemi náhodného výběru rozsahu 30 z tohoto normálního

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 < ,0500(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

Proměnná	Parciální korelace (psenice.sta) Označ. korelace jsou významné na hlad. p < ,0500(N=30 (Celé případy vynechány u ChD) 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

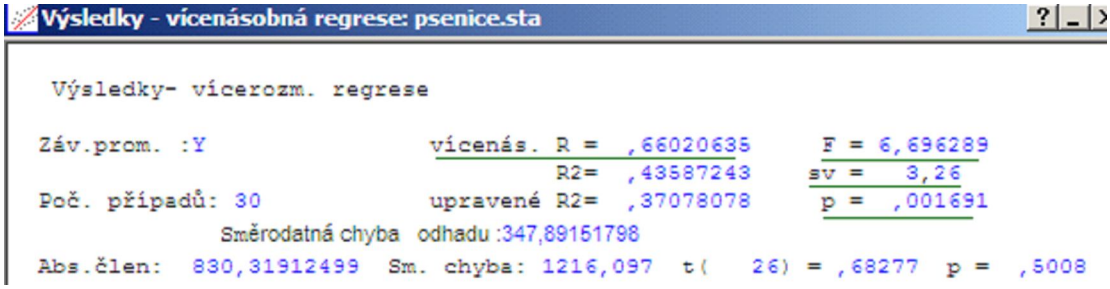
	Parciální korelace (psenice.sta) Označ. korelace jsou významné na hlad. p < ,0500(N=30 (Celé případy vynechány u ChD)	
Proměnná	Y	X2
Y: výnos	1,0000	,0322
	p= ---	p=,868
X2: letni teploty	,0322	1,0000
	p=,868	p= ---

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 2.: U 19 vzorků potravinářské pšenice byl zjišťován obsah zinku v zrnu (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) Vypočtěte výběrovou korelační matici všech čtyř proměnných a pro $\alpha = 0,05$ otestujte významnost jednotlivých korelačních koeficientů.

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

e) V první fázi zpracování předpokládejte, že je vhodný regresní model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$. Vypočtěte index determinace a interpretujte ho. Proveďte celkový F-test. Odhadněte parametry regresního modelu. Proveďte dílčí t-testy pro regresní koeficienty. Zjistěte odhad rozptylu. Vypočtěte parciální indexy determinace. (Hladinu významnosti volte $\alpha = 0,05$.)

f) Posuďte pomocí beta koeficientů vliv jednotlivých nezávisle proměnných veličin na regresní model.

g) Z regresního modelu odstraňte ty proměnné, jejichž regresní koeficienty se neprokázaly významné pro $\alpha = 0,05$. Sestavte nový regresní model a proveďte v něm tytéž úkoly jako v bodě e).

h) Normalitu reziduí v tomto novém regresním modelu posuďte K-S testem na hladině významnosti $\alpha = 0{,}05$.

i) V novém regresním modelu najděte 95% interval spolehlivosti pro teoretickou regresní funkci a 95% predikční interval.

Řešení: Načteme datový soubor zinek.sta.

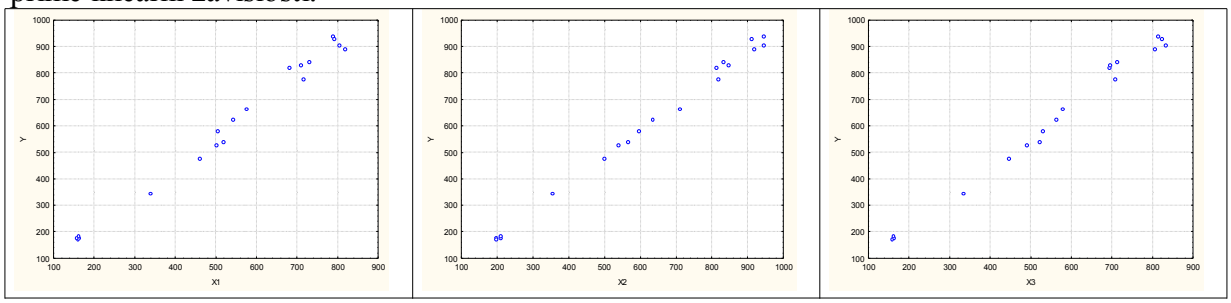
ad a) Výsledky K-S testu normality

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, X_1, X_2, X_3 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 $r_{Y,X_1.(X_2,X_3)}$

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 $r_{Y,X_1.(X_2,X_3)}$ je -0,039.
Pokud eliminujeme vliv proměnných X_2, X_3, tak mezi proměnnými Y a X_1 existuje velmi slabá nepřímá lineární závislost, která není na hladině 0,05 významná.

Výběrový koeficient parciální korelace $r_{Y,X_2.(X_1,X_3)}$

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 $r_{Y,X_2.(X_1,X_3)}$ poklesl na 0,7515.
Pokud eliminujeme vliv proměnných X_1, X_3, tak mezi proměnnými Y a X_2 existuje silná přímá lineární závislost, která je na hladině 0,05 významná.

Výběrový koeficient parciální korelace $r_{Y,X_3.(X_1,X_2)}$

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 $r_{Y,X_3.(X_1,X_2)}$ 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. Lze tedy soudit na existenci multikolinearity. O tom svědčí ikoeficienty VIF:

	Statistiky kolineace za daných podmínek (zinek.sta) Sigma-omezená parametrizace							
Efekt	Toler.	Rozptyl Infl fak	R^2	Y Beta v	Y Parciál.	Y Semipar.	Y t	Y p
X1	0,003802	262,9861	0,996198	-0,037425	-0,038960	-0,002308	-0,151006	0,881983
X2	0,007214	138,6290	0,992786	0,793836	0,751501	0,067422	4,411716	0,000505
X3	0,003120	320,5035	0,996880	0,242409	0,223005	0,013540	0,886006	0,389598

ad e) Výsledky pro regresní model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

	Výsledky regrese se závislou proměnnou : Y (zinek.sta) R= ,99824679 R2= ,99649665 Upravené R2= ,99579598 F(3,15)=1422,2 p<,00000 Směrod. chyba odhadu : 18,094					
N=19	Beta	Sm.chyba beta	B	Sm.chyba B	t(15)	Úroveň p
Abs.člen			-28,7607	10,60478	-2,71205	0,016066
X1	-0,037425	0,247835	-0,0439	0,29089	-0,15101	0,881983
X2	0,793836	0,179938	0,8079	0,18312	4,41172	0,000505
X3	0,242409	0,273598	0,2802	0,31623	0,88601	0,389598

Adjustovaný index determinace je 0,9958, tedy zvolený regresní model s proměnnými X_1, X_2, X_3 vysvětluje variabilitu proměnné Y z 99,58%. Testová statistika pro celkový F-test nabývá hodnoty

1422,2, odpovídající p-hodnota je velmi blízká 0, tedy model jako celek je významný na hladině 0,05.
Odhad rozptylu získáme z tabulky analýzy rozptylu:

Efekt	Součet čtverců	sv	Průměr čtverců	F	Úroveň p
Regres.	1396846	3	465615,2	1422,205	0,000000
Rezid.	4911	15	327,4		
Celk.	1401757				

$s^2 = 327,4$

Odhadnutá regresní funkce má tvar: $\hat{Y} = -28,7607 - 0,0439x_1 + 0,8079x_2 + 0,2802x_3$.
Dílčí t-testy pro jednotlivé regresní koeficienty:
testová statistika pro test hypotézy H_0: $\beta_0 = 0$ je -2,71205, p-hodnota je 0,016066, tedy H_0 zamítáme na hladině významnosti 0,05;
testová statistika pro test hypotézy H_0: $\beta_1 = 0$ je -0,15101, p-hodnota je 0,881983, tedy H_0 nezamítáme na hladině významnosti 0,05;
testová statistika pro test hypotézy H_0: $\beta_2 = 0$ je 4,41172, p-hodnota je 0,000505, tedy H_0 zamítáme na hladině významnosti 0,05;
testová statistika pro test hypotézy H_0: $\beta_3 = 0$ je 0,88601, p-hodnota je 0,389598, tedy H_0 nezamítáme na hladině významnosti 0,05.
Výpočet parciálních indexů determinace:
$r_{Y,X_1}^{\ 2} = 0,9947^2 = 0,9894$ (Pokud do modelu $Y = \beta_0 + \varepsilon$ zařadíme veličina X_1, pak bude vysvětlovat variabilitu hodnot veličiny Y z 98,94%.)
$r_{Y,X_2.X_1}^{\ 2} = 0,8079^2 = 0,6527$ (Pokud do modelu $Y = \beta_0 + \beta_1 x_1 + \varepsilon$ zařadíme veličinu X_2, pak bude vysvětlovat variabilitu hodnot veličiny Y z 65,27%.)
$r_{Y,X_3.(X_1,X_2)}^{\ 2} = 0,223^2 = 0,0497$ (Pokud do modelu $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ zařadíme veličinu X_3, pak bude vysvětlovat variabilitu hodnot veličiny Y z 4,97%.)

ad f) Interpretace beta koeficientů:
beta1 = -0,037425, beta2 = 0,793836, beta3 = 0,242409. V absolutní hodnotě je největší beta2, tedy obsah zinku v otrubách má největší vliv na obsah zinku v zrnu.

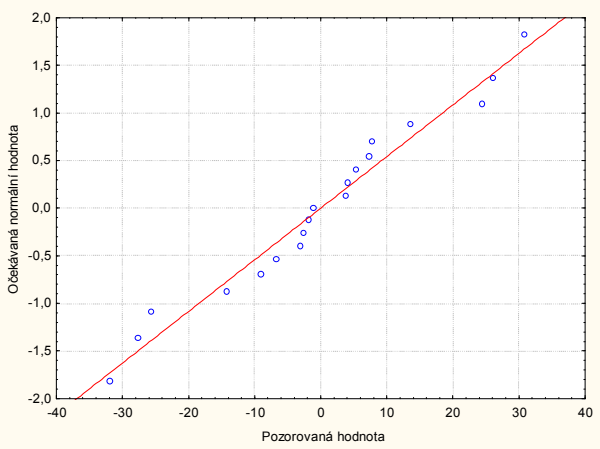
ad g) Protože dílčí t-testy prokázaly, že na hladině 0,05 nejsou proměnné X_1 a X_3 významné, sestavíme nový regresní model $Y = \beta_0 + \beta_2 x_2 + \varepsilon$.

N=19	Výsledky regrese se závislou proměnnou : Y (zinek.sta) R= ,99807615 R2= ,99615600 Upravené R2= ,99592988 F(1,17)=4405,5 p<0,0000 Směrod. chyba odhadu : 17,803					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(17)	Úroveň p
Abs.člen			-30,2507	10,31117	-2,93378	0,009274
X2	0,998076	0,015037	1,0157	0,01530	66,37372	0,000000

Adjustovaný index determinace je 0,9959, tedy zvolený regresní model s proměnnou X_2 vysvětluje variabilitu proměnné Y z 99,59%. Testová statistika pro celkový F-test nabývá hodnoty 4405,5, odpovídající p-hodnota je velmi blízká 0, tedy model jako celek je významný na hladině

0,05.
Vidíme, že $\widehat{Y} = -30{,}2507 + 1{,}0157x_2$.
Dílčí t-testy pro jednotlivé regresní koeficienty:
testová statistika pro test hypotézy H_0: $\beta_0 = 0$ je -2,93378, p-hodnota je 0,009274, tedy H_0 zamítáme na hladině významnosti 0,05;
testová statistika pro test hypotézy H_0: $\beta_2 = 0$ je 66,37372, p-hodnota je 0,000000, tedy H_0 zamítáme na hladině významnosti 0,05.

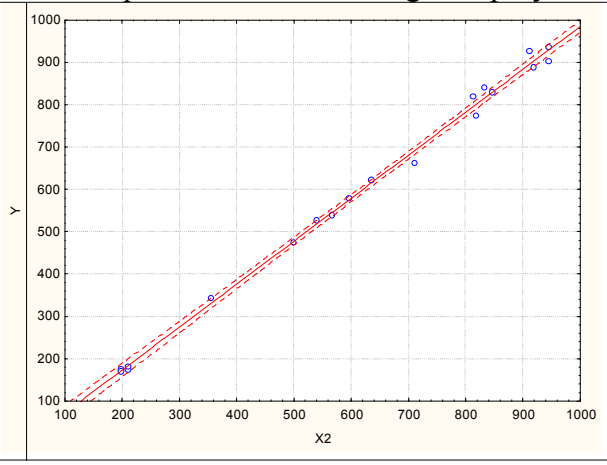
ad h) Ověření normality reziduí
Abychom mohli analyzovat rezidua, musíme je uložit. Ve výstupní tabulce zvolíme Rezidua/předpoklady/předpovědi - Reziduální analýza - Uložit - Uložit rezidua& předpovědi - OK.
Testová statistika pro K-S test nabývá hodnoty 0,1163, odpovídající p-hodnota je větší než 0,20, tedy hypotézu o normalitě reziduí nezamítáme na hladině významnosti 0,05.
Pro úplnost ještě posoudíme vzhled N-P plotu:



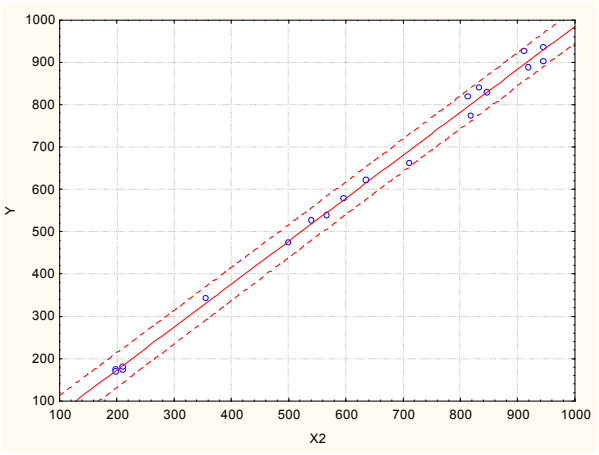
N-P plot svědčí o tom, že rozložení reziduí se příliš neliší od normálního rozložení.

ad i) Intervaly spolehlivosti pro regresní funkci a pro predikci získáme pomocí dvourozměrných tečkových diagramů, kde v Detailech vybereme lineární proložení a zvolíme regresní pásy.

95% interval spolehlivosti pro regresní funkci



95% interval spolehlivosti pro predikci


1000
900
800
700
600
500
400
300
200
100
Y
100
200
300
400
500
600
700
800
900
1000
X2