Cvičení 6.: Mnohonásobná lineární regrese

Příklad: 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 Lilieforsova varianty 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.

j) Proveďte regresi metodou STPEPVISE, a to jak Forward, tak Backward.

Řešení: Načteme datový soubor zinek.sta.

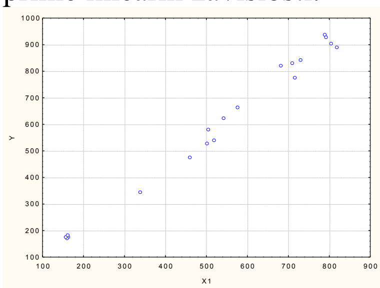
ad a) Výsledky Lilieforsova varianty K-S testu normality

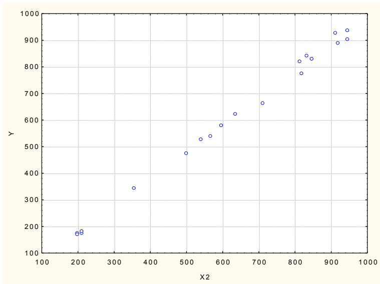
proměnná	testová statistika	p-hodnota
Y	0,15792	> 0,2
X_1	0,15613	> 0,2
X_2	0,18177	< 0,1
X_3	0,16420	< 0,2

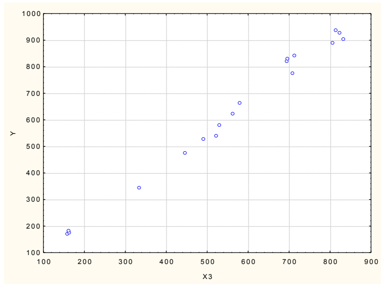
Na hladině významnosti 0,05 nelze ani v jednom případě zamítnout hypotézu o normalitě.

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%.)

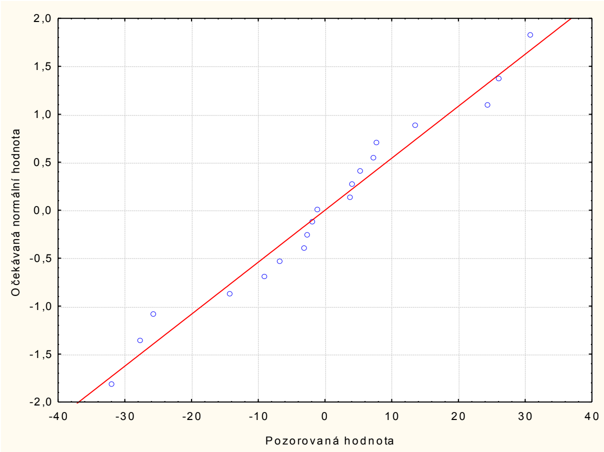
ad f) Interpretace beta koeficientů:
beta1 = -0,037425, beta2 = 0,793836, beta3 = 0,242409. V absolutní hodnotě je největší beta2, tedy obsah zinku v otrubách má největší vliv na obsah zinku v zrnu.

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

	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					
N=19	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 $\hat{Y} = -30{,}2507 + 1{,}0157 x_2$.
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.

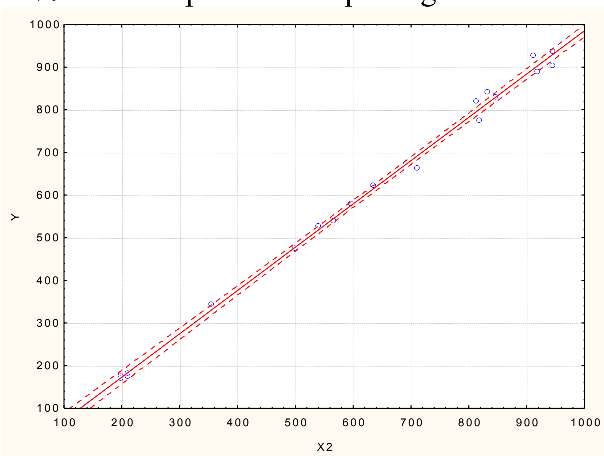
ad h) Ověření normality reziduí
Abychom mohli analyzovat rezidua, musíme je uložit. Ve výstupní tabulce zvolíme Rezidua/předpoklady/předpovědi – Reziduální analýza – Uložit – Uložit rezidua& předpovědi - OK.
Testová statistika pro K-S test nabývá hodnoty 0,1163, odpovídající p-hodnota je větší než 0,20,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



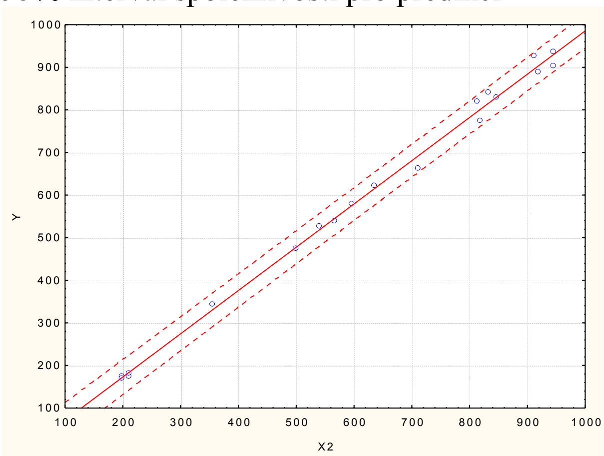
N-P plot svědčí o tom, že rozložení reziduí se příliš neliší od normálního rozložení.

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



ad j) Nejprve aplikujeme metodu Forward:
Statistiky – Vícerozměrná regrese – Proměnné – Závisle proměnná Y, Nezávisle proměnné X1, X2, X3 – OK – Detailní nastavení – zaškrtneme Další možnosti – OK – Metoda – zvolíme Kroková dopředná – na záložce Metoda zvolíme Zobrazit výsledky Po každém kroku – OK (V kroku 0 nejsou v regresní rovnici žádné proměnné.) Klikneme na Další – Výpočet:Výsledky regrese.

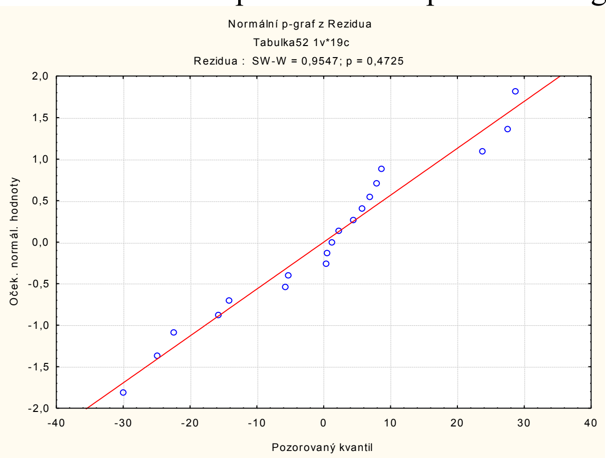
	Výsledky regrese se závislou proměnnou : Y (zinek.sta) R= ,99807615 R2= ,99615600 Upravené R2= ,99592988 F(1,17)=4405,5 p<0,0000 Směrod. chyba odhadu : 17,803					
N=19	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-30,2507	10,31117	-2,93378	0,009274
X2	0,998076	0,015037	1,0157	0,01530	66,37372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

V prvním kroku byla vybrána proměnná X2. Opět klikneme na Další a dostaneme výsledky kroku 2, který je již konečný:

N=19	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					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(15)	p-hodn.
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

Empirická regresní funkce má tvar $\hat{Y} = -28,9426 + 0,802x_2 + 0,2436x_4$.
Model jako celek je významný na hladině 0,05, avšak nezávisle proměnná X_3. Přispívá však k vysvětlení variability hodnot závisle proměnné veličiny Y. Adjustovaný index determinace je 0,9961. V modelu s nezávisle proměnnou X_2 byl 0,9959 a v modelu se všemi třemi nezávisle proměnnými byl 0,9958.

Normalitu reziduí prozkoumáme pomocí N-P grafu a S-W testu:



Rezidua neporušují předpoklad normality.

Nyní provedeme metodu Backward:
Statistiky – Vícerozměrná regrese – Proměnné – Závisle proměnná Y, Nezávisle proměnné X1, X2, X3 – OK – Detailní nastavení – zaškrtneme Další možnosti – OK – Metoda – zvolíme Kroková zpětná – na záložce Metoda zvolíme Zobrazit výsledky Po každém kroku – OK – Výpočet:Výsledky regrese.

N=19	Výsledky regrese se závislou proměnnou : Y (zinek.sta) R= ,99824412 R2= ,99649132 Upravené R2= ,99605274 F(2,16)=2272,1 p<0,0000 Směrod. chyba odhadu : 17,533					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(16)	p-hodn.
Abs.člen			-28,9426	10,20929	-2,83493	0,011948
X2	0,788109	0,170440	0,8020	0,17345	4,62396	0,000282
X3	0,210764	0,170440	0,2436	0,19700	1,23659	0,234086

V prvním kroku byly zařazeny všechny proměnné.

Klikneme na Další – Výpočet: Výsledky regrese.

N=19	Výsledky regrese se závislou proměnnou : Y (zinek.sta) R= ,99824412 R2= ,99649132 Upravené R2= ,99605274 F(2,16)=2272,1 p<0,0000 Směrod. chyba odhadu : 17,533					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(16)	p-hodn.
Abs.člen			-28,9426	10,20929	-2,83493	0,011948
X2	0,788109	0,170440	0,8020	0,17345	4,62396	0,000282
X3	0,210764	0,170440	0,2436	0,19700	1,23659	0,234086

V tomto kroku byly vyloučena proměnná X1.

Opět klikneme na Další – Výpočet: Výsledky regrese a dostaneme konečnou tabulku:

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					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-30,2507	10,31117	-2,93378	0,009274
X2	0,998076	0,015037	1,0157	0,01530	66,37372	0,000000

Vidíme, že metoda STEPWISE, Backward poskytla stejné výsledky jako metoda ENTER.