Cvičení 4.: Vícenásobná 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). Údaje jsou uvedeny v mg/kg.

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_l, X_2, X_3 posuďte pomocí Lilieforsova 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 koeficienty VIF a ukazatele tolerance pro vysvětlující proměnné X_l, X_2, X_3.
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. (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 Lilieforsovým 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 STEPWISE, a to jak Forward, tak Backward.

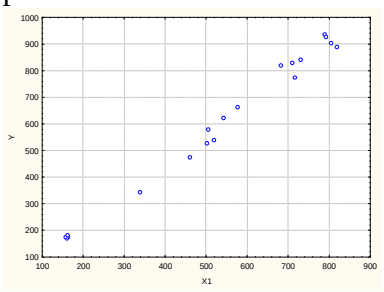
Řešení: Načteme datový soubor zinek.sta.

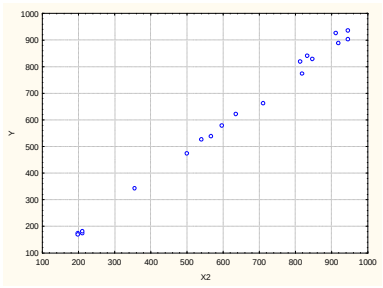
ad a) Výsledky Lilieforsova testu normality

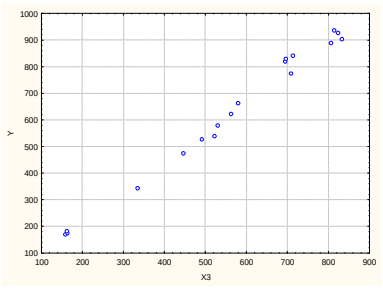
proměnná	testová statistika	p-hodnota
Y	0,15792	> 0,2
X_1	0,15613	> 0,2
X_2	0,18177	< 0,1
X_3	0,16420	< 0,2

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) Výpočet koeficientů VIF a ukazatelů tolerance:
Statistiky - Pokročilé lineární/nelineární modely – Obecné regresní modely – OK – Proměnné – Závislá Y, Spojité nezávisle proměnné X1, X2, X3 – OK – Matice – Parciální korelace.

	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

O existenci multikolinearity svědčí extrémně vysoké koeficienty VIF a velmi malé ukazatele tolerance.

ad e) Výsledky pro regresní model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ získáme takto: Statistiky – Vícenásobná regrese – Proměnné – Závislá proměnná Y, seznam nezáv. proměnných X1, X2, X3 – OK – OK.

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					
	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.

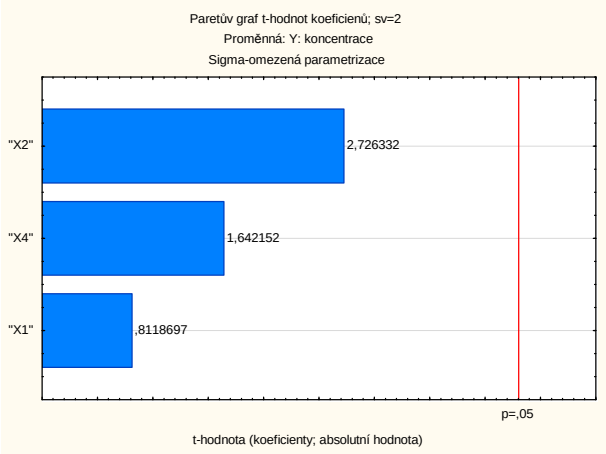
Odhad rozptylu získáme z tabulky analýzy rozptylu, kterou dostaneme pomocí cesty Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

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				

Vidíme, že $s^2 = 327,4$

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.

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.
Znázornění beta koeficientů pomocí Paretova diagramu:
Statistiky - Pokročilé lineární/nelineární modely – Obecné regresní modely – OK – Proměnné – Závislá Y, Spojité nezávisle proměnné X1, X2, X3 – OK – Paretův graf.



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;

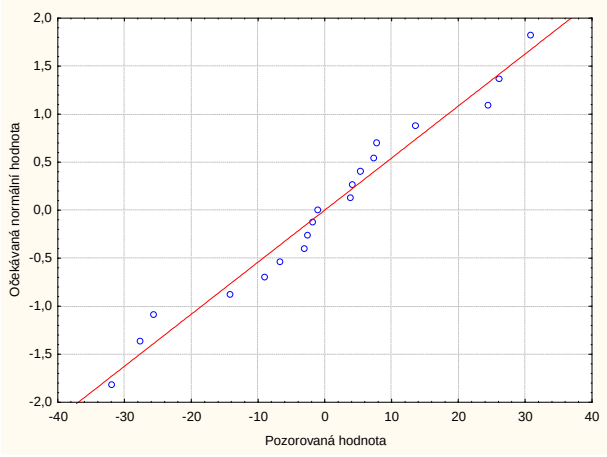
testová statistika pro test hypotézy H_0: $\beta_2 = 0$ je 66,37372, p-hodnota je 0,000000, tedy H_0 zamítáme na hladině významnosti 0,05.

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 Lilieforsův test nabývá hodnoty 0,1163, odpovídající p-hodnota je větší než 0,20, tedy hypotézu o normalitě reziduí nezamítáme na hladině významnosti 0,05.

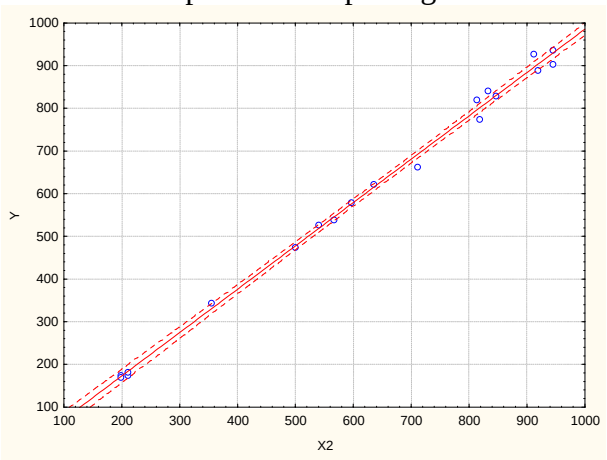
Pro úplnost ještě posoudíme vzhled N-P plotu:



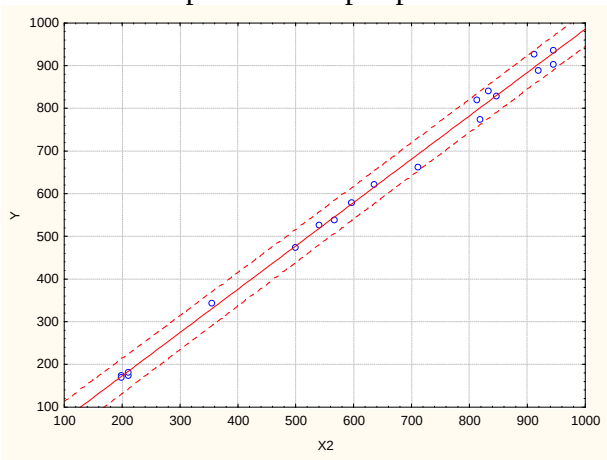
N-P plot svědčí o tom, že rozložení reziduí se příliš neliší od normálního rozložení.

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.

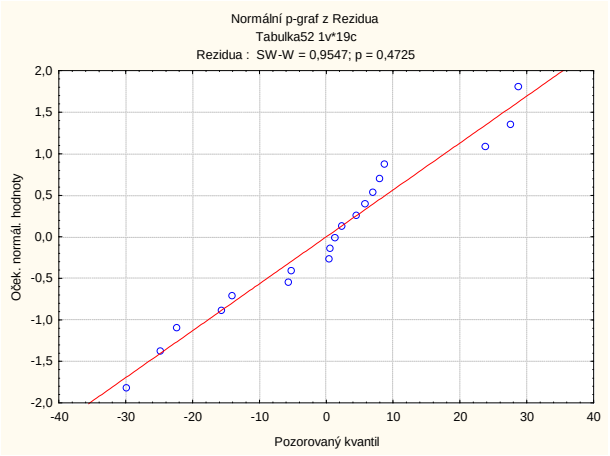
N=19	Výsledky regrese se závislou proměnnou : Y (zinek.sta) R= ,99807615 R2= ,99615600 Upravené R2= ,99592988 F(1,17)=4405,5 p<0,0000 Směrod. chyba odhadu : 17,803					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-30,2507	10,31117	-2,93378	0,009274
X2	0,998076	0,015037	1,0157	0,01530	66,37372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

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= ,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

Empirická regresní funkce má tvar $\widehat{Y} = -28{,}9426 + 0{,}802x_2 + 0{,}2436x_3$.
Model jako celek je významný na hladině 0,05, avšak nezávisle proměnná X_3 významná není. 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= ,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

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 byla 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.

Příklad k samostatnému řešení:
Byla změřena výška 20 osmnáctiletých chlapců (proměnná Y) a dále byly zjištěny výšky jejich příbuzných ve věku 18 let: X1 … výška matky, X2 … výška otce, X3 … výška babičky z matčiny strany, X4 … výška dědečka z otcovy strany, X5 … výška babičky z otcovy strany, X6 … výška dědečka z otcovy strany, X7 … výška chlapce při narození.
Data jsou uložena v souboru vysky_pribuznych.sta.
Nejprve proveďte korelační analýzu: vypočtěte koeficienty korelace proměnné Y se všemi nezávisle proměnnými.

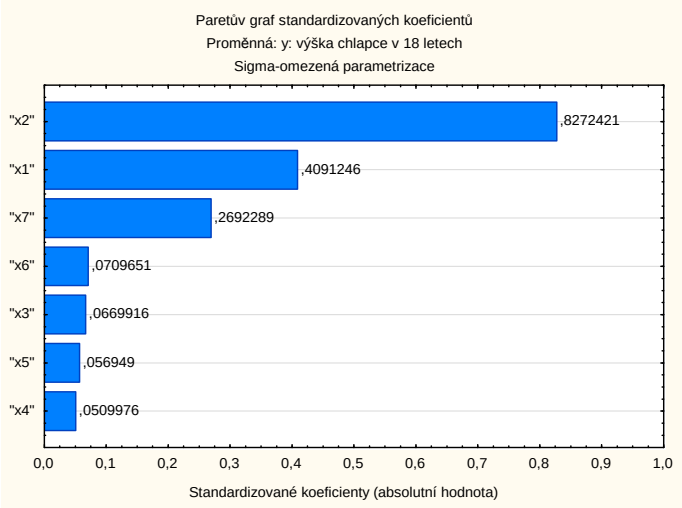
Proměnná	Korelace (vysky_pribuznych.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=20 (Celé případy vynechány u ChD)						
	x1	x2	x3	x4	x5	x6	x7
y	0,395117	0,791297	-0,211974	-0,260861	0,193163	-0,230270	0,264300

Pomocí koeficientů VIF prověřte, zda v modelu, který vysvětluje proměnnou Y pomocí proměnných X1 až X7, mezi nezávisle proměnnými veličinami existuje multikolinearita.

Efekt	Statistiky kolineace za daných podmínek (vysky_pribuznych.sta) Sigma-omezená parametrizace							
	Toler.	Rozptyl Infl fak	R^2	y Beta v	y Parciál.	y Semipar.	y t	y p
"x1"	0,6093157	1,6411853	0,3906843	0,4091246	0,7398021	0,3193573	3,8089422	0,0024890
"x2"	0,6095369	1,6405897	0,3904631	0,8272421	0,9120212	0,6458515	7,7030062	0,0000055
"x3"	0,8091856	1,2358105	0,1908144	-0,0669916	-0,2031556	-0,0602621	-0,7187401	0,4860614
"x4"	0,8192387	1,2206455	0,1807613	0,0509976	0,1569551	0,0461588	0,5505319	0,5920558
"x5"	0,6902695	1,4487094	0,3097305	0,0569490	0,1607847	0,0473146	0,5643166	0,5829329
"x6"	0,6513272	1,5353267	0,3486728	-0,0709651	-0,1934629	-0,0572722	-0,6830803	0,5075302
"x7"	0,6796494	1,4713469	0,3203506	0,2692289	0,6071912	0,2219546	2,6472305	0,0212876

V uvedeném regresním modelu posuďte pomocí beta koeficientů vliv jednotlivých nezávisle proměnných na Y. Použijte také Paretův diagram.

N=20	b*
Abs.člen	
x1	0,409125
x2	0,827242
x3	-0,066992
x4	0,050998
x5	0,056949
x6	-0,070965
x7	0,269229



Nyní pro výstavbu modelu použijte dopřednou i zpětnou krokovou metodu a jejich výsledky porovnejte.

Dopředná metoda:

N=20	Výsledky regrese se závislou proměnnou : y (vysky_pribuznych.sta) R= ,95111334 R2= ,90461658 Upravené R2= ,88673219 F(3,16)=50,581 p<,00000 Směrod. chyba odhadu : 2,3719					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(16)	p-hodn.
Abs.člen			-199,722	33,79003	-5,91069	0,000022
x2	0,873161	0,078496	1,106	0,09942	11,12369	0,000000
x1	0,364507	0,085585	0,688	0,16156	4,25898	0,000600
x7	0,263581	0,086598	1,373	0,45103	3,04372	0,007740

Zpětná metoda:

N=20	Výsledky regrese se závislou proměnnou : y (vysky_pribuznych.sta) R= ,92162267 R2= ,84938835 Upravené R2= ,83166934 F(2,17)=47,937 p<,00000 Směrod. chyba odhadu : 2,8915					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-156,313	37,34335	-4,18582	0,000620
x1	0,474630	0,094553	0,896	0,17849	5,01972	0,000105
x2	0,836417	0,094553	1,059	0,11976	8,84599	0,000000

Mezi těmito dvěma modely rozhodněte na základě reziduální analýzy a adjustovaného indexu determinace.