9. Mnohonásobná lineární regrese

Popis modelu mnohonásobné lineární regrese

Budeme zkoumat lineární závislost veličiny Y na p nezávisle proměnných veličinách $x_1, ..., x_p$. Omezíme se pouze na model tvaru

$$Y_i = \beta_0 + \beta_1 x_{il} + ... + \beta_p x_{ip} + \varepsilon_i,\ i = 1, ..., n.$$

Parametr β_0 interpretujeme jako teoretickou hodnotu závisle proměnné veličiny při nulových hodnotách všech nezávisle proměnných veličin. Parametr β_i, j = l, ..., p interpretujeme jako přírůstek teoretické hodnoty závisle proměnné veličiny odpovídající jednotkové změně j-té nezávisle proměnné veličiny při konstantní úrovni ostatních nezávisle proměnných.
Geometricky tento model představuje regresní nadrovinu. Lze ho formálně ztotožnit s lineárním regresním modelem z kapitoly „Jednoduchá lineární regrese“, kde položíme $f_1(x_i) = x_{i1}, ..., f_p(x_i) = x_{ip}$, i = 1,..., n. Dostáváme tedy maticový tvar $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, kde regresní matice

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix},$$ přičemž $h(\mathbf{X}) = p+1 < n$ a $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\mathbf{I})$.

Všechny výsledky uvedené v kapitole „Jednoduchá lineární regrese“ zůstávají v platnosti.

Míra lineární závislosti veličiny Y na veličinách $x_1, ..., x_p$

Jak bylo uvedeno v kapitole „Jednoduchá, mnohonásobná a parciální korelace“, mírou těsnosti lineární závislosti náhodné veličiny Y na vektoru $\mathbf{X} = (X_1,..., X_p)'$ je koeficient mnohonásobné korelace $\rho_{Y.\mathbf{X}}$:
$\rho_{Y.\mathbf{X}}^2 = \mathrm{cor}(Y, \mathbf{X})\ \mathrm{cor}(\mathbf{X})^{-1}\ \mathrm{cor}(\mathbf{X}, Y)$, kde
$\mathrm{cor}(Y, \mathbf{X})$ je korelační matice veličiny Y a vektoru $\mathbf{X}$ (v tomto případě se redukuje na vektor $(\rho_{YX_1},...,\rho_{YX_p})$),
$\mathrm{cor}(\mathbf{X})$ je korelační matice vektoru $\mathbf{X}$.
Výběrovým protějškem koeficientu $\rho_{Y.\mathbf{X}}$ je výběrový koeficient mnohonásobné korelace $r_{Y.\mathbf{X}}$:
$r_{Y.\mathbf{X}}^2 = R_{Y\mathbf{X}}\ \mathbf{R}^{-1}\ R_{\mathbf{X}Y}$, kde
$R_{Y\mathbf{X}}$ je výběrová korelační matice veličiny Y a vektoru $\mathbf{X}$ (v tomto případě se redukuje na vektor $(r_{YX_1},...,r_{YX_p})$),
$\mathbf{R}$ je výběrová korelační matice vektoru $\mathbf{X}$.
V regresním modelu se mu říká index korelace. (V případě regresní přímky se jedná o obyčejný párový koeficient korelace r_{YX}.) Jeho kvadrát odpovídá inde-

xu determinace v regresním modelu $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. Formálně je tedy celkový F-test rovnocenný s testem o nulové hodnotě koeficientu mnohonásobné korelace. Stojí za zmínku, že vypočtená hodnota testové statistiky F by měla být aspoň 4x větší než příslušný kvantil Fisherova - Snedecorova rozložení, aby bylo možné prohlásit zvolený regresní model za skutečně kvalitní.

Posouzení vlivu jednotlivých nezávisle proměnných v modelu

Chceme-li porovnávat vliv, jaký mají proměnné $x_1, ..., x_p$ v modelu $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, můžeme spočítat tzv. standardizované regresní parametry, kterým se také říká B-koeficienty. Zavedeme proto standardizované veličiny

$$Z_i = \frac{Y_i - m_Y}{s_Y}, v_{ij} = \frac{x_{ij} - m_{x_j}}{s_{x_j}}, j = 1, ..., p, i = 1, ..., n$$

a vytvoříme regresní model s těmito standardizovanými proměnnými. Odhady regresních parametrů v tomto novém modelu jsou B-koeficienty, které pak vyjadřují intenzitu vlivu jednotlivých nezávisle proměnných veličin na veličinu Y.

Použití parciálních korelačních koeficientů v modelu mnohonásobné lineární regrese

Uvažme model $Y_i = \beta_0 + \beta_1 x_{il} + ... + \beta_p x_{ip} + \varepsilon_i$, i = 1, ..., n. Druhá mocnina výběrového parciálního korelačního koeficientu $r_{Y,x_j.(x_1...x_{j-1})}$, j = 2, ..., p se nazývá parciální index determinace. Lze ho interpretovat jako „čistý" přínos proměnné x_j do modelu, který obsahoval proměnné $x_1, ..., x_{j-1}$. Čím větší je závislost mezi x_j a $(x_1, ..., x_{j-1})'$, tím menší se tento "čistý" přínos ukáže.
Výběrový parciální korelační koeficient $r_{Y,x_j.(x_1...x_{j-1})}$ měří „čistou“ korelaci mezi Y a X_j, když se eliminuje vliv náhodného vektoru $(X_1, \ldots, X_{j-1})$.
Protože v klasickém modelu lineární regrese je $S_T = S_R + S_E$, je pokles reziduálního součtu čtverců při zařazení nové proměnné do modelu roven růstu regresního součtu čtverců a naopak. Vzhledem k dříve zařazeným proměnným je tedy parciální index determinace mírou relativního zvýšení regresního součtu čtverců (poklesu reziduálního součtu čtverců) v důsledku zařazení nové proměnné.

Multikolinearita v modelu mnohonásobné regrese

O multikolinearitě hovoříme tehdy, když mezi některými sloupci regresní matice existuje silná lineární závislost, což svědčí o tom, že regresní model obsahuje nadbytečné vysvětlující proměnné.
Důsledky multikolinearity: matice **X'X** je blízká singulární matici => kvalita odhadu **b** je nízká => rozptyly odhadů $b_0, b_1, ..., b_p$ jsou velké => intervaly spolehlivosti pro $\beta_0, \beta_1, ..., \beta_p$ jsou široké.
Signály upozorňující na existenci multikolinearity:

- vysoké absolutní hodnoty výběrových korelačních koeficientů nezávisle proměnných (orientačně > 0,75)
- celkový F-test je významný, ale dílčí t-testy nikoliv.

Odstranění multikolinearity: do modelu se zařadí jen ty proměnné, které významně zlepšují odhad regresních parametrů. Jednou z metod výběru nejlepší podmnožiny proměnných je step-wise regression (postupná regrese). Úkolem postupné regrese je najít ty prediktory, které co nejlépe vystihují variabilitu závisle proměnné veličiny a získat odhady parametrů lineární regresní funkce, s jejíž pomocí pak lze uspokojivě predikovat hodnoty závisle proměnné veličiny.
Postupná regrese se používá ve dvou variantách – dopředná (forward) a zpětná (backward). Při metodě forward se prediktory postupně přidávají, při metodě backward se nejdříve zařadí všechny prediktory a pak se postupně odebírají. Princip postupné regrese spočívá v tom, že regresní model je budován krok po kroku tak, že v každém kroku zkoumáme všechny prediktory a zjišťujeme, který z nich nejlépe vystihuje variabilitu závisle proměnné veličiny. Zařazování prediktoru do modelu či jeho vylučování se děje pomocí sekvenčních F-testů. Sekvenční F-test je založen na statistice F, která je podílem přírůstku regresního součtu čtverců při zařazení daného prediktoru do modelu a reziduálního součtu čtverců. Jestliže je tato statistika větší než hodnota zvaná „F to enter“ (česky „F na zahrnutí“, ve STATISTICE implicitně 1, v SPSS 3,84), je prediktor zařazen. Je-li statistika F menší než hodnota zvaná „F to remove“ (česky „F na vyjmutí“, ve STATISTICE implicitně 0, v SPSS 2,71), je již dříve zařazený prediktor z modelu vyloučen. Po vybrání proměnných do modelu jsou odhadnuty parametry lineární regresní funkce a kvalita regrese je posouzena indexem determinace. Do modelu se postupně přidávají další proměnné tak dlouho, pokud se zvyšuje podíl vysvětlené variability hodnot veličiny Y.

Algoritmus postupné regrese:

1. krok: Vypočteme výběrové korelační koeficienty mezi závisle proměnnou Y a regresory $x_1, ..., x_p$. Do modelu vybereme ten regresor x_i, pro který je absolutní hodnota korelačního koeficientu největší.

2. krok: Sestavíme model $Y = \beta_0 + \beta_1 x_i$, MNČ odhadneme regresní koeficienty, vypočteme regresní a reziduální součty čtverců S_R a S_E a testové kritérium $F = \frac{S_R}{\frac{S_E}{n-2}}$. Pokud $F \geq F_{1-\alpha}(1,n-2)$, pak regresor x_i zařadíme do modelu.

3. krok: Vypočteme výběrové parciální korelační koeficienty mezi závisle proměnnou a regresory dosud nezařazenými do modelu a vyloučením vlivu regresoru x_i. Vybereme ten regresor x_j, pro který je absolutní hodnota parciálního korelačního koeficientu největší.

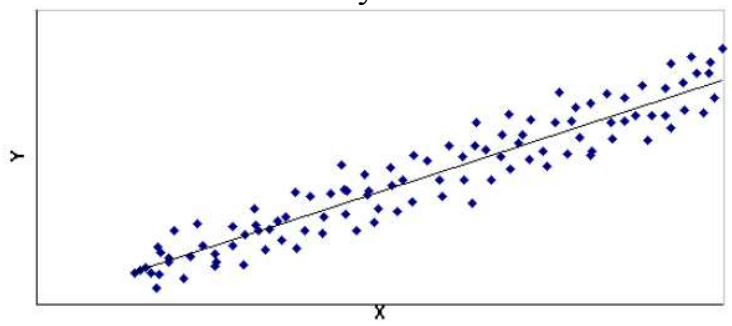
4. krok: Sestavíme model $Y = \beta_0 + \beta_1 x_i + \beta_2 x_j$, MNČ odhadneme regresní koeficienty, vypočteme regresní a reziduální součty čtverců S_R a S_E a testové kritérium $F = \frac{\Delta S_R}{\frac{S_E}{n-3}}$, kde ΔS_R je přírůstek regresního součtu čtverců při zařazení regresoru x_j do modelu. Pokud $F \geq F_{1-\alpha}(1,n-3)$, pak regresor x_j zařadíme do modelu.

5. krok: Vypočteme výběrové parciální korelační koeficienty mezi závisle proměnnou a regresory dosud nezařazenými do modelu s vyloučením vlivu regresorů x_i a x_j a podle kroků 3 a 4 postupujeme dále, až vyčerpáme všechny regresory.

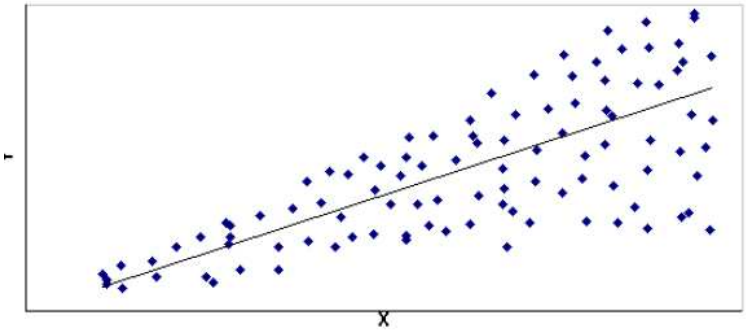
Postup při budování modelu mnohonásobné lineární regrese

1. Sestrojíme dvourozměrné tečkové diagramy dvojic (Y,x_j), $j = 1, ..., p$. Lze-li diagramem uspokojivě proložit přímku, svědčí to o tom, že Y lineárně závisí na x_j. Objeví-li se náhodný mrak bodů, Y na x_j záviset nebude. Obrazce jiných tvarů svědčí o problémech. Například trojúhelníkový tvar dvourozměrného tečkového diagramu indikuje heteroskedasticitu (tzn. že je porušena podmínka (d) v modelu klasické lineární regrese, tedy náhodné odchylky nemají týž rozptyl). Poučení o heteroskedasticitě lze nalézt např. v knize J. Hebák, J. Hustopecký: Vícerozměrné statistické metody s aplikacemi, SNTL 1987, Praha, kde je popsána zobecněná metoda nejmenších čtverců.

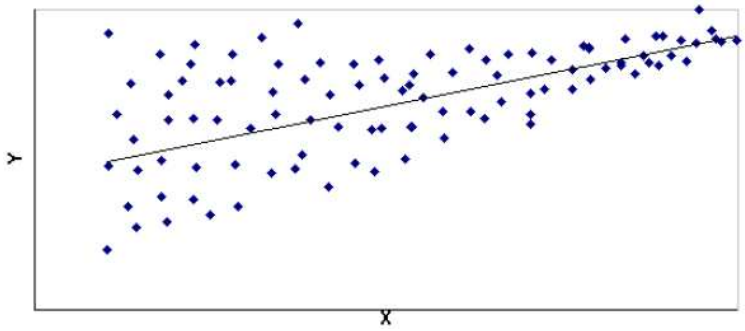
Ukázka homoskedastických dat:



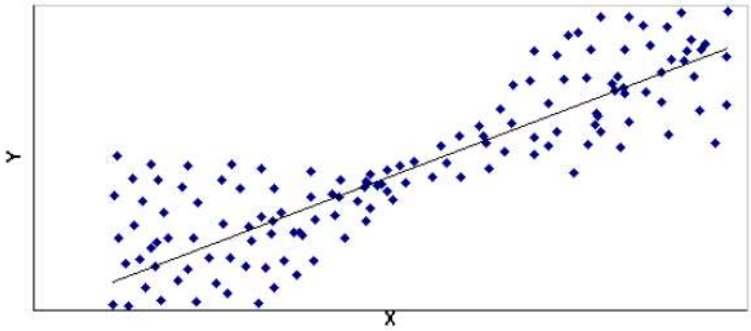
Ukázka dat s rostoucí heteroskedasticitou:



Ukázka dat s klesající heteroskedasticitou:



Ukázka dat s proměnlivou heteroskedasticitou:



2. Vypočteme výběrové párové korelační koeficienty, abychom posoudili sílu případné lineární závislosti Y na x_i . Dále vypočteme všechny výběrové parciální korelační koeficienty, abychom posoudili sílu „čisté" lineární závislosti mezi Y a x_i při vyloučení vlivu ostatních proměnných. Budou-li velké rozdíly mezi párovými a parciálními korelačními koeficienty, svědčí to o existenci multikolinearity.

3. V modelu $Y_i = \beta_0 + \beta_1 x_{il} + ... + \beta_p x_{ip} + \varepsilon_i$, i = l, ..., n získáme bodové a intervalové odhady regresních parametrů β_0, β_l, ..., β_p, index determinace, odhad rozptylu. Provedeme dílčí t-testy a celkový F-test. Vliv jednotlivých proměnných posoudíme pomocí B-koeficientů.

4. Z modelu vyloučíme ty nezávisle proměnné, pro něž byly dílčí t-testy nevýznamné.

Příklad

Šest studentů gymnázia absolvovalo čtyři testy, které měří následující veličiny: X_1 - přírodovědné vědomosti, X_2 – literární vědomosti, X_3 – schopnost koncentrace, X_4 – logické myšlení. Testy se hodnotí na škále od 1 do 10 (1 = špatný výsledek, 10 = výborný výsledek).

student	X_1	X_2	X_3	X_4
1	7	9	10	8
2	9	8	8	10
3	4	3	1	2
4	2	3	2	2
5	3	1	2	4
6	1	1	1	4

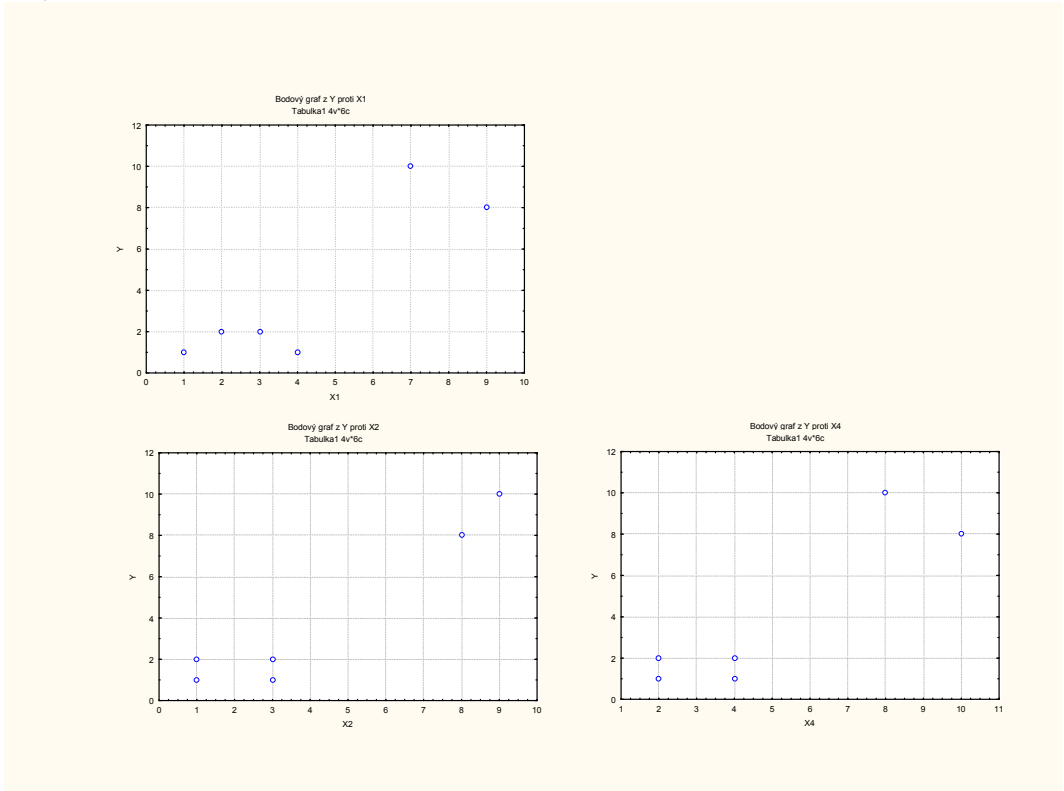
Zajímá nás, kolik bodů můžeme očekávat v testu koncentračních schopností studenta, jestliže známe výsledky testů pro literární schopnosti, přírodovědné schopnosti a logické myšlení.

Řešení pomocí systému STATISTICA:

V tomto problému je proměnná X_3 závislá (označíme ji Y) a ostatní proměnné jsou nezávislé.

Sestavíme regresní model $Y_i = \beta_0 + \beta_1 x_{il} + \beta_2 x_{i2} + \beta_4 x_{i4} + \varepsilon_i$, i = 1, ..., 6.

Nejprve sestrojíme dvourozměrné tečkové diagramy vyjadřující závislost Y na X_1, X_2 a X_4.



Dále spočteme výběrové korelační koeficienty r_{Y,X_1}, r_{Y,X_2}, r_{Y,X_4} a výběrové parciální korelační koeficienty $r_{Y,X_1.X_2}$, $r_{Y,X_1.X_4}$, $r_{Y,X_2.X_1}$, $r_{Y,X_2.X_4}$, $r_{Y,X_4.X_1}$, $r_{Y,X_4.X_2}$.

	Korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)		
Proměnná	X1	X2	X4
Y	0,87	0,96	0,89

Vidíme, že korelace dvojic (Y, X_1), (Y, X_2), (Y, X_4) jsou vysoké.

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	X1	Y
X1	1,0000	0,0273
Y	0,0273	1,0000

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	X1	Y
X1	1,0000	0,4275
Y	0,4275	1,0000

Parciální korelace dvojice (Y, X_1) při vyloučení vlivu veličiny X_2 je pouze 0,0273 a při vyloučení vlivu veličiny X_4 je 0,4275, tedy mnohem slabší než párová korelace, která činila 0,87.

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	X2	Y
X2	1,0000	0,8108
Y	0,8108	1,0000

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	X2	Y
X2	1,0000	0,8773
Y	0,8773	1,0000

Parciální korelace dvojice (Y, X_2) při vyloučení vlivu veličiny X_1 resp. X_4 je stále silná, jen o něco menší než párová korelace (ta byla 0,96).

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	Y	X4
Y	1,0000	0,5586
X4	0,5586	1,0000

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	Y	X4
Y	1,0000	0,6590
X4	0,6590	1,0000

Parciální korelace dvojice (Y, X_4) při vyloučení vlivu veličiny X_1 resp. X_2 je o dost menší než párová korelace (ta byla 0,89), ale pokles není tak výrazný jako u dvojice (Y, X_1) při vyloučení vlivu veličiny X_2 resp. X_4.

Z těchto analýz vyplývá, že největší roli v modelu lineární regresní závislosti Y na X_1, X_2 a X_4 bude hrát proměnná X_2, podstatně menší X_4 a role X_1 bude zřejmě jen nepatrná.

Metodou nejmenších čtverců získáme odhady regresních parametrů.

	Výsledky regrese se závislou proměnnou : Y (ctyritesty.sta) R= .98240301 R2= .96511567 Upravené R2= .91278918 F(3,2)=18.444 p<.05187 Směrod. chyba odhadu : 1.1664					
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(2)	Úroveň p
Abs.člen			#######	0.941927	#######	0.366858
X1	########	0.368366	#######	0.472872	#######	0.502130
X2	0.864242	0.316998	0.97862	0.358949	2.72633	0.112320
X4	0.445257	0.271142	0.53513	0.325873	1.64215	0.242263

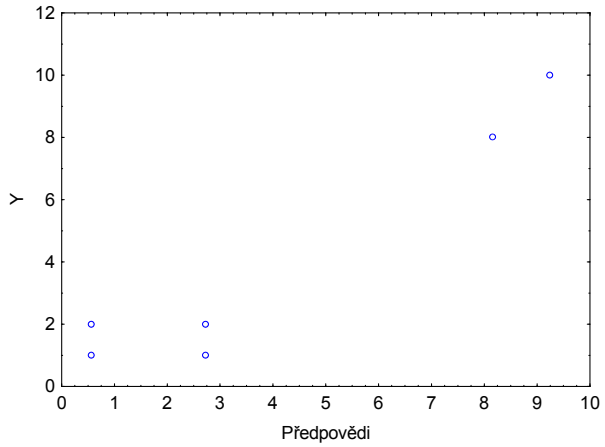
Empirická regresní funkce má tedy tvar $\hat{Y} = -1{,}09 - 0{,}38x_1 + 0{,}98x_2 + 0{,}54x_4$. Variabilita proměnné Y je z 96,5% vysvětlená zvoleným regresním modelem. Pro $\alpha = 0{,}05$ je celkový F-test nevýznamný, všechny dílčí t-testy rovněž. Podíváme-li se na beta koeficienty, vidíme, že největší vliv má proměnná X_2. Sestavíme tedy nový model $Y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i$, $i = 1, \ldots, 6$. Metodou nejmenších čtverců opět získáme odhady regresních parametrů.

	Výsledky regrese se závislou proměnnou : Y (ctyritesty.sta) R= .95813306 R2= .91801897 Upravené R2= .89752371 F(1,4)=44.792 p<.00259 Směrod. chyba odhadu : 1.2644					
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			########	0.850099	########	0.573413
X2	#######	0.143162	1.084932	0.162108	6.692666	0.002593

Nyní má empirická regresní funkce tvar $\hat{Y} = -0{,}52 + 1{,}08x_2$, model jako celek je významný a nezávisle proměnná X_2 rovněž.
Pro kontrolu kvality regrese porovnáme zjištěné a predikované hodnoty veličiny Y.

	1 student1	2 student2	3 student3	4 student4	5 student5	6 student6
Skutečnost	10.0	8.0	1.0	2.0	2.0	1.0
Predikce	9.2	8.2	2.7	2.7	0.6	0.6

Vztah mezi naměřenými a předikovanými hodnotami znázorníme pomocí dvourozměrného tečkového diagramu.



Nyní aplikujeme dopřednou metodu postupné regrese:
Statistiky – Vícerozměrná regrese – Proměnné – Závisle proměnná Y, Nezávisle proměnné X1, X2, X4 – 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.

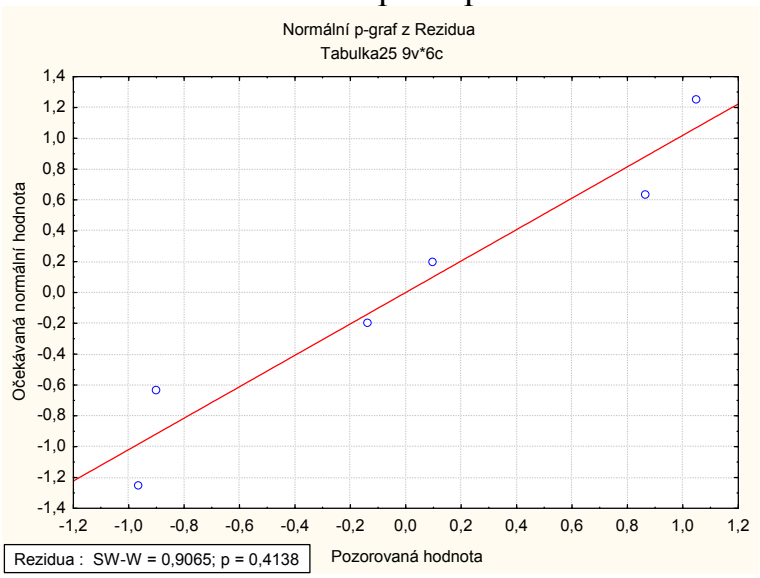
N=6	Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta) R= ,95813306 R2= ,91801897 Upravené R2= ,89752371 F(1,4)=44,792 p<,00259 Směrod. chyba odhadu : 1,2644					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			-0,520548	0,850099	-0,612338	0,573413
X2	0,958133	0,143162	1,084932	0,162108	6,692666	0,002593

V prvním kroku byla vybrána proměnná X_2. Opět klikneme na Další a dostaneme výsledky kroku 2, který je již konečný:

N=6	Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta) R= ,97653416 R2= ,95361897 Upravené R2= ,92269829 F(2,3)=30,841 p<,00999 Směrod. chyba odhadu : 1,0981					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(3)	Úroveň p
Abs.člen			-1,22615	0,872554	-1,40524	0,254603
X2	0,687789	0,217256	0,77881	0,246007	3,16580	0,050644
X4	0,329675	0,217256	0,39622	0,261109	1,51745	0,226436

Empirická regresní funkce má tvar $\hat{Y} = -1{,}23 + 0{,}78x_2 + 0{,}4x_4$, model jako celek je významný na hladině 0,05, avšak nezávisle proměnná X_2 a X_4 nikoliv. Přispívají však k vysvětlení variability hodnot závisle proměnné veličiny Y. Adjusto-

vaný index determinace je 0,9227. V modelu s nezávisle proměnnou X_2 byl 0,8975 a v modelu se všemi třemi nezávisle proměnnými byl 0,9128.
V tomto výsledném modelu uložíme rezidua a predikované hodnoty:
Rezidua/předpoklady/předpovědi – Reziduální analýza – Uložit rezidua & předpovědi – OK
Pomocí S-W testu a N-P plotu prozkoumáme normalitu reziduí:



Vidíme, že rozložení reziduí je blízké normálnímu rozložení.

Zkusíme ještě zpětnou metodu postupné regrese:
Na záložce Metoda zvolíme Metoda – zvolíme Kroková zpětná. V nultém kroku jsou do modelu zařazeny všechny nezávisle proměnné:

N=6	Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta) R= ,98240301 R2= ,96511567 Upravené R2= ,91278918 F(3,2)=18,444 p<,05187 Směrod. chyba odhadu : 1,1664					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(2)	Úroveň p
Abs.člen			-1,08961	0,941927	-1,15679	0,366858
X1	-0,299065	0,368366	-0,38391	0,472872	-0,81187	0,502130
X2	0,864242	0,316998	0,97862	0,358949	2,72633	0,112320
X4	0,445257	0,271142	0,53513	0,325873	1,64215	0,242263

V 1. kroku je z modelu vyřazena proměnná X_1:

N=6	Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta) R= ,97653416 R2= ,95361897 Upravené R2= ,92269829 F(2,3)=30,841 p<,00999 Směrod. chyba odhadu : 1,0981					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(3)	Úroveň p
Abs.člen			-1,22615	0,872554	-1,40524	0,254603
X2	0,687789	0,217256	0,77881	0,246007	3,16580	0,050644
X4	0,329675	0,217256	0,39622	0,261109	1,51745	0,226436

Ve 2. kroku, který je současně poslední, je vyřazena proměnná X_4:

N=6	Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta) R= ,95813306 R2= ,91801897 Upravené R2= ,89752371 F(1,4)=44,792 p<,00259 Směrod. chyba odhadu : 1,2644					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			-0,520548	0,850099	-0,612338	0,573413
X2	0,958133	0,143162	1,084932	0,162108	6,692666	0,002593

Metoda zpětné postupné regrese tedy jako optimální našla model regresní přímky s nezávisle proměnnou X_2.
Upozornění: Pokud bychom na záložce Metoda ručně změnili hodnoty „F na zahrnutí“ a „F na vyjmutí“, mohli bychom dostat jiné výsledky.

Řešení pomocí systému SPSS

Nejprve sestrojíme dvourozměrné tečkové diagramy vyjadřující závislost Y na X_1, X_2 a X_4: Graphs - Legacy Dialogs – Scatter/Dot – Define – Y Axis Y, X Axis X1 (resp. X2 resp. X4)

Výpočet výběrových korelačních koeficientů r_{Y,X_1}, r_{Y,X_2}, r_{Y,X_4} a výběrových parciálních korelačních koeficientů $r_{Y,X_1.X_2}$, $r_{Y,X_1.X_4}$, $r_{Y,X_2.X_1}$, $r_{Y,X_2.X_4}$, $r_{Y,X_4.X_1}$, $r_{Y,X_4.X_2}$:
Analyze – Correlate – Bivariate – Variables Y, X1 (analogický výpočet pro dvojice Y, X2 a Y, X4)

Correlations

		X1	Y
X1	Pearson Correlation	1,000	,872*
	Sig. (2-tailed)		,023
	N	6	6
Y	Pearson Correlation	,872*	1,000
	Sig. (2-tailed)	,023	
	N	6	6

*. Correlation is significant at the 0.05 level (2-tailed).

Analyze – Correlate – Partial – Variables Y, X1, Controlling for X2 – OK

Correlations

Control Variables			Y	X1
X2	Y	Correlation	1,000	,027
		Significance (2-tailed)	.	,965
		df	0	3
	X1	Correlation	,027	1,000
		Significance (2-tailed)	,965	.

Correlations

Control Variables			Y	X1
X2	Y	Correlation	1,000	,027
		Significance (2-tailed)	.	,965
		df	0	3
	X1	Correlation	,027	1,000
		Significance (2-tailed)	,965	.
		df	3	0

Analogicky pro ostatní parciální koeficienty korelace.

Metodou nejmenších čtverců získáme odhady regresních parametrů.
Analyze – Regression – Linear – Dependent Y, Independents X1, X2, X4

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1,090	,942		-1,157	,367
	X1	-,384	,473	-,299	-,812	,502
	X2	,979	,359	,864	2,726	,112
	X4	,535	,326	,445	1,642	,242

a. Dependent Variable: Y

Vidíme, že dílčí t-testy nezamítají hypotézy o nevýznamnosti jednotlivých regresních koeficientů na hladině významnosti 0,05, všechny čtyři p-hodnoty jsou větší než 0,05. Výsledek celkového F-testu najdeme v tabulce ANOVA:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	75,279	3	25,093	18,444	,052[a]
	Residual	2,721	2	1,360		
	Total	78,000	5			

a. Predictors: (Constant), X4, X2, X1

b. Dependent Variable: Y

Testová statistika se realizuje hodnotou 18,444, odpovídající p-hodnota je 0,052, tedy na hladině významnosti 0,05 nezamítáme hypotézu, že dostačující je model konstanty.

Index determinace a adjustovaný index determinace je uveden v tabulce Model Summary:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,982[a]	,965	,913	1,166

a. Predictors: (Constant), X4, X2, X1

V tabulce „Coefficients“ ve sloupci Beta zjistíme, že největší absolutní hodnota koeficientu beta je u proměnné X2. Sestavíme tedy nový model s nezávisle proměnnou X2 a dostaneme tyto výsledky:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,958[a]	,918	,898	1,264

a. Predictors: (Constant), X2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71,605	1	71,605	44,792	,003[a]
	Residual	6,395	4	1,599		
	Total	78,000	5			

a. Predictors: (Constant), X2

b. Dependent Variable: Y

Coefficients[a]

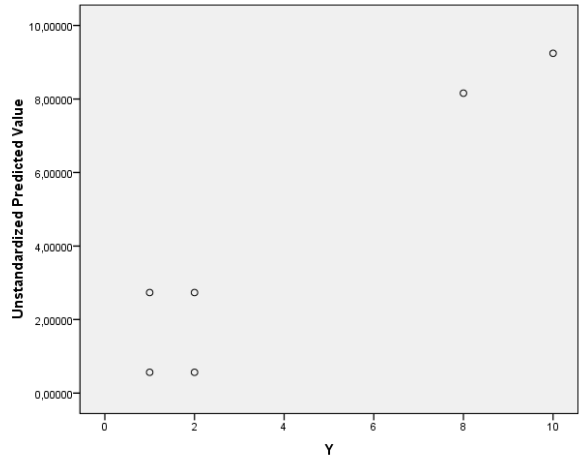
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-,521	,850		-,612	,573
	X2	1,085	,162	,958	6,693	,003

a. Dependent Variable: Y

Vidíme, že model jako celek je významný na hladině významnosti 0,05 a nezávisle proměnná X_2 rovněž.

Porovnání zjištěné a predikované hodnoty veličiny Y:

V menu Linear regression klikneme na Save, ve sloupcích Predicted values a Residulas zaškrtneme Unstandardized – Continue – OK. K datovému souboru se přidají dvě nové proměnné PRE_1 a RES_1.
Graficky znázorníme závislost predikovaných hodnot na původních hodnotách:



Normalitu reziduí posoudíme pomocí Lilieforsovy varianty K-S testu a pomocí S-W testu:
Analyze – Descriptive Statistics – Explore – Dependent list RES_1 – OK.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,150	6	,200*	,984	6	,971

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

V tabulce Test of normality zjistíme, že ani jeden z testů nezamítá hypotézu o normalitě reziduí na hladině významnosti 0,05.

Dopředná metoda postupné regrese v SPSS:
Analyze – Regression – Linear – Dependent Y, Independents X1, X2, X4 - Continue, Method Forward – OK
Tabulka s indexem determinace:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,958[a]	,918	,898	1,264

a. Predictors: (Constant), X2

Tabulka s výsledky ANOVY:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71,605	1	71,605	44,792	,003[a]
	Residual	6,395	4	1,599		
	Total	78,000	5			

a. Predictors: (Constant), X2

b. Dependent Variable: Y

Tabulka s odhady regresních koeficientů:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-,521	,850		-,612	,573
	X2	1,085	,162	,958	6,693	,003

a. Dependent Variable: Y

Tabulka s vyloučenými proměnnými:

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	X1	,019[a]	,047	,965	,027	,177
	X4	,330[a]	1,517	,226	,659	,328

a. Predictors in the Model: (Constant), X2

b. Dependent Variable: Y