Osnova přednášky Mnohonásobná lineární regrese

1. Popis modelu mnohonásobné lineární regrese

Budeme zkoumat lineární závislost veličiny Y na p nezávisle proměnných veličinách (regresorech) $X_1, ..., X_p$.

Omezíme se pouze na model tvaru

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

Interpretace parametrů:

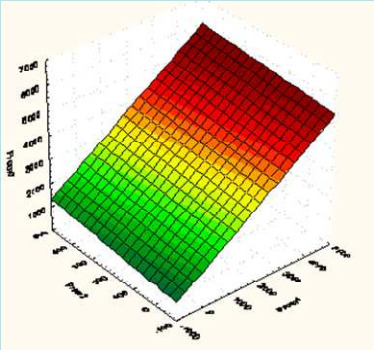
β_0 … teoretická hodnota závisle proměnné veličiny při nulových hodnotách všech nezávisle proměnných veličin,

β_j … přírůstek teoretické hodnoty závisle proměnné veličiny odpovídající jednotkové změně j-té nezávisle proměnné veličiny při konstantní úrovni ostatních nezávisle proměnných, j = l, ..., p.

Parametry $\beta_1, …, \beta_p$ se nazývají parciální regresní koeficienty.

Geometricky tento model představuje regresní nadrovinu.

Ilustrace pro dva regresory:



Model $Y_i = \beta_0 + \beta_1 X_{il} + ... + \beta_p X_{ip} + \varepsilon_i$, i = 1, ..., n lze formálně ztotožnit s lineárním regresním modelem z přednášky „Jednoduchá lineární regrese I“:

$Y_i = \beta_0 + \beta_1 f_1(x_i) + ... + \beta_p f_p(x_i) + \varepsilon_i$, i = 1, ..., n,

kde položíme $f_1(x_i) = x_{i1}$, ..., $f_p(x_i) = x_{ip}$, i = 1,..., n.

Dostáváme tedy maticový tvar $\mathbf{Y = X\beta + \varepsilon}$, 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 přednášce „Jednoduchá lineární regrese“ zůstávají v platnosti.

Příklad:

Při zkoumání závislosti hodinové výkonnosti dělníka (veličina Y – v kusech) na jeho věku (veličina X_1 – v letech) a době zapracovanosti (veličina X_2 – v letech) byly u 10 náhodně vybraných dělníků zjištěny tyto údaje:

Y	67	65	75	66	77	84	69	60	70	66
X_1	43	40	49	46	41	41	48	34	32	42
X_2	6	8	14	14	8	12	16	1	5	7

Najděte regresní matici a vektor regresních parametrů.

Řešení:

$$\mathbf{X} = \begin{pmatrix} 1 & 43 & 6 \\ 1 & 40 & 8 \\ 1 & 49 & 14 \\ 1 & 46 & 14 \\ 1 & 41 & 8 \\ 1 & 41 & 12 \\ 1 & 48 & 16 \\ 1 & 34 & 1 \\ 1 & 32 & 5 \\ 1 & 42 & 7 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

2. Specifika modelu mnohonásobné lineární regrese

2.1. Kroky před prováděním mnohonásobné lineární regrese

a) Musíme prozkoumat, zda naše data splňují předpoklady pro regresní analýzu.

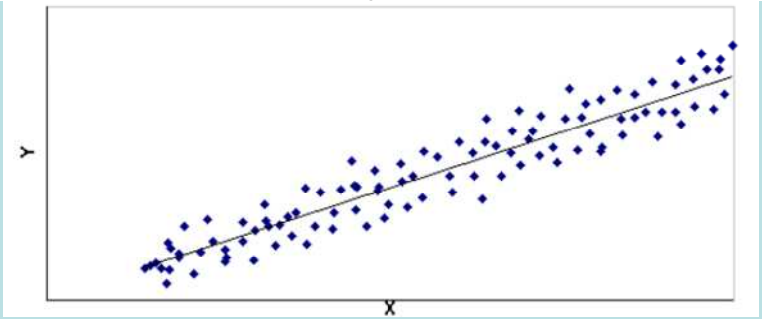
b) Pokud je nesplňují, posoudíme, jak vážné je porušení těchto předpokladů.

c) Je-li porušení předpokladů vážné, musíme s daty provést některé operace, abychom porušení předpokladů odstranili (nebo aspoň zmírnili).

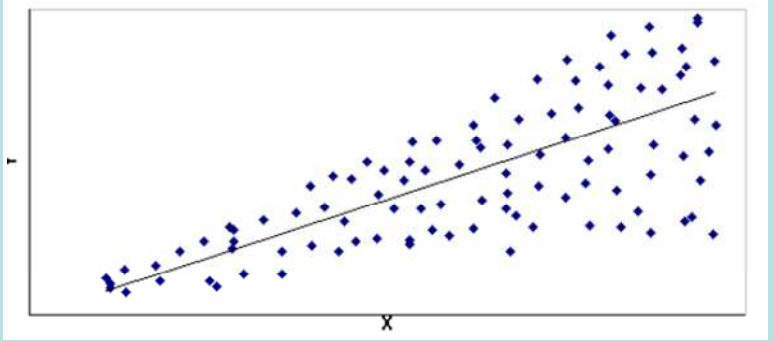
2.2. Sedm hlavních předpokladů regresní analýzy

1. Závisle proměnná Y musí být proměnná aspoň intervalového typu. (Pokud není, musíme použít logistickou regresi.)
2. Nezávisle proměnné $X_{1,}$ ..., X_p jsou rovněž aspoň intervalového typu. Mohou to být i proměnné alternativní.
3. Nezávisle proměnné by neměly být mezi sebou příliš vysoce korelovány. Pokud v datech existuje multikolinearita, výsledky regrese jsou nespolehlivé. Vysoká multikolinearita zvyšuje pravděpodobnost, že důležitá nezávisle proměnná bude shledána statisticky nevýznamná a bude vyřazena z modelu.
4. V datech nesmějí být odlehlé či extrémní hodnoty, neboť na ty je regresní analýza citlivá. Odlehlé hodnoty mohou vážně narušit kvalitu odhadů regresních parametrů.
5. Proměnné musejí být v lineárním vztahu. Vícenásobná lineární regrese je založena Pearsonově korelačním koeficientu, takže neexistence linearity způsobuje, že i důležité vztahy mezi proměnnými, pokud nejsou lineární, zůstanou neodhaleny.
6. Proměnné mají normální rozložení. Význam tohoto předpokladu ustupuje do pozadí, máme-li dostatečně velký datový soubor, kde se již uplatňuje působení centrální limitní věty.
7. Proměnné vykazují homoskedasticitu, tedy homogenitu rozptylu. (Opakem homoskedasticity je heteroskedasticita.)

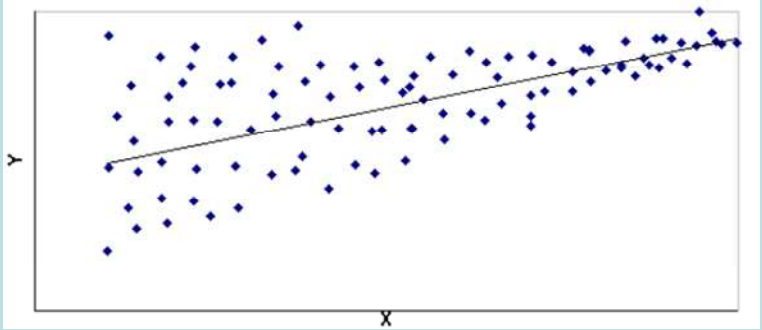
Ukázka homoskedastických dat:



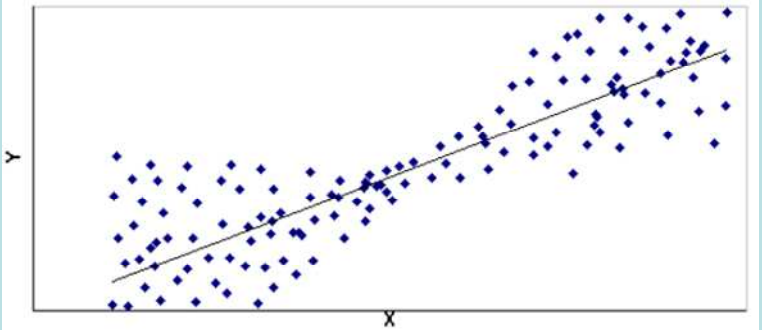
Ukázka dat s rostoucí heteroskedasticitou:



Ukázka dat s klesající heteroskedasticitou:



Ukázka dat s proměnlivou heteroskedasticitou:

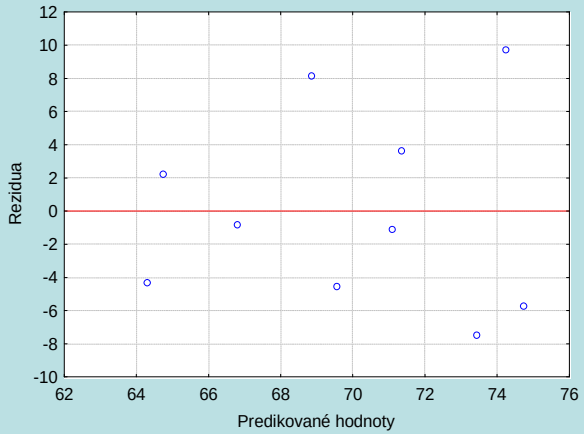


2.3. Ověřování předpokladů modelu

Ověřování normality:

- jednorozměrná: použijeme např. N-P plot a S-W test či Lilieforsův test.

- vícerozměrná: sestrojíme graf závislosti reziduí na predikovaných hodnotách. Tečky by měly být rovnoměrně rozptýleny po obou stranách vodorovné osy.



Odhalení multikolinearity:

- Vysoké absolutní hodnoty výběrových korelačních koeficientů nezávisle proměnných (orientačně > 0,75).
- Velké rozdíly mezi párovými a parciálními korelačními koeficienty.
- Celkový F-test je významný, ale dílčí t-testy nikoliv.

Při použití statistického software lze informace o multikolinearitě získat pomocí koeficientu VIF (Variance inflation factor). Má-li koeficient VIF hodnotu 1, pak příslušná nezávisle proměnná není korelovaná s ostatními nezávisle proměnnými, jestliže $1 < VIF < 5$, pak existuje mírná korelace, pro $VIF > 5$ vysoká korelace a pro $VIF > 10$ extrémní multikolinearita.

Odstranění multikolinearity:

- Je-li multikolinearita způsobena silnou lineární závislostí dvou proměnných, vypustíme jednu z nich z analýzy. Tím se nedopustíme žádné závažné chyby, neboť když máme dvě vysoce vzájemně korelované proměnné, velmi často to znamená, že obě indikují podobný jev. Tím, že jednu z těchto proměnných z regresního modelu vyřadíme, nijak jej neoslabíme.
- Je-li multikolinearita zapříčiněna vzájemnou korelovaností několika proměnných, nabízí se řešení zkombinovat je do jedné nové proměnné. Tu vytvoříme např. s pomocí analýzy hlavních komponent.

Příklad: Při zkoumání závislosti hodinové výkonnosti dělníka (veličina Y – v kusech) na jeho věku (veličina X_1 – v letech) a době zapracovanosti (veličina X_2 – v letech) byly u 10 náhodně vybraných dělníků zjištěny tyto údaje:

Y	67	65	75	66	77	84	69	60	70	66
X_1	43	40	49	46	41	41	48	34	32	42
X_2	6	8	14	14	8	12	16	1	5	7

Posuďte pomocí koeficientu VIF, zda proměnné věk a doba zapracovanosti mohou způsobit multikolinearitu v modelu $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Řešení:
Statistiky - Pokročilé lineární/nelineární modely – Obecné regresní modely – OK – Proměnné – Závislá Y, Spojité nezávisle proměnné X_1, X_2 – OK – Matice – Parciální korelace.

Efekt	Toler.	Rozptyl Infl fak	R^2	Y Beta v	Y Parciál.	Y Semipar.	Y t	Y p
"X1"	0,282545	3,539258	0,717455	-0,550937	-0,328630	-0,292850	-0,920604	0,387883
"X2"	0,282545	3,539258	0,717455	0,920415	0,502564	0,489246	1,537994	0,167937

Koeficient VIF je 3,54, tedy mezi věkem a dobou zapracovanosti existuje jen mírná korelace.

Odhalení nelinearity vztahů:

Pomocí tečkového diagramu prozkoumáme závislost reziduí na hodnotách závisle proměnné veličiny Y. Pokud tečky vytvoří nelineární obrazec, pak buď jedna z nezávisle proměnných nebo kombinace nezávisle proměnných mají nelineární vztah se závisle proměnnou veličinou Y. Tento graf nám také pomůže odhalit případnou heteroskedasticitu v datech.

Odstranění nelinearity vztahů:

Doporučuje se ty proměnné, u nichž jsme detekovali nelinearitu, transformovat pomocí logaritmické nebo odmocninové transformace. Pokud tento postup nepomůže, musíme použít nelineární regresi.

Odhalení odlehlých hodnot:

Použijeme krabicové grafy nebo pravidlo 3 sigma. Odlehlé hodnoty mají velký vliv na kvalitu odhadu regresních parametrů.

Způsoby řešení problému odlehlých hodnot:

Ověříme, zda při zadávání hodnot dané proměnné nedošlo k překlepu;

proměnnou transformujeme;

upravíme hodnotu odlehlého případu;

odstraníme případy s odlehlou hodnotou;

proměnnou vymažeme.

2.4. Posouzení vlivu jednotlivých nezávisle proměnných v modelu

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

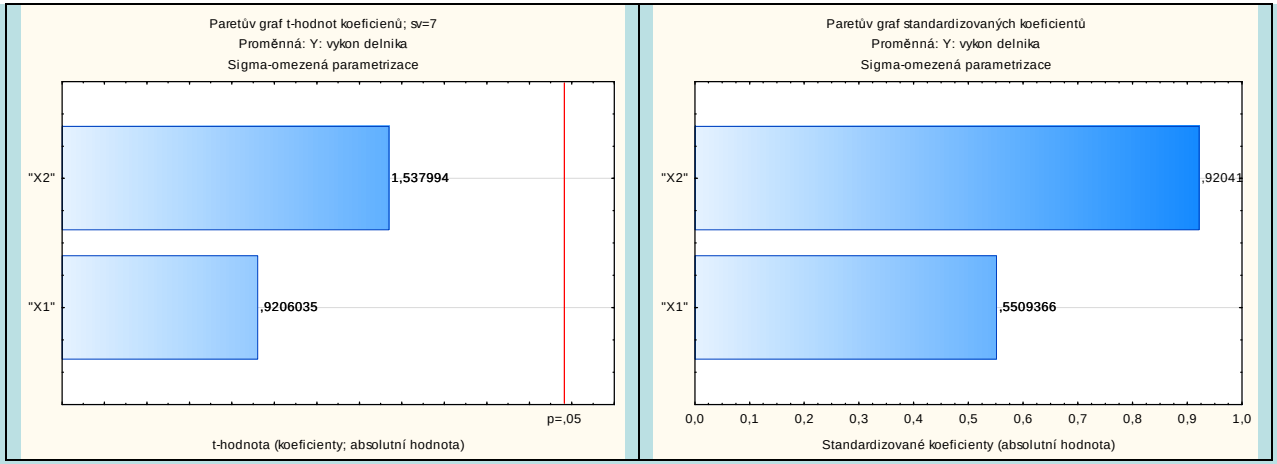
$$Z_i = \frac{Y_i - m_Y}{s_Y}, v_{ij} = \frac{x_{ij} - m_{x_j}}{s_{x_j}}, j = 1, ..., p, i = 1, ..., n$$

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.

V sytému STATISTICA jsou B-koeficienty značeny b*.

Graficky lze absolutní hodnoty standardizovaných regresních parametrů (nebo absolutní hodnoty testových statistik dílčích t-testů) znázornit pomocí Paretových grafů.

Statistiky - Pokročilé lineární/nelineární modely – Obecné regresní modely – OK – Proměnné – Závislá Y, Spojité nezávisle proměnné X_1, X_2 – OK – Paretův graf (pokud ponecháme zaškrtnuto t-hodn., dostaneme graf pro absolutní hodnoty testových statistik, pokud tuto volbu vypneme, získáme graf pro absolutní hodnoty standardizovaných regresních parametrů).



Příklad: Při zkoumání závislosti hodinové výkonnosti dělníka (veličina Y – v kusech) na jeho věku (veličina X_1 – v letech) a době zapracovanosti (veličina X_2 – v letech) byly u 10 náhodně vybraných dělníků zjištěny tyto údaje:

Y	67	65	75	66	77	84	69	60	70	66
X_1	43	40	49	46	41	41	48	34	32	42
X_2	6	8	14	14	8	12	16	1	5	7

Posuďte vliv věku a doby zapracovanosti na výkon dělníka pomocí standardizovaných regresních parametrů.

Řešení:
Statistiky – Vícenásobná regrese – Proměnné – Závislá proměnná Y, seznam nezáv. proměnných X1, X2 – OK – OK.

	Výsledky regrese se závislou proměnnou : Y (vykony delniku.sta) R= ,54005243 R2= ,29165662 Upravené R2= ,08927280 F(2,7)=1,4411 p<,29913 Směrod. chyba odhadu : 6,6491					
N=10	b*	Sm.chyba z b*	b	Sm.chyba z b	t(7)	p-hodn.
Abs.člen			86,74217	25,32397	3,425299	0,011056
X1	-0,550937	0,598452	-0,70031	0,76071	-0,920604	0,387883
X2	0,920415	0,598452	1,35062	0,87817	1,537994	0,167937

Standardizované regresní parametry jsou uvedeny ve sloupci b^*. Pro věk má tento parametr hodnotu -0,5509 a pro dobu zapracovanosti 0,9204. V absolutní hodnotě je vyšší parametr pro dobu zapracovanosti, tedy tato proměnná má vyšší vliv na výkon než věk.

3. Dvě hlavní metody při provádění mnohonásobné lineární regrese

3.1. Metoda ENTER

Tato metoda je standardní metoda, do modelu vstupují všechny nezávisle proměnné najednou.

Metodu ENTER použijeme v případě,

- kdy chceme popsat, jak velký podíl rozptylu závisle proměnné veličiny Y je vysvětlen nezávisle proměnnými veličinami X_1, …, X_p (zajímá nás index determinace),
- kdy chceme zjistit, jak velký vliv má každá z nezávisle proměnných na proměnnou závislou při kontrole vlivu působení ostatních proměnných (interpretujeme nestandardizované odhady regresních parametrů),
- kdy nás zajímá, jaká je relativní důležitost každé z nezávisle proměnných (posuzujeme standardizované odhady regresních parametrů).

Při regresi založené na metodě ENTER by mělo na každou proměnnou připadat minimálně dvacet případů (poměr tedy **1:20**). Budou-li v našem modelu např. čtyři proměnné, datový soubor by měl mít minimálně 80 případů

Nejnižší možný poměr proměnná/počet případů je **1:5**. V tom případě ale platí silný požadavek na normalitu – rozložení reziduí by mělo být normální.

3.2. Metoda STEPWISE

Metoda STEPWISE (postupná regrese) je metoda nalezení „nejlepšího“ modelu (co nejmenší počet nezávisle proměnných veličin, co nejkvalitněší predikce).

Uživatel nekontroluje pořadí proměnných, jak postupně vstupují do modelu, to provádí samotný program, který pracuje podle jistého algoritmu.

Používá se 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.

Pořadí vkládání nezávisle proměnných je důležité, neboť může vést k různým odhadům jejich důležitosti v modelu. Proto je při mnohonásobné regresi vždy nutné si dobře rozmyslet, jakou metodu vkládání proměnných zvolíme.

Při regresi založené na metodě STEPWISE by mělo na každou proměnnou připadat minimálně čtyřicet případů (poměr tedy **1:40**). Budou-li v našem modelu např. čtyři proměnné, datový soubor by měl mít minimálně 160 případů.

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 pro dopřednou metodu, 11 pro zpětnou), je prediktor zařazen.

Je-li statistika F menší než hodnota zvaná „F to remove“ (česky „F na vyjmutí“, ve STATISTICE implicitně 0 pro dopřednou metodu, 10 pro zpětnou), 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, pokud se zvyšuje podíl vysvětlené variability hodnot veličiny Y.

3.3. Postup při budování modelu mnohonásobné lineární regrese

Metoda ENTER

1. Ověříme předpoklady modelu: normalitu, homoskedasticitu, prozkoumáme existenci případné multikolinearity, prověříme linearitu vztahů, detekujeme případná vybočující pozorování.

2. 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.

3. Z modelu vyloučíme ty nezávisle proměnné, pro něž byly dílčí t-testy nevýznamné a odhadneme parametry výsledného modelu.

4. Provedeme reziduální analýzu.

Metoda STEPWISE

1. Ověření předpokladů modelu.

2. Zvolíme dopřednou nebo zpětnou metodu Stepwise, nastavíme hladinu významnosti, hodnoty F na zahrnutí a F na vyjmutí (nebo ponecháme implicitně nastavené hodnoty 0,05, 1, 0.

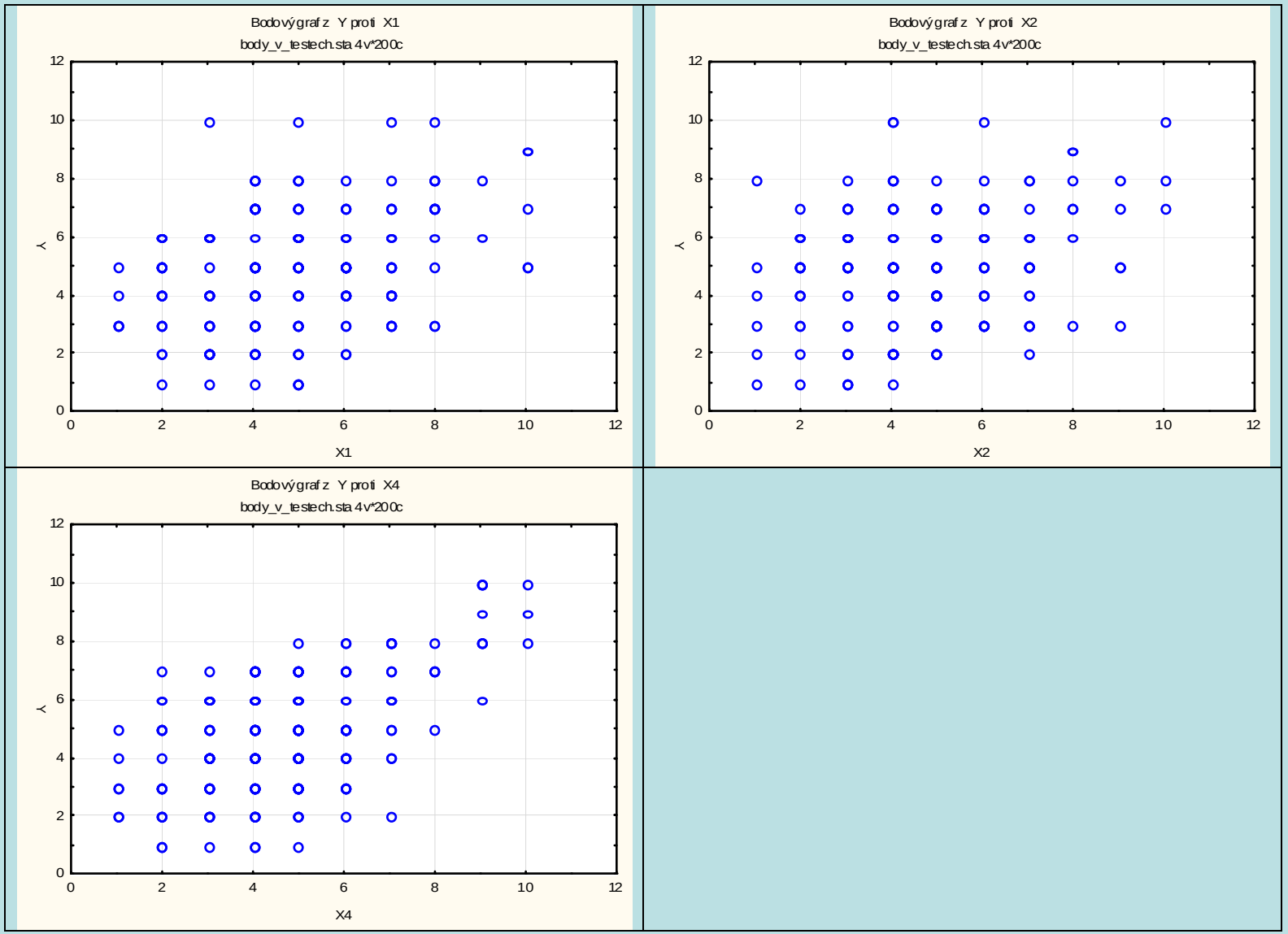
3. Pro výsledný model provedeme reziduální analýzu.

4. Příklad:

200 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).

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, ..., 200.
Nejprve sestrojíme dvourozměrné tečkové diagramy vyjadřující závislost Y na X_1, X_2 a X_4.


Bodový graf z Y proti X1
body_v_testech.sta 4v*200c
Y
X1
Bodový graf z Y proti X2
body_v_testech.sta 4v*200c
Y
X2
Bodový graf z Y proti X4
body_v_testech.sta 4v*200c
Y
X4

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

Párové korelace:

	Korelace (body_v_testech.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=200 (Celé případy vynechány u ChD)		
Proměnná	X1	X2	X4
Y	0,334160	0,252031	0,545681

Vidíme, že korelace dvojic (Y, X_1), (Y, X_2), (Y, X_4) nejsou příliš vysoké, ale jsou významné na hladině významnosti 0,05.

Parciální korelace pro dvojici (Y, X_1):

	Parciální korelace (body_v_testech.sta) S vyloučením vlivu:X2 Označ. korelace jsou významné na hlad. p < ,05000 N=200 (Celé případy vynechány u ChD)	
Proměnná	Y	X1
Y	1,000000	0,248934
X1	0,248934	1,000000

	Parciální korelace (body_v_testech.sta) S vyloučením vlivu:X4 Označ. korelace jsou významné na hlad. p < ,05000 N=200 (Celé případy vynechány u ChD)	
Proměnná	Y	X1
Y	1,000000	0,195610
X1	0,195610	1,000000

Parciální korelace dvojice (Y, X_1) při vyloučení vlivu veličiny X_2 je 0,2489 a při vyloučení vlivu veličiny X_4 je 0,1956, tedy poněkud slabší než párová korelace, která činila 0,3342.

Parciální korelace pro dvojici (Y, X_2):

	Parciální korelace (body_v_testech.sta) S vyloučením vlivu:X1 Označ. korelace jsou významné na hlad. p < ,05000 N=200 (Celé případy vynechány u ChD)	
Proměnná	Y	X2
Y	1,000000	0,105507
X2	0,105507	1,000000

	Parciální korelace (body_v_testech.sta) S vyloučením vlivu:X4 Označ. korelace jsou významné na hlad. p < ,05000 N=200 (Celé případy vynechány u ChD)	
Proměnná	Y	X2
Y	1,000000	0,133832
X2	0,133832	1,000000

Parciální korelace dvojice (Y, X_2) při vyloučení vlivu veličiny X_1 je 0,1055 a při vyloučení vlivu veličiny X_4 je 0,1338, tedy o dost slabší než párová korelace, která činila 0,252.

Parciální korelace pro dvojici (Y, X_4):

	Parciální korelace (body_v_testech.sta) S vyloučením vlivu:X1 Označ. korelace jsou významné na hlad. p < ,05000 N=200 (Celé případy vynechány u ChD)	
Proměnná	Y	X4
Y	1,000000	0,489638
X4	0,489638	1,000000

	Parciální korelace (body_v_testech.sta) S vyloučením vlivu:X2 Označ. korelace jsou významné na hlad. p < ,05000 N=200 (Celé případy vynechány u ChD)	
Proměnná	Y	X4
Y	1,000000	0,513389
X4	0,513389	1,000000

Parciální korelace dvojice (Y, X_4) při vyloučení vlivu veličiny X_1 je 0,4896 a při vyloučení vlivu veličiny X_2 je 0,5134, tedy jen nepatrně slabší než párová korelace, která činila 0,5457.

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_4, podstatně menší X_1 a role X_2 bude zřejmě jen nepatrná.

Posoudíme, zda v modelu existuje multikolinearita:

Efekt	Statistiky kolineace za daných podmínek (body_v_testech.sta) Sigma-omezená parametrizace							
	Toler.	Rozptyl Infl fak	R^2	Y Beta v	Y Parciál.	Y Semipar.	Y t	Y p
X1	0,7120782	1,4043400	0,2879218	0,1506085	0,1530374	0,1270905	2,1680623	0,0313589
X2	0,7428373	1,3461897	0,2571627	0,0500399	0,0524801	0,0431284	0,7357350	0,4627715
X4	0,8785432	1,1382479	0,1214568	0,4829905	0,4830165	0,4527100	7,7228682	0,0000000

Koeficienty VIF jsou jen o málo větší než 1, stupeň multikolinearity je tedy mírný.

Metodou nejmenších čtverců získáme odhady regresních parametrů.

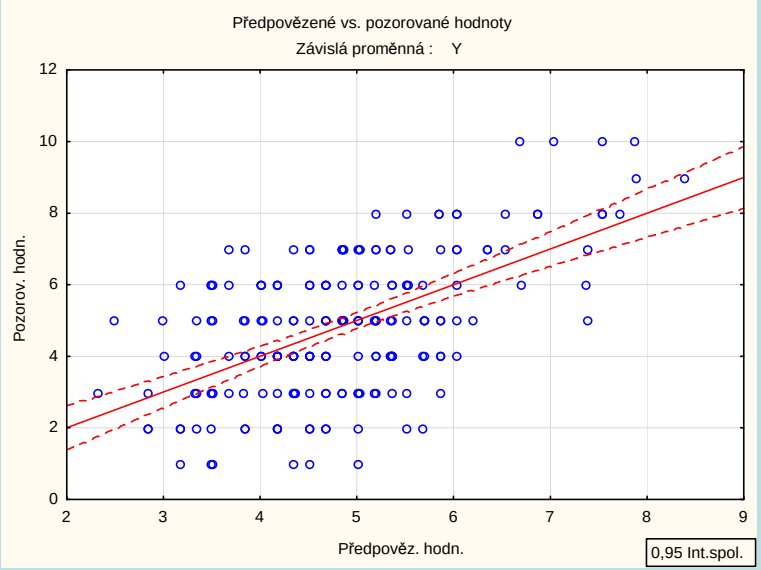
N=200	Výsledky regrese se závislou proměnnou : Y (body_v_testech.sta) R= ,57139981 R2= ,32649775 Upravené R2= ,31618904 F(3,196)=31,672 p<,00000 Směrod. chyba odhadu : 1,6247					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(196)	p-hodn.
Abs.člen			1,533188	0,394207	3,889301	0,000138
X1	0,150608	0,069467	0,148320	0,068411	2,168062	0,031359
X2	0,050040	0,068013	0,052542	0,071414	0,735735	0,462772
X4	0,482990	0,062540	0,497254	0,064387	7,722868	0,000000

Empirická regresní funkce má tedy tvar $\widehat{Y} = 1{,}5332 + 0{,}1483x_1 + 0{,}0525x_2 + 0{,}4973x_4$. Variabilita proměnné Y je z 32,6 % vysvětlená zvoleným regresním modelem. Pro $\alpha = 0{,}05$ je celkový F-test významný, dílčí t-testy až na β_2 rovněž. Sestavíme tedy nový model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_4 x_{i4} + \varepsilon_i$, $i = 1, \ldots, 200$. Metodou nejmenších čtverců opět získáme odhady regresních parametrů.

N=200	Výsledky regrese se závislou proměnnou : Y (body_v_testech.sta) R= ,56976986 R2= ,32463769 Upravené R2= ,31778122 F(2,197)=47,348 p<,00000 Směrod. chyba odhadu : 1,6228					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(197)	p-hodn.
Abs.člen			1,642339	0,364800	4,502032	0,000012
X1	0,173561	0,061995	0,170925	0,061053	2,799595	0,005626
X4	0,488638	0,061995	0,503068	0,063826	7,881865	0,000000

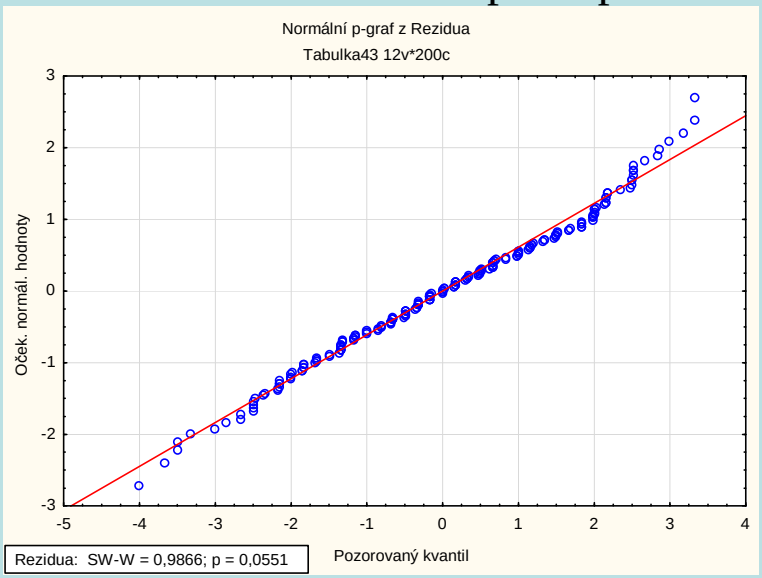
Nyní má empirická regresní funkce tvar $\widehat{Y} = 1{,}6423 + 0{,}1709x_1 + 0{,}5031x_4$, model jako celek je významný a nezávisle proměnné X_1, X_4 rovněž.

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



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:



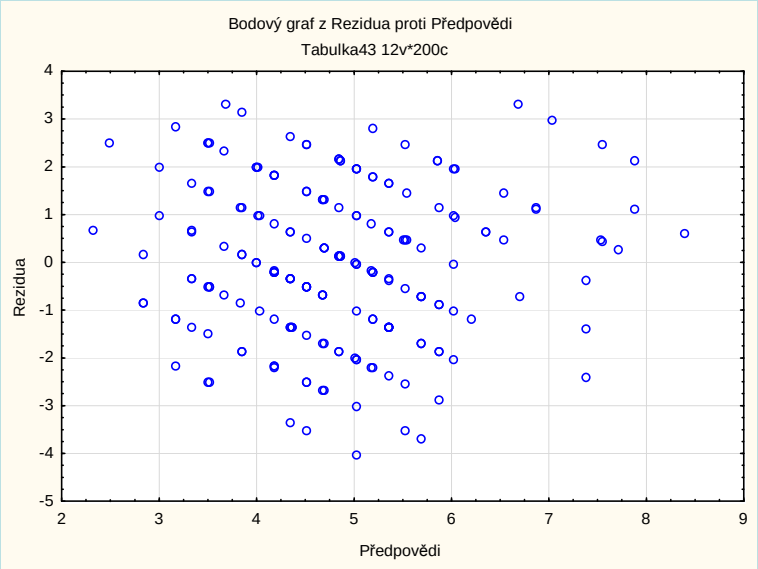
Vidíme, že rozložení reziduí je blízké normálnímu rozložení, p-hodnota S-W testu je větší než 0,05.

Případnou autokorelaci reziduí posoudíme pomocí Durbinovy – Watsonovy statistiky:
Rezidua/předpoklady/předpovědi – Reziduální analýza – Detaily – Durbin-Watsonova statistika:

	Durbin-Watsonovo d (body_v_testech.sta) a sériové korelace reziduí	
	Durbin-Watson.d	Sériové korelace
Odhad	1,833944	0,075507

D-W statistika je blízká 2, rezidua tudíž nebudou vykazovat autokorelaci.

Homoskedasticitu reziduí prozkoumáme pomocí grafu závislosti reziduí na predikovaných hodnotách:



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.

	Výsledky regrese se závislou proměnnou : Y (body_v_testech.sta) R= ,54568132 R2= ,29776811 Upravené R2= ,29422148 F(1,198)=83,958 p<,00000 Směrod. chyba odhadu : 1,6506					
N=200	b*	Sm.chyba z b*	b	Sm.chyba z b	t(198)	p-hodn.
Abs.člen			2,196128	0,311760	7,044297	0,000000
X4	0,545681	0,059554	0,561797	0,061312	9,162868	0,000000

V prvním kroku byla vybrána proměnná X_4. Opět klikneme na Další a dostaneme výsledky kroku 2, který je již konečný:

	Výsledky regrese se závislou proměnnou : Y (body_v_testech.sta) R= ,56976986 R2= ,32463769 Upravené R2= ,31778122 F(2,197)=47,348 p<,00000 Směrod. chyba odhadu : 1,6228					
N=200	b*	Sm.chyba z b*	b	Sm.chyba z b	t(197)	p-hodn.
Abs.člen			1,642339	0,364800	4,502032	0,000012
X4	0,488638	0,061995	0,503068	0,063826	7,881865	0,000000
X1	0,173561	0,061995	0,170925	0,061053	2,799595	0,005626

Empirická regresní funkce má tvar $\widehat{Y} = 1{,}6423 + 0{,}1709x_1 + 0{,}5031x_4$. Dostali jsme stejný model jako v případě, kdy jsme proměnné vybírali na základě výsledků dílčích t-testů.

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=200	Výsledky regrese se závislou proměnnou : Y (body_v_testech.sta) R= ,57139981 R2= ,32649775 Upravené R2= ,31618904 F(3,196)=31,672 p<,00000 Směrod. chyba odhadu : 1,6247					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(196)	p-hodn.
Abs.člen			1,533188	0,394207	3,889301	0,000138
X1	0,150608	0,069467	0,148320	0,068411	2,168062	0,031359
X2	0,050040	0,068013	0,052542	0,071414	0,735735	0,462772
X4	0,482990	0,062540	0,497254	0,064387	7,722868	0,000000

V 1. kroku je z modelu vyřazena proměnná X_2:

N=200	Výsledky regrese se závislou proměnnou : Y (body_v_testech.sta) R= ,56976986 R2= ,32463769 Upravené R2= ,31778122 F(2,197)=47,348 p<,00000 Směrod. chyba odhadu : 1,6228					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(197)	p-hodn.
Abs.člen			1,642339	0,364800	4,502032	0,000012
X1	0,173561	0,061995	0,170925	0,061053	2,799595	0,005626
X4	0,488638	0,061995	0,503068	0,063826	7,881865	0,000000

Ve 2. kroku, který je současně poslední, je vyřazena proměnná X_1:

N=200	Výsledky regrese se závislou proměnnou : Y (body_v_testech.sta) R= ,54568132 R2= ,29776811 Upravené R2= ,29422148 F(1,198)=83,958 p<,00000 Směrod. chyba odhadu : 1,6506					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(198)	p-hodn.
Abs.člen			2,196128	0,311760	7,044297	0,000000
X4	0,545681	0,059554	0,561797	0,061312	9,162868	0,000000

Metoda zpětné postupné regrese tedy jako optimální našla model regresní přímky s nezávisle proměnnou X_1.
Upozornění: Pokud bychom na záložce Metoda ručně změnili hodnoty „F na zahrnutí“ a „F na vyjmutí“, mohli bychom dostat jiné výsledky.