Mnohonásobná lineární regrese

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:

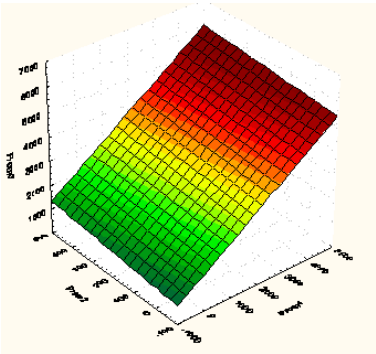
β_0 … teoretická hodnota závisle proměnné veličiny při nulových hodnotách všech nezávisle proměnných veličin,

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 β_1, …, β_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 přednáškách „Jednoduchá lineární regrese I“ a „Jednoduchá lineární regrese II“ 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:

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

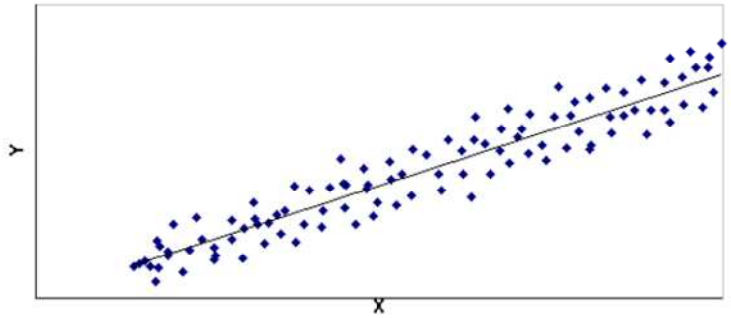
Kroky před prováděním mnohonásobné lineární regrese

1. Musíme prozkoumat, zda naše data splňují předpoklady pro regresní analýzu.
2. Pokud je nesplňují, posoudíme, jak vážné je porušení těchto předpokladů.
3. Je-li porušení předpokladů vážné, musíme s daty provést některé operace, abychom porušení předpokladů odstranili (nebo aspoň zmírnili).

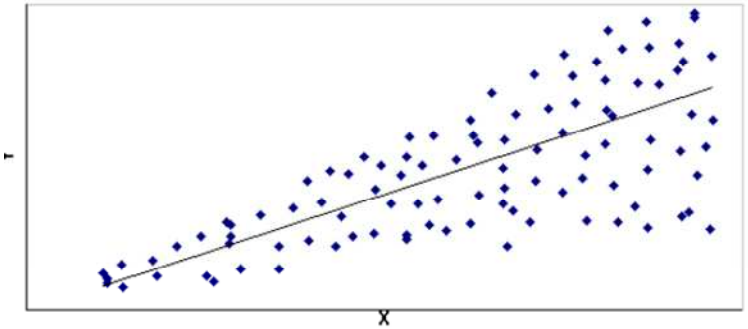
Sedm hlavních předpokladů regresní analýzy

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.)

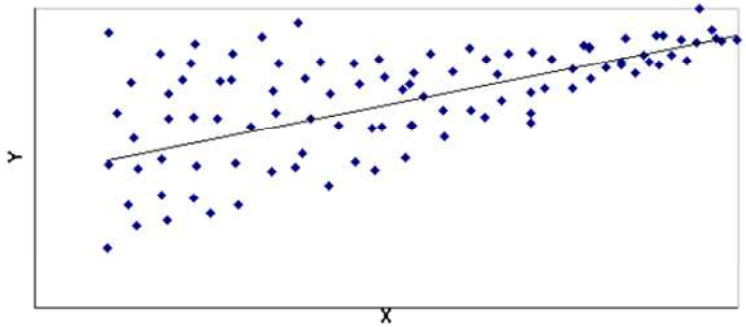
Ukázka homoskedastických dat:



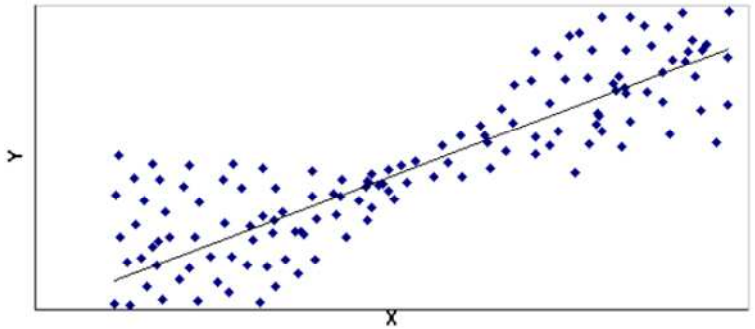
Ukázka dat s rostoucí heteroskedasticitou:



Ukázka dat s klesající heteroskedasticitou:



Ukázka dat s proměnlivou heteroskedasticitou:



Ověřování předpokladů modelu

Při ověřování předpokladů modelu mnohonásobné lineární regrese hraje důležitou úlohu **koeficient parciální korelace**.

Nechť Y, Z jsou náhodné veličiny a $\mathbf{X} = (X_1, \ldots, X_p)'$ je náhodný vektor. Korelační koeficient $\rho(Y,Z)$ udává míru těsnosti lineárního vztahu mezi veličinami Y a Z. Ta však může být ovlivněna i tím, že mezi veličinami $X_1, \ldots, X_p$ existují veličiny, které silně korelují jak s Y, tak se Z. Zajímá nás proto, jaká je „čistá“ korelace mezi Y a Z, když se eliminuje vliv náhodného vektoru **X**.
Pokud se omezíme na lineární vztahy, můžeme vliv vektoru **X** na veličinu Y popsat lineární regresní funkcí $\hat{Y} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$.

Tu část veličiny Y, kterou vektor **X** nevysvětlí, si můžeme představit jako reziduum $Y - \hat{Y}$.

Analogicky pro veličinu Z dostáváme $\hat{Z} = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_j$,

tudíž reziduum $Z - \hat{Z}$ chápeme jako tu část veličiny Z, kterou vektor **X** nevysvětlí.

Korelační koeficient mezi rezidui $Y - \hat{Y}$ a $Z - \hat{Z}$ se nazývá parciální korelační koeficient mezi náhodnými veličinami Y a Z při pevně daném vektoru **X** a značí se $\rho_{Y,Z.\mathbf{X}}$.

Tedy $\rho_{Y,Z.\mathbf{X}} = R(Y - \hat{Y}, Z - \hat{Z})$.

Nechť náhodný vektor $(Y, Z, X_1, \ldots, X_p)'$ pochází z (p+2)-rozměrného rozložení, které má parciální korelační koeficient $\rho_{Y,Z.\mathbf{X}}$.

Nechť je dán náhodný výběr $(Y_1, Z_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, Z_n, X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení. Musí platit $n > p+2$. Jako odhad $\rho_{Y,Z.\mathbf{X}}$ slouží výběrový parciální korelační koeficient

$$r_{Y,Z.\mathbf{X}} = \frac{S_{12}}{S_1 S_2}, \text{ kde}$$

S_{12} je výběrová kovariance dvojic $(Y_i - \hat{Y}_i, Z_i - \hat{Z}_i)$, i = 1, …, n,

S_1 je výběrová směrodatná odchylka hodnot $Y_i - \hat{Y}_i$, i = 1, …, n,

S_2 je výběrová směrodatná odchylka hodnot $Z_i - \hat{Z}_i$, i = 1, …, n.

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

Vypočtěte výběrové parciální korelační koeficienty $r_{Y,X_1.X_2}$, $r_{Y,X_2.X_1}$, interpretujte je a porovnejte je s obyčejnými výběrovými korelačními koeficienty r_{YX_1}, r_{YX_2}.

Výpočet pomocí systému STATISTICA

Nejprve vypočteme koeficient korelace mezi výkonem a věkem.
Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 2 seznamy – 1. seznam Y, 2. seznam X_1, X_2 – Výpočet.

Proměnná	X1
Y	0,2287

Dále vypočteme parciální korelační koeficient mezi výkonem a věkem při vyloučení vlivu doby zapracovanosti.
Statistiky – Základní statistiky/tabulky – Korelační matice – OK – na záložce Detaily zvolíme Parciální korelace – 1. seznam proměnných Y, X1, druhý seznam proměnných X2 – OK

	Parciální korelace (vykony delniku.sta) S vyloučením vlivu:X2 Označ. korelace jsou významné na hlad. p < ,05000 N=10 (Celé případy vynechány u ChD)	
Proměnná	Y	X1
Y	1,000000	-0,328630
X1	-0,328630	1,000000

Korelační koeficient mezi výkonem a věkem vyšel 0,2287, tedy s rostoucím věkem roste výkon.
Parciální korelační koeficient mezi výkonem a věkem při vyloučení vlivu doby zapracovanosti vyšel -0,3286, tedy u dělníků se stejnou dobou zapracovanosti klesá s rostoucím věkem výkon.

Nyní vypočteme koeficient korelace mezi výkonem a dobou zapracovanosti:

Proměnná	X2
Y	0,4538

Dále vypočteme parciální korelační koeficient mezi výkonem a dobou zapracovanosti při vyloučení vlivu věku pracovníka.

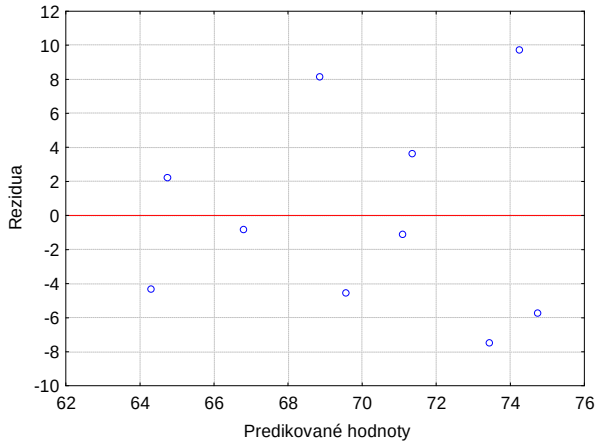
	Parciální korelace (vykony delniku.sta) S vyloučením vlivu:X1 Označ. korelace jsou významné na hlad. p < ,05000 N=10 (Celé případy vynechány u ChD)	
Proměnná	Y	X2
Y	1,000000	0,502564
X2	0,502564	1,000000

Korelační koeficient mezi výkonem a dobou zapracovanosti vyšel 0,4538, tedy čím delší doba zapracovanosti, tím lepší výkon dělník podává. Parciální korelační koeficient mezi výkonem a dobou zapracovanosti při vyloučení vlivu věku vyšel 0,5026, tedy u stejně starých dělníků je poněkud silnější přímá lineární vazba mezi výkonem a dobou zapracovanosti.

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 závislou 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.

Pravidla o počtu případů připadajících na jednu proměnnou

Při provádění mnohonásobné lineární regrese se používají dvě hlavní metody:

metoda ENTER (standardní metoda) – do modelu vstupují všechny nezávisle proměnné najednou,

metoda STEPWISE (postupná regrese) - 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 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

Při regresi počítané metodu STEPWISE by měl být poměr **1:40**. Pro model se čtyřmi proměnnými tak budeme požadovat minimálně 160 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.

Metoda ENTER

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).

Metoda STEPWISE

Je to metoda nalezení „nejlepšího“ modelu (co nejmenší počet regresorů, co nejkvalitnější predikce).

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

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), je prediktor zařazen.

Je-li statistika F menší než hodnota zvaná „F to remove“ (česky „F na vyjmutí“, ve STATISTICE implicitně 0), 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.

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 s 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.

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 systému STATISTICA se značí b*.

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:

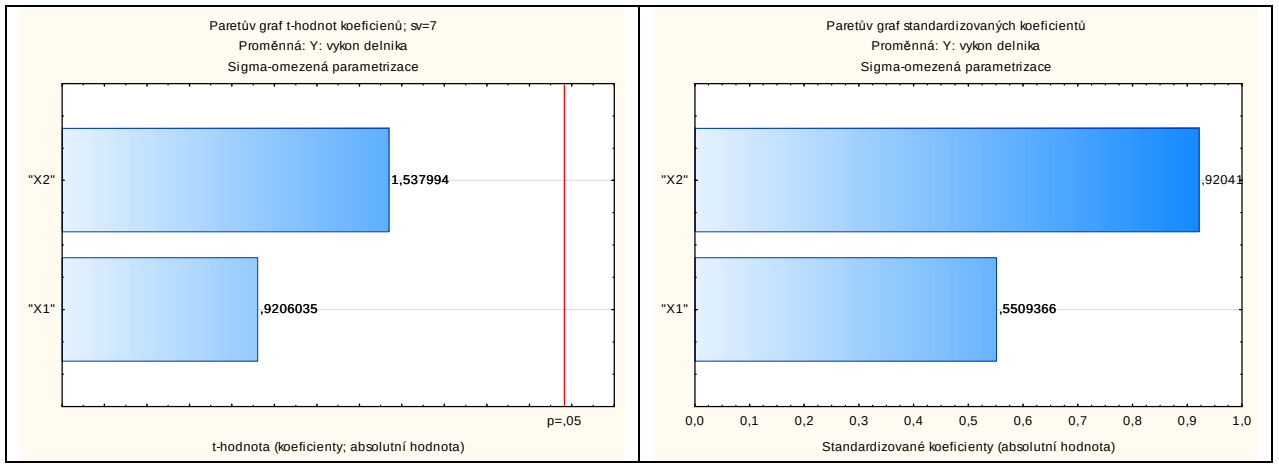
Statistiky – Vícenásobná regrese – Proměnné – Závislá proměnná Y, seznam nezáv. proměnných X1, X2 – OK – OK.

N=10	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					
	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.

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).



Postup při budování modelu mnohonásobné lineární regrese metodou 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.

Postup při budování modelu mnohonásobné lineární regrese metodou 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.

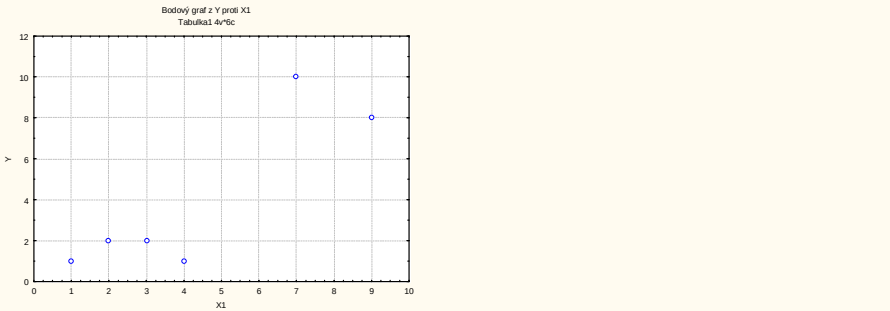
Příklad: Šest studentů gymnázia absolvovalo čtyři testy, které měří následující veličiny: X_1 - přírodovědné vědomosti, X_2 – literární vědomosti, X_3 – schopnost koncentrace, X_4 – logické myšlení. Testy se hodnotí na škále od 1 do 10 (1 = špatný výsledek, 10 = výborný výsledek).

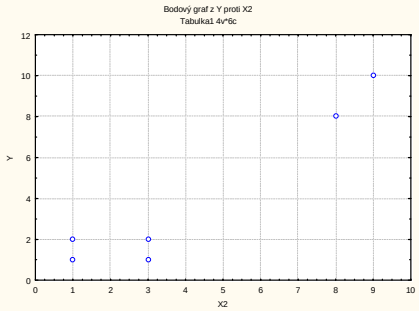
student	X_1	X_2	X_3	X_4
1	7	9	10	8
2	9	8	8	10
3	4	3	1	2
4	2	3	2	2
5	3	1	2	4
6	1	1	1	4

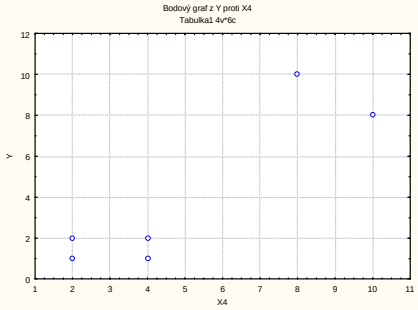
Zajímá nás, kolik bodů můžeme očekávat v testu koncentračních schopností studenta, jestliže známe výsledky testů pro literární schopnosti, přírodovědné schopnosti a logické myšlení.

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.

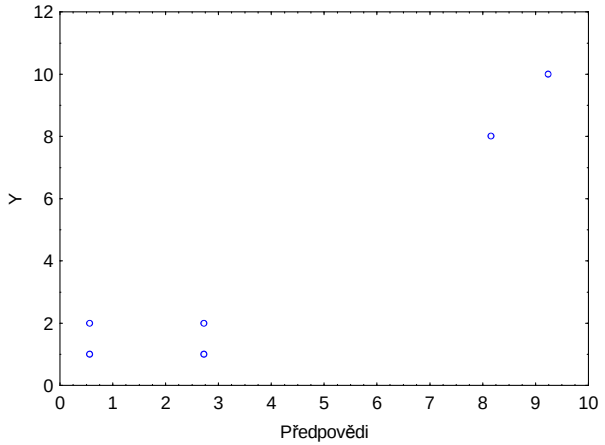
N=6	Výsledky regrese se závislou proměnnou : y (ctyri_testy.sta) R= ,98240301 R2= ,96511567 Upravené R2= ,91278918 F(3,2)=18,444 p<,05187 Směrod. chyba odhadu : 1,1664					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(2)	p-hodn.
Abs.člen			-1,08961	0,941927	-1,15679	0,366858
X1	-0,299065	0,368366	-0,38391	0,472872	-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

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, ..., 6. Metodou nejmenších čtverců opět získáme odhady regresních parametrů.

N=6	Výsledky regrese se závislou proměnnou : X3 (ctyri testy.sta) R= ,95813306 R2= ,91801897 Upravené R2= ,89752371 F(1,4)=44,792 p<,00259 Směrod. chyba odhadu : 1,2644					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(4)	p-hodn.
Abs.člen			-0,520548	0,850099	-0,612338	0,573413
X2	0,958133	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.

Vztah mezi naměřenými a predikovanými 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.

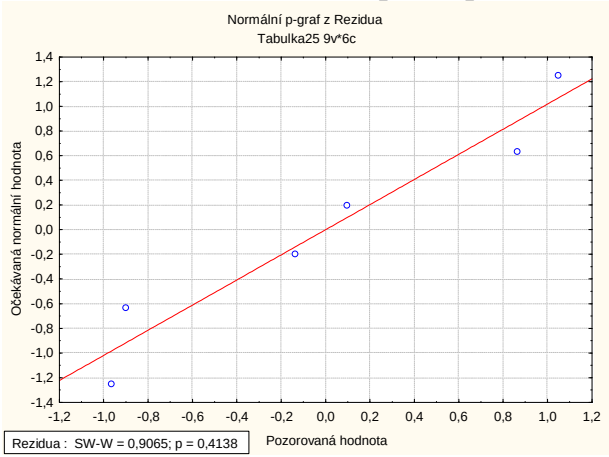
N=6	Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta) R= ,95813306 R2= ,91801897 Upravené R2= ,89752371 F(1,4)=44,792 p<,00259 Směrod. chyba odhadu : 1,2644					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			-0,520548	0,850099	-0,612338	0,573413
X2	0,958133	0,143162	1,084932	0,162108	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 proměnné X_2 a X_4 nikoliv. Přispívají však k vysvětlení variability hodnot závisle proměnné veličiny Y. Adjustovaný 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.