Osnova přednášky Vícenásobná lineární regrese

1. Popis modelu vícenásobné 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:

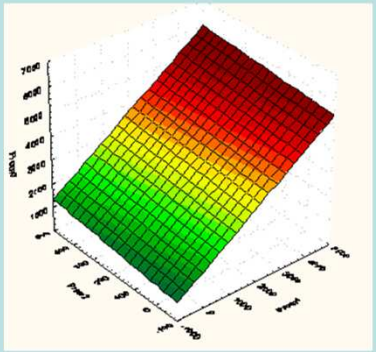
β_0 … teoretická hodnota závisle proměnné veličiny při nulových hodnotách všech nezávisle proměnných veličin,

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“:

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

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.

Příklady vícenásobné regrese

Lékaře zajímá, jak krevní tak Y závisí na věku pacienta X_1, na jeho BMI X_2 a na množství vypitého alkoholu X_3.

Majitele realitní kanceláře zajímá, jak cena bytu Y závisí na velikosti bytu X_1, na počtu pokojů X_2, vzdálenosti bytu od centra města X_3 a existenci vlastního parkovacího místa X_4 (1 – ano, 0 – ne).

Pěstitele brambor zajímá, jak výnos Y jisté odrůdy brambor závisí na množství dodaného hnojiva X_1, na množství srážek X_2 ve vegetačním období a na teplotě půdy X_3.

Ekonoma zajímá, jak výdaje domácnosti za potraviny a nápoje Y závisí na čistém příjmu domácnosti X_1 a na počtu členů domácnosti X_2.

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 vícenásobné lineární regrese

2.1. Kroky před prováděním vícenásobné lineární regrese

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

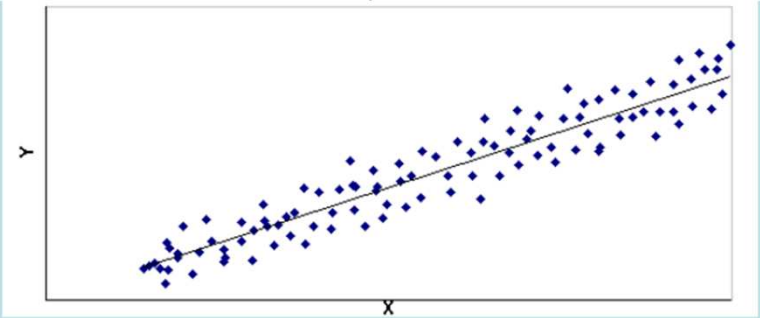
b) Pokud je nesplňují, posoudíme, jak vážné je porušení těchto předpokladů.

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

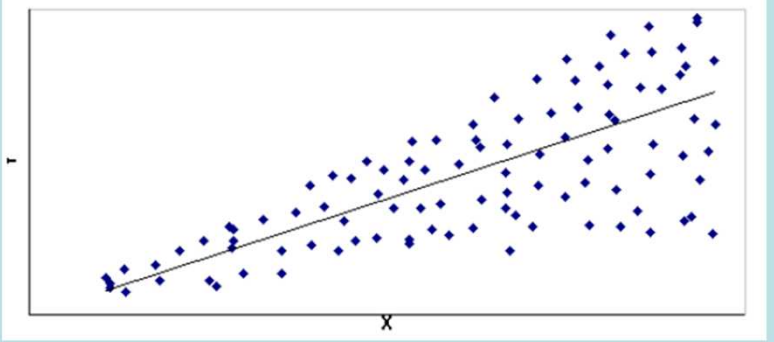
2.2. Sedm hlavních předpokladů regresní analýzy

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.)

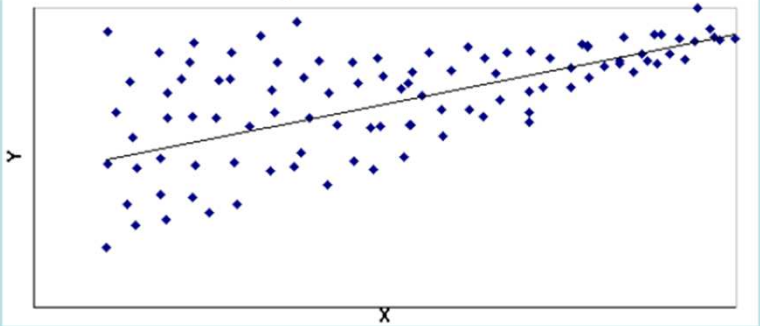
Ukázka homoskedastických dat:



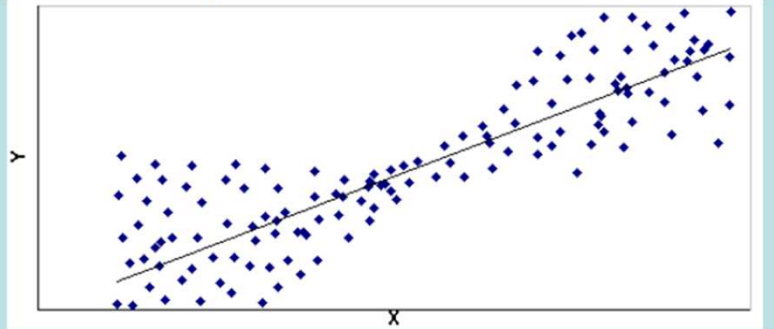
Ukázka dat s rostoucí heteroskedasticitou:



Ukázka dat s klesající heteroskedasticitou:



Ukázka dat s proměnlivou heteroskedasticitou:

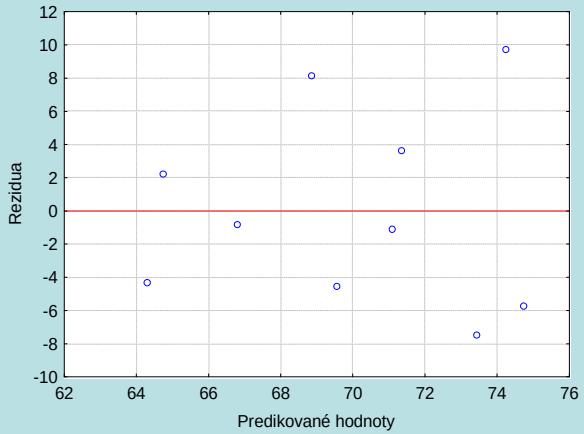


2.3. Ověřování předpokladů modelu

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.

Informace o multikolinearitě poskytuje 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. Užitečný je též ukazatel tolerance. Hodnoty pod 0,2 svědčí o multikolinearitě.

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

Řešení:

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, ukazatel tolerance je větší než 0,2, 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

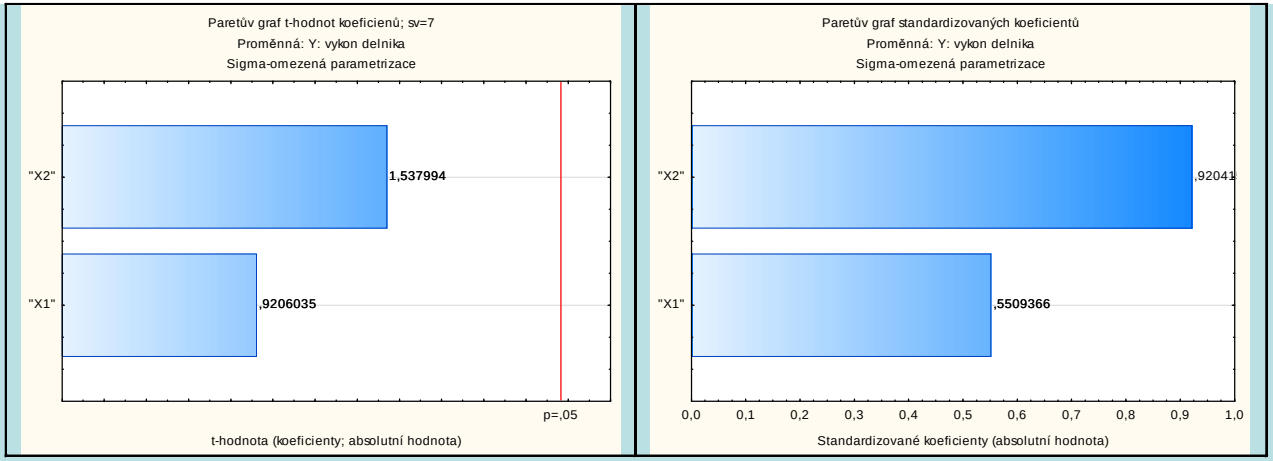
Chceme-li porovnávat vliv, jaký mají proměnné $x_1, ..., x_p$ v modelu $\mathbf{Y = X\beta + \varepsilon}$, můžeme spočítat tzv. standardizované regresní parametry, kterým se také říká B-koeficienty (nebo také beta koeficienty). Zavedeme proto standardizované veličiny

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

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.



2.5. Pravidla pro stanovení počtu pozorování v závislosti na počtu prediktorů

a) Pokud nás přednostně zajímá variabilita vysvětlená modelem (tj. index determinace), mělo by platit $n \geq 50 + 8p$. Např. pro 4 prediktory bychom měli mít aspoň 82 pozorování.

b) Zajímají-li nás odhady regresních parametrů (tj. jde nám především o predikci), mělo by platit $n \geq 104 + p$. Např. pro 4 prediktory bychom měli mít aspoň 108 pozorování.

c) Nejnižší možný poměr proměnná/počet případů je **1:5**. V tom případě ale platí silný požadavek na normalitu – rozložení reziduí by mělo být normální. Při čtyřech prediktorech bychom měli mít aspoň 20 pozorování.

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 odhadů standardizovaných regresních parametrů a interpretujte nestandardizované odhady regresních parametrů.

Řešení:

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

Odhady standardizovaných regresních parametrů 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.
Odhad konstanty je 86,74, což znamená, že dělníka s věkem 0 a dobou zapracovanosti 0 by hodinový výkon byl 86,74.
Odhad parametru pro věk je -0,7, tedy při konstantní době zapracovanosti by se při zvýšení věku o rok výkon snížil o 0,7.
Odhad parametru pro dobu zapracovanosti je 1,35, což znamená, že při konstantním věku by se při zvýšení doby zapracovanosti o rok zvýšil výkon o 1,35.

Příklad řešený v systému R

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

Úkolem je provést v systému R regresní analýzu modelu $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Data jsou uložena v souboru vykony_delniku.txt.

Načtení dat a pojmenování proměnných:

```
data<-read.delim('vykony delniku.txt',sep=' ', header=T)
Y<-data$Y
X1<-data$X1
X2<-data$X2
```

Výpočet korelační matice:

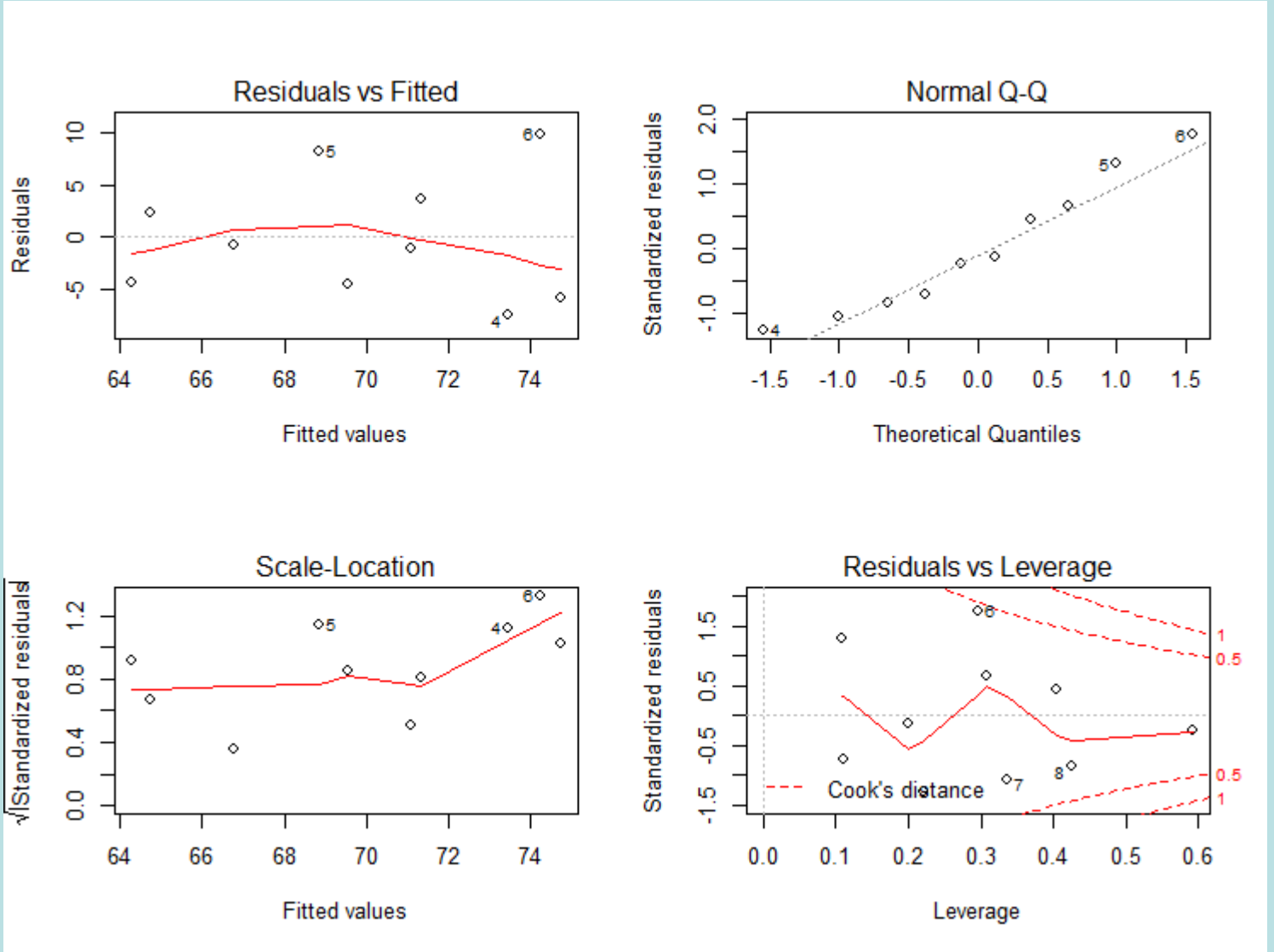
```
cor(data)
          Y        X1        X2
Y  1.000000 0.2286800 0.4537570
X1 0.228680 1.0000000 0.8470271
X2 0.453757 0.8470271 1.0000000
```

Vytvoření modelu a jeho výstup:

```
vystup1<-lm(Y~X1+X2)
> summary(vystup1)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4367 -4.4717 -0.9345  3.3150  9.7630

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.7422    25.3240   3.425   0.0111 *
X1           -0.7003     0.7607  -0.921   0.3879
X2            1.3506     0.8782   1.538   0.1679
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.649 on 7 degrees of freedom
Multiple R-squared:  0.2917,     Adjusted R-squared:  0.08927
F-statistic: 1.441 on 2 and 7 DF,  p-value: 0.2991
```

Ověření předpokladů modelu:

```
par(mfrow=c(2,2))
plot(vystup1)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Testování normality reziduí:

```
shapiro.test(vystup1$residuals)

      Shapiro-Wilk normality test

data:  vystup1$residuals
W = 0.9355, p-value = 0.5041
```

Provedení Durbinova – Watsonova testu autokorelovanosti reziduí:

```
durbinWatsonTest(vystup1)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1974273      2.376259   0.446
 Alternative hypothesis: rho != 0
```

Výpočet koeficientu VIF:

```
library(car)
vif(vystup1)
      X1       X2
3.539258 3.539258
```

Pro posouzení vlivu jednotlivých nezávisle proměnných v modelu vytvoříme standardizovaná data:

```
sdata<-scale(data)
```

Pojmenujeme jednotlivé standardizované proměnné:

```
sY<-sdata[,1]
sX1<-sdata[,2]
sX2<-sdata[,3]
```

Vytvoříme model se standardizovanými proměnnými a podíváme se na jeho výstup:

```
svystup1<-lm(sY~sX1+sX2)
summary(svystup1)

Call:
lm(formula = sY ~ sX1 + sX2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0674 -0.6418 -0.1341  0.4758  1.4012

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.118e-15  3.018e-01   0.000    1.000
sX1         -5.509e-01  5.985e-01  -0.921    0.388
sX2          9.204e-01  5.985e-01   1.538    0.168

Residual standard error: 0.9543 on 7 degrees of freedom
Multiple R-squared:  0.2917,    Adjusted R-squared:  0.08927
F-statistic: 1.441 on 2 and 7 DF,  p-value: 0.2991
```

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).

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.

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 vícenásobné 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:

U 15 žen ve věkové kategorii 20 – 50 let byly zjištěny tyto údaje: tělesná hmotnost (v kg – závisle proměnná veličina Y), tělesná výška (v cm – nezávisle proměnná veličina neboli regresor X_1), věk (v letech, regresor X_2):

	1 Y	2 X1	3 X2
1	52,2	147,3	45
2	53,1	149,9	25
3	54,4	152,4	39
4	55,8	154,9	43
5	57,1	157,5	35
6	58,5	160,0	38
7	59,9	162,6	36
8	61,2	165,1	47
9	63,0	167,6	34
10	64,4	170,2	23
11	66,2	172,7	37
12	68,0	175,3	42
13	69,8	177,8	32
14	72,1	180,3	49
15	74,4	182,9	26

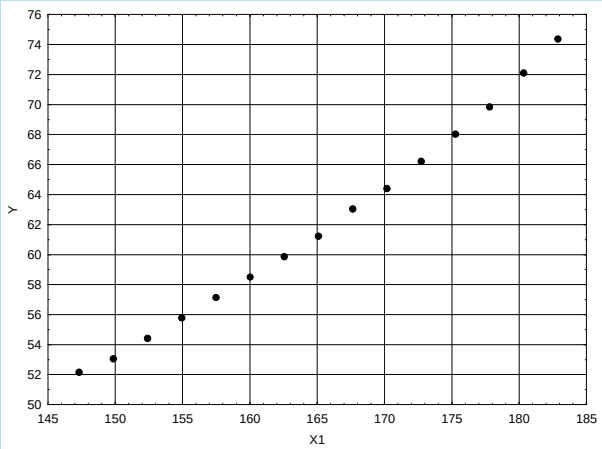
Zajímá nás, jakou hmotnost u ženy můžeme očekávat, jestliže známe její výšku a věk

Řešení:

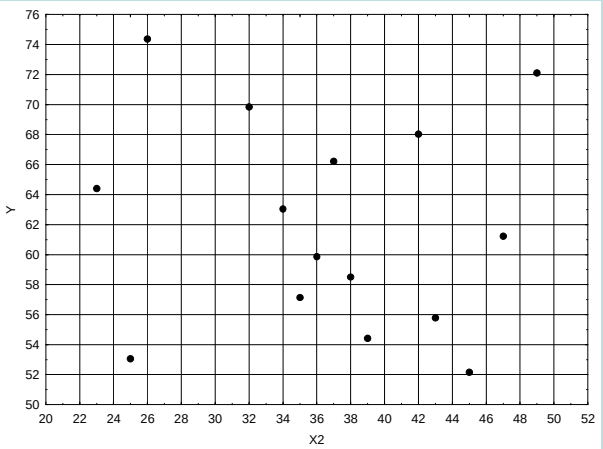
Sestavíme regresní model $Y_i = \beta_0 + \beta_1 x_{il} + \beta_2 x_{i2} + \varepsilon_i$, i = 1, ..., 15.

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

Závislost hmotnosti na výšce



Závislost hmotnosti na věku



Dále spočteme koeficienty korelace

mezi hmotností a výškou: $r_{YX1} = 0{,}9955$

mezi hmotností a věkem: $r_{YX2} = -0{,}108$

mezi výškou a věkem: $r_{X1X2} = -0{,}109$

Vidíme, že korelace mezi hmotností a výškou je přímá a velmi silná, avšak mezi hmotností a věkem či mezi výškou a věkem je nepřímá a jen slabá.

Metodou nejmenších čtverců získáme odhady regresních parametrů.

N=15	Výsledky regrese se závislou proměnnou : Y (Tabulka1) R= ,99549503 R2= ,99101036 Upravené R2= ,98951208 F(2,12)=661,43 p<,00000 Směrod. chyba odhadu : ,71983					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(12)	p-hodn.
Abs.člen			-39,7220	3,051469	-13,0174	0,000000
X1	0,995574	0,027535	0,6160	0,017038	36,1565	0,000000
X2	0,000729	0,027535	0,0006	0,024510	0,0265	0,979327

Empirická regresní funkce má tvar $\hat{Y} = -39{,}722 + 0{,}616x_1 + 0{,}0006x_2$. Variabilita proměnné Y je z 99,1 % vysvětlená zvoleným regresním modelem. Pro $\alpha = 0{,}05$ je celkový F-test významný, dílčí t-testy pro β_0 a β_1 rovněž, avšak parametr β_2 je nevýznamný. Podíváme-li se na beta koeficienty, vidíme, že největší vliv má proměnná X_1. Sestavíme tedy nový model $Y_i = \beta_0 + \beta_2 x_{i1} + \varepsilon_i$, i = 1, ..., 15. Metodou nejmenších čtverců opět získáme odhady regresních parametrů.

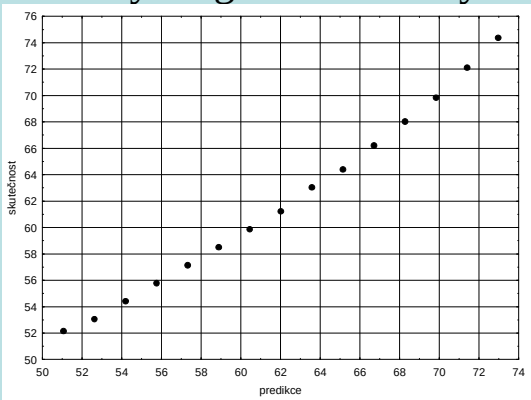
N=15	Výsledky regrese se závislou proměnnou : Y (Tabulka1) R= ,99549477 R2= ,99100983 Upravené R2= ,99031828 F(1,13)=1433,0 p<,00000 Směrod. chyba odhadu : ,69161					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(13)	p-hodn.
Abs.člen			-39,6901	2,692492	-14,7410	0,000000
X1	0,995495	0,026297	0,6160	0,016272	37,8553	0,000000

Nyní má empirická regresní funkce tvar $\hat{Y} = -39{,}6901 + 0{,}616x_1$, model jako celek je významný a nezávisle proměnná X_1 rovněž. Adjustovaný index determinace je 0,9903, což je větší než adjustovaný index determinace 0,9895 v původním modelu.

Pro kontrolu kvality regrese porovnáme zjištěné a predikované hodnoty veličiny Y.

	1 skutečnost	2 predikce
1	52,2	51,1
2	53,1	52,6
3	54,4	54,2
4	55,8	55,8
5	57,1	57,3
6	58,5	58,9
7	59,9	60,4
8	61,2	62,0
9	63,0	63,6
10	64,4	65,1
11	66,2	66,7
12	68,0	68,3
13	69,8	69,8
14	72,1	71,4
15	74,4	73,0

Tečkový diagram naměřených a predikovaných hodnot:



Provedeme-li krokovou dopřednou metodu, skončí hned v 1. kroku a vybere X_1.
Provedeme-li krokovou zpětnou metodu, skončí také v 1. kroku a vyloučí proměnnou X_2.