KORELACIJA

Korelacija slučajnih varijabli

Za slučajne varijable X i Y **kovarijanca** se definira kao

$$Cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right].$$

Kovarijanca mjeri zajedničku varijaciju varijabli X i Y.

Kovarijanca je izražena u mjernim jedinicama slučajnih varijabli X i Y.

Korelacija:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Korelacija nema mjernu jedinicu.

$$-1 \leq Corr(X, Y) \leq 1$$

Ukoliko su X i Y **nezavisne** slučajne varijable, tada je

$$E(X \cdot Y) = E(X) \cdot E(Y),$$

a posebno je

$$\begin{aligned} Cov(X, Y) &= E\left[(X - E(X))(Y - E(Y))\right] = \\ &= E\left[X - E(X)\right] \cdot E\left[Y - E(Y)\right] = \\ &= 0 \cdot 0 = 0. \end{aligned}$$

Isto vrijedi i za korelaciju:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y} = 0.$$

Ukoliko su X i Y **linearno povezane** slučajne varijable:

$$Y = a \cdot X + b$$

tada je

$$Var(Y) = a^2 Var(X), \quad \text{tj.} \quad \sigma_Y = |a|\sigma_X$$

i

$$\begin{aligned} Cov(X, Y) &= E\left[(X - E(X))(Y - E(Y))\right] = \\ &= E\left[(X - E(X))(a \cdot X + b - a \cdot E(X) - b)\right] = \\ &= E\left[(X - E(X))(a \cdot X - a \cdot E(X)\right] = \\ &= aE\left[(X - E(X))(X - E(X)\right] = \\ &= a \cdot Var(X). \end{aligned}$$

Korelacija:

$$
\begin{aligned}
Corr(X,Y) &= \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \\
&= \frac{a \cdot Var(X)}{\sigma_X \, |a| \, \sigma_X} = \\
&= \frac{a}{|a|} = \begin{cases} 1, & \text{za } a > 0; \\ -1, & \text{za } a < 0. \end{cases}
\end{aligned}
$$

Slučajne varijable X i Y **nezavisne** $\rightarrow$ $Corr(X, Y) = 0$.

Slučajne varijable X i Y **linearno zavisne** $\rightarrow$ $Corr(X, Y) = \pm 1$.

Uočimo da $Corr(X, Y) = 0$ ne znači da su X i Y nezavisne slučajne varijable.

Pearsonov koeficijent korelacije

Promatramo dva obilježja X i Y u populaciji veličine N.

Pearsonov koeficijent korelacije obilježja X i Y:

$$\rho = \frac{\frac{1}{N}\sum(x_i - \mu_x)(y_i - \mu_y)}{\sigma_X \cdot \sigma_Y}$$

σ_X - standardna devijacija obilježja X

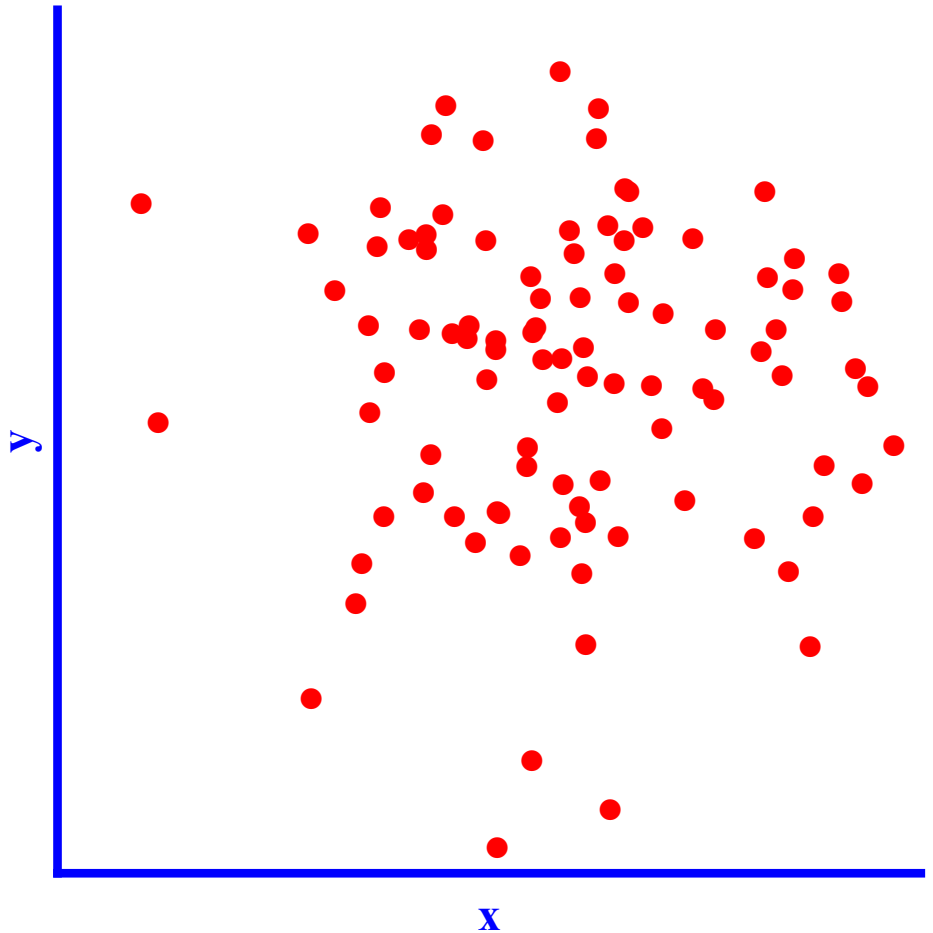
σ_Y - standardna devijacija obilježja Y

σ_X - standardna devijacija obilježja X

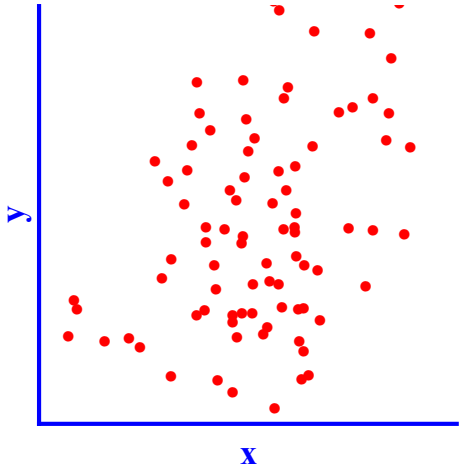
σ_Y - standardna devijacija obilježja Y

Dijagram raspršenja

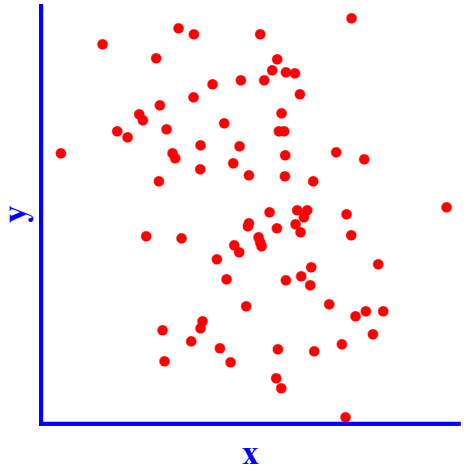
Za verijable X i Y promatramo (X, Y)-graf:



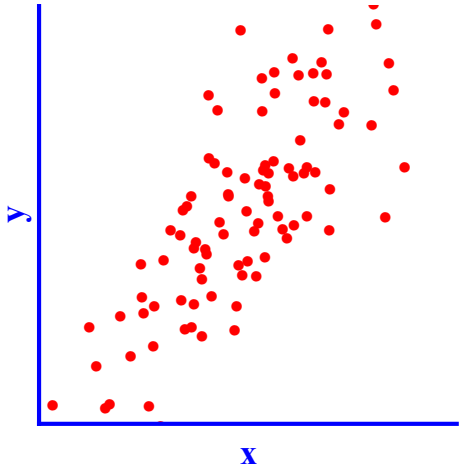
$\rho = 0$



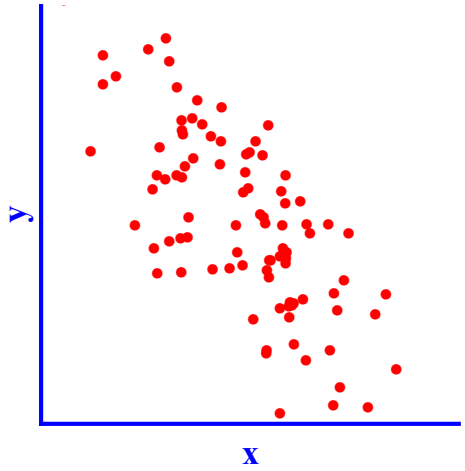
$\rho = 0.47$



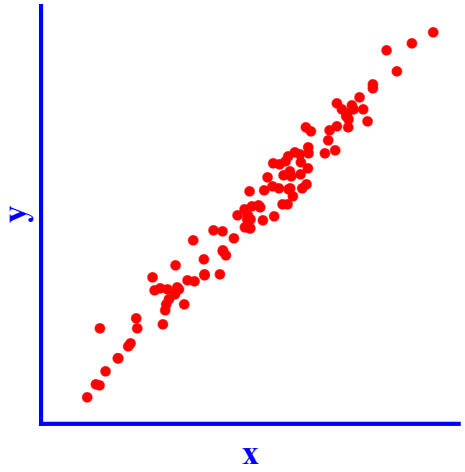
$\rho = -0.46$



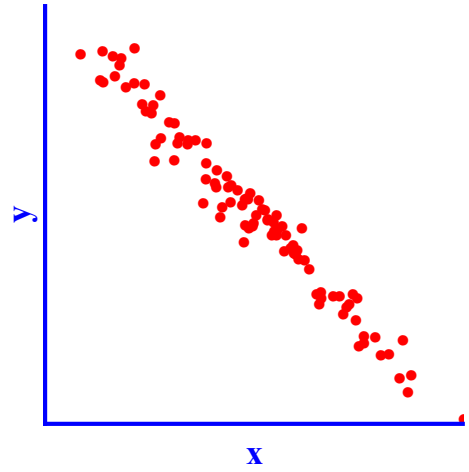
$\rho = 0.80$



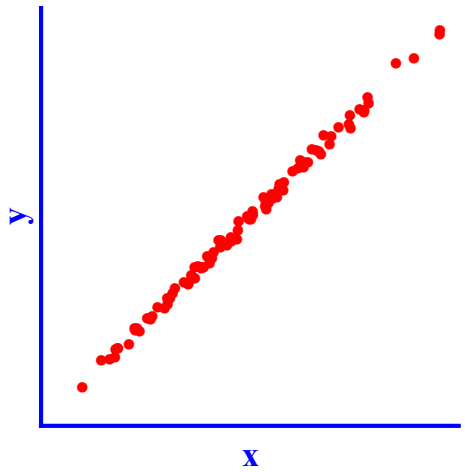
$\rho = -0.82$



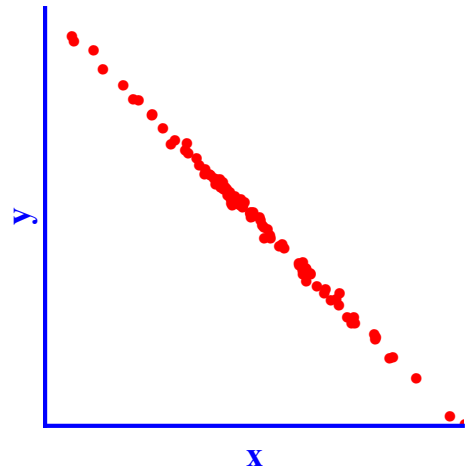
$\rho = 0.98$



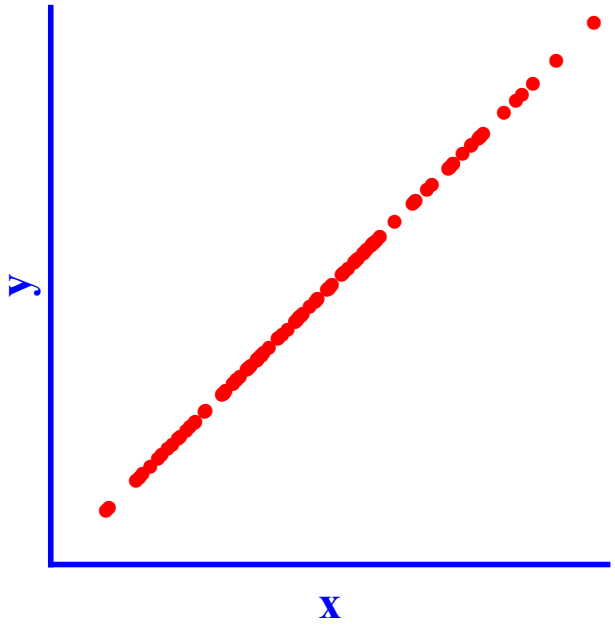
$\rho = -0.98$



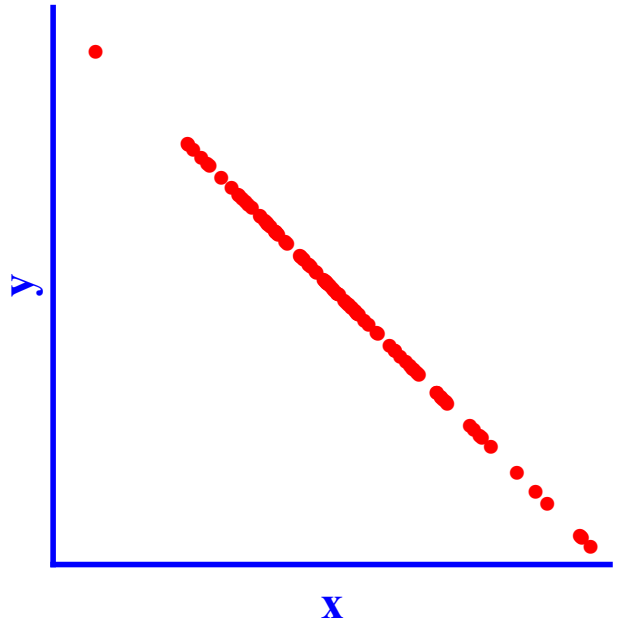
$\rho = 0.999$



$\rho = -0.999$



$\rho = 1$



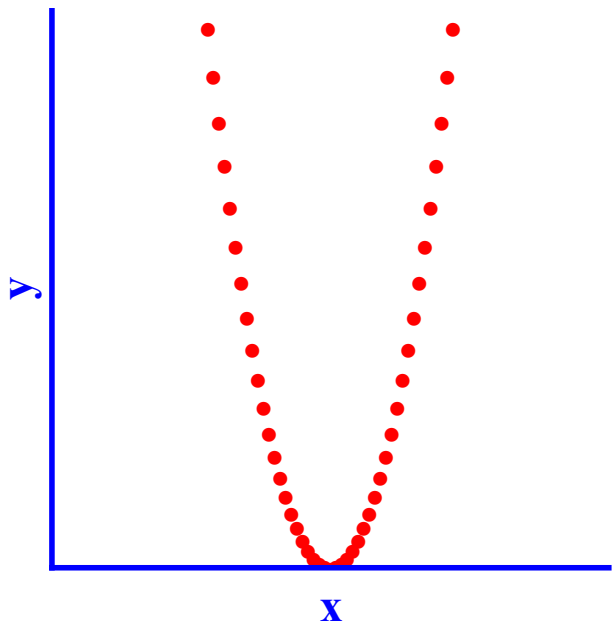
$\rho = -1$

Potpuna linearna povezanost!

Pearsonov koeficijent koleracije mjeri **linearnu** povezanost dvije varijable.

$\longrightarrow$ **koeficijent linearne korelacije**

Primjer nelinearne povezanost:



$$\rho = 0, \quad y = x^2$$

Pearsonov koeficijent korelacije:

- broj iz intervala $[-1, 1]$
- iskazuje smjer i jakost linarne statističke veze između dvije pojave
- r bliži -1 ili 1 $\longrightarrow$ jača korelacija
- $r = 1$ ili $r = -1 \longrightarrow$ potpuna povezanost, funkcionalna povezanost
- $r > 0 \longrightarrow$ pozitivna korealcija (veći $x \rightarrow$ veći y)
- $r < 0 \longrightarrow$ negativna korealcija (veći $x \rightarrow$ manji y)
-
  - $0 - 0.25$ - linearna korelacija slaba
  - $0.25 - 0.64$ - korelacija srednje jačine
  - $0.64 - 1$ - čvrsta korelacija

Procjena Pearsonova koeficijenta korelacije

n - veličina uzorka

Uzorak: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

$S_X = \sqrt{S_X^2}$ - procjena standardne devijacije za obilježje X

$S_Y = \sqrt{S_Y^2}$ - procjena standardne devijacije za obilježje Y

Procjena Pearsonova koeficijenta korelacije:

$$r = \frac{\frac{1}{n-1} \sum_i \left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right)}{S_X \cdot S_Y}$$

Testiranje hipoteze o koeficijentu korelacije

Na osnovu uzorka $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ možemo testirati hipotezu

$$H_0: \quad \rho = 0$$

gdje je ρ Pearsonov koeficijent korelacije (za populaciju).

Statistika

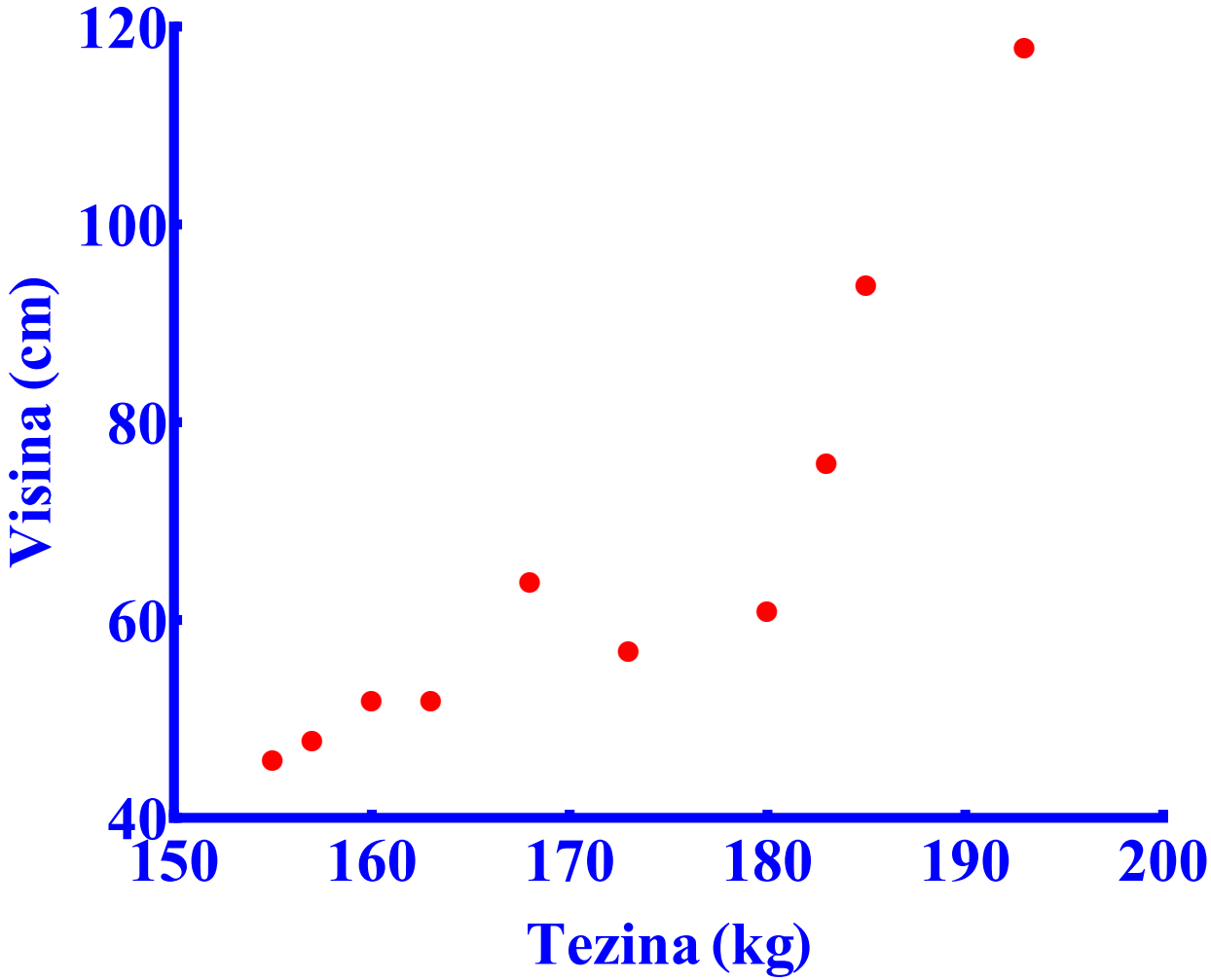
$$t = \frac{\sqrt{n-2} \cdot r}{\sqrt{1-r^2}}$$

je distribuirana prema Studentovoj razdiobi: $t \sim t(n-2)$.

Primjer. Na osnovu uzorka od 10 osoba procijenite koeficijent korelacije za visinu i težinu.

Visina (cm)	Težina (kg)
183	76
163	52
180	61
168	64
160	52
157	48
185	94
155	46
193	118
173	57

Dijagram raspršenja:



$r = 0.89$

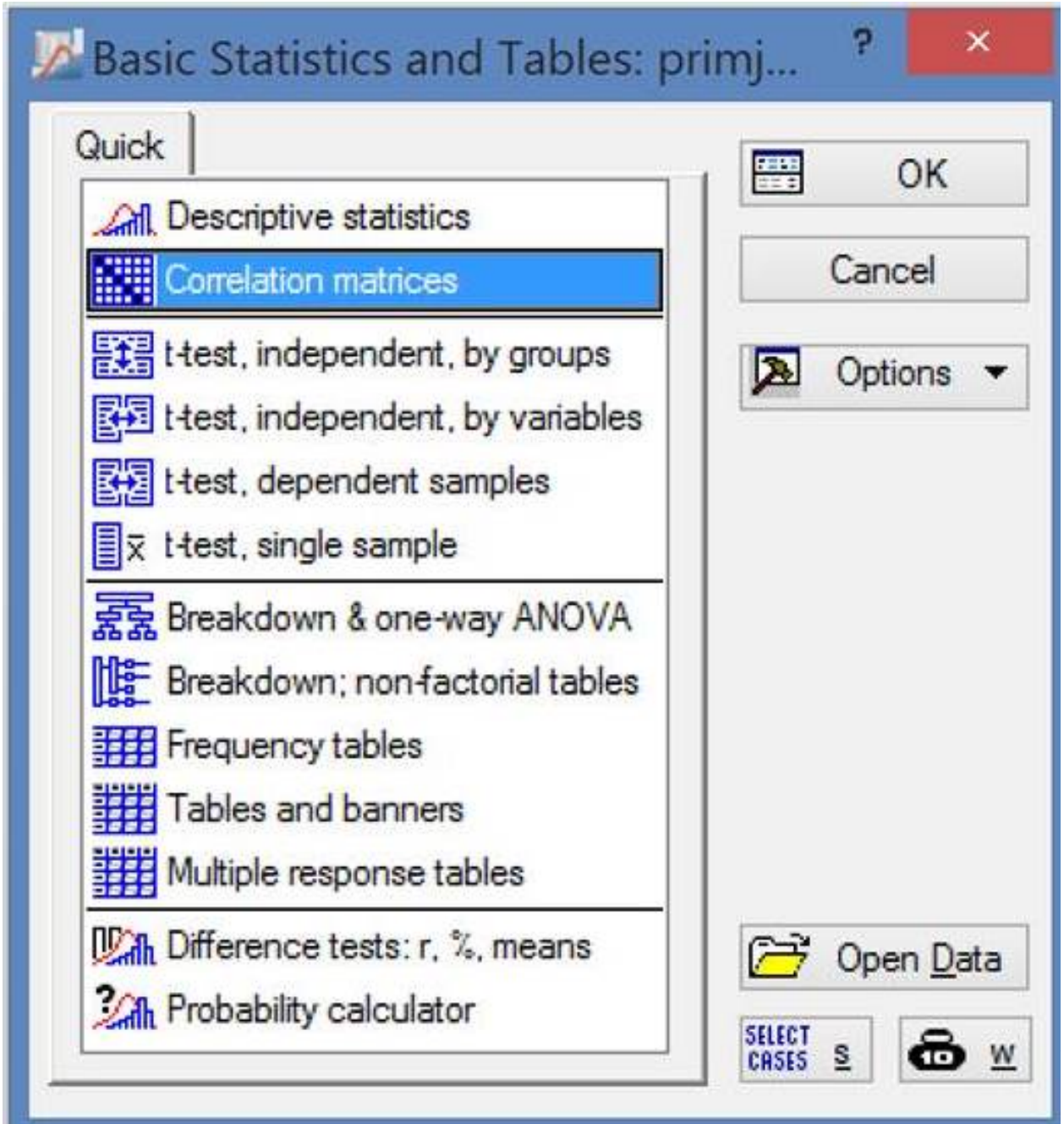
Statistica

Podaci:

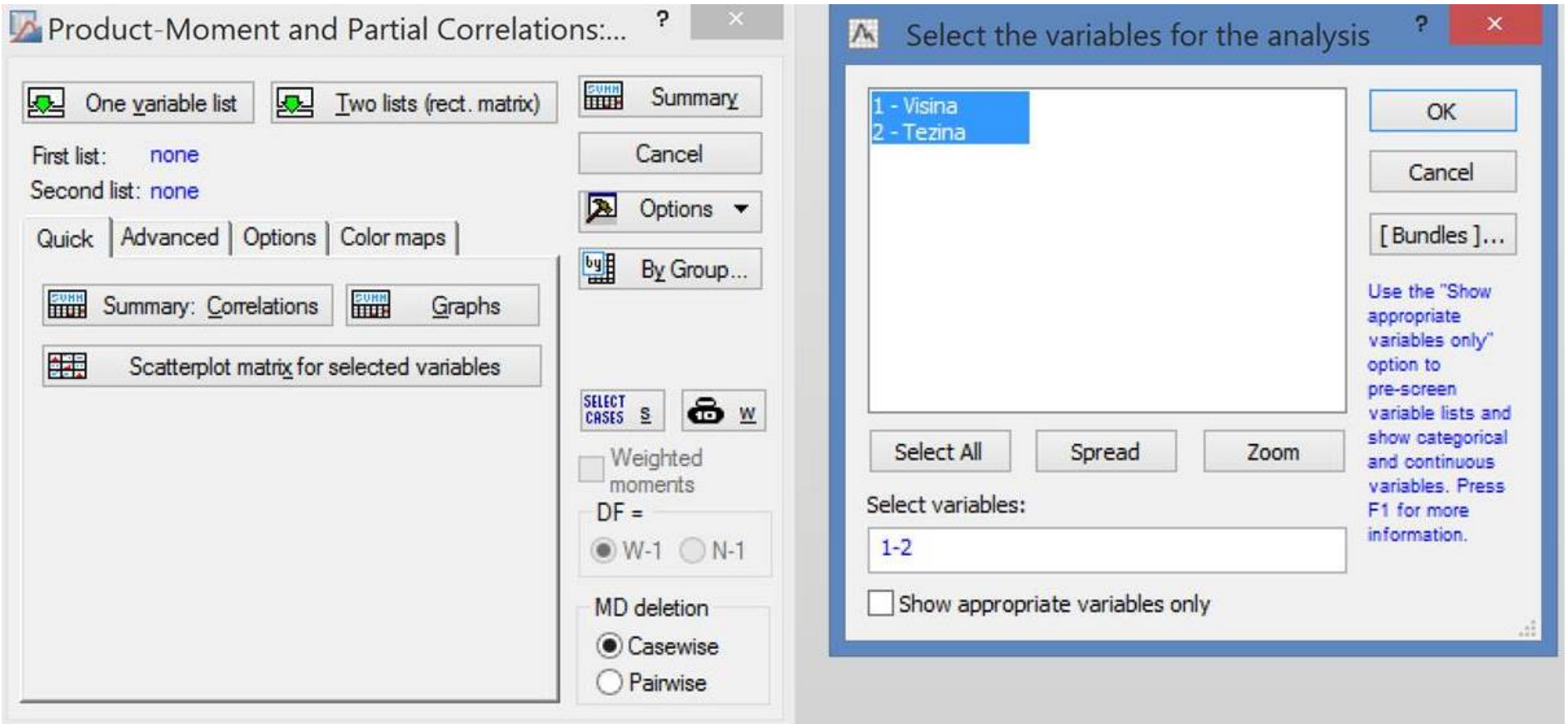
Data: primjer-7-1.sta* (2v ...

	1 Visina	2 Tezina
1	183	76
2	163	52
3	180	61
4	168	64
5	160	52
6	157	48
7	185	94
8	155	46
9	193	118
10	173	57

Basic Statistics and Tables:



Izbor varijabli:



Rezultat:

Workbook2* - Correlations (primjer-7-1.sta)

Correlations (primjer-7-1.sta)
Marked correlations are significant at p < ,05000
N=10 (Casewise deletion of missing data)

Variable	**Means**	Std.Dev.	Visina	Tezina
Visina	171,7000	13,12377	1,000000	0,890661
Tezina	66,8000	23,11229	0,890661	1,000000

Interpretacija:

- Kod osobe s većom visinom očekujemo i veću težinu (pozitivna koreliranost)
- Kod osobe s većom težinom očekujemo i veću visinu (pozitivna koreliranost)
- Korelacija ne pokazuje uzročno-posljedičnu povezanost!
- Povećanjem težine nećemo povećati visinu.

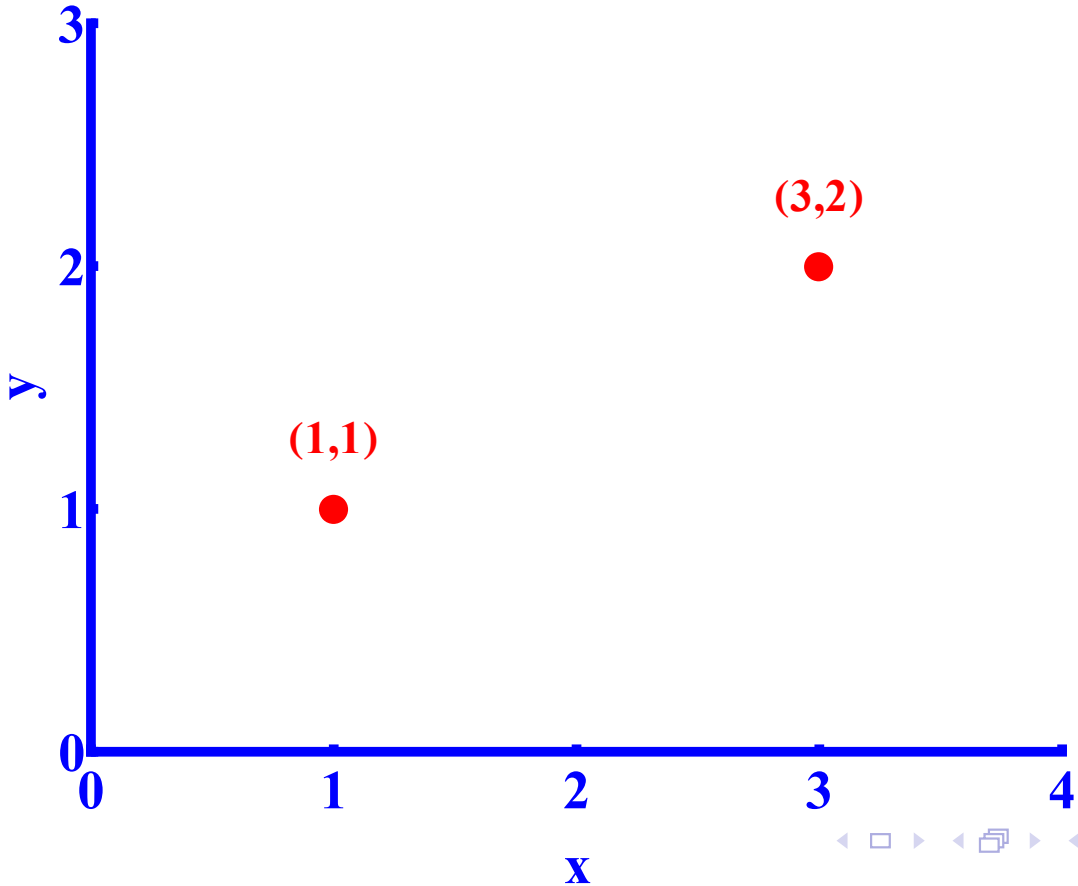
LINEARNA REGRESIJA

Jednostavna linearna regresija

Pravac

Primjer. Nacrtajte pravac koji prolazi kroz točke $(1, 1)$ i $(3, 2)$.

Nacrtamo točke:

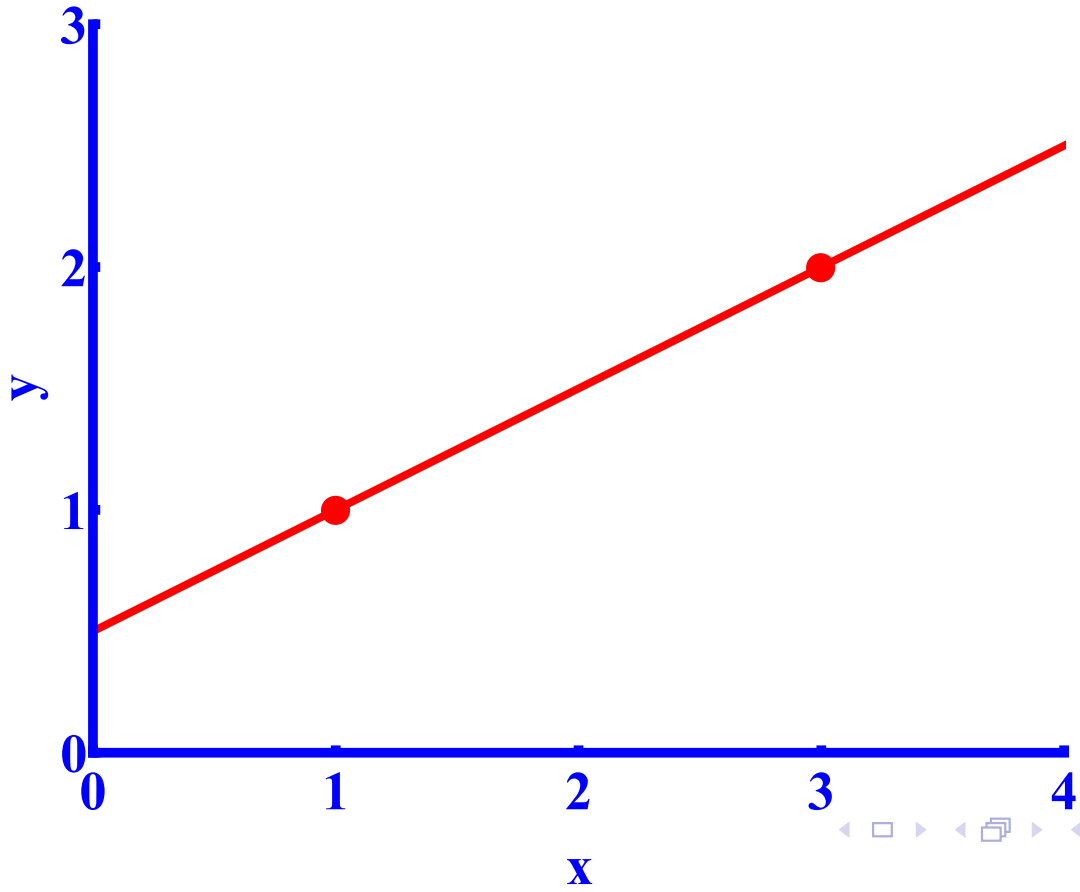


Jednostavna linearna regresija

Pravac

Primjer. Nacrtajte pravac koji prolazi kroz točke $(1, 1)$ i $(3, 2)$.

Nacrtamo točke i provučemo pravac kroz njih:



Primjer. Nacrtajte pravac $y = 2 \cdot x - 1$.

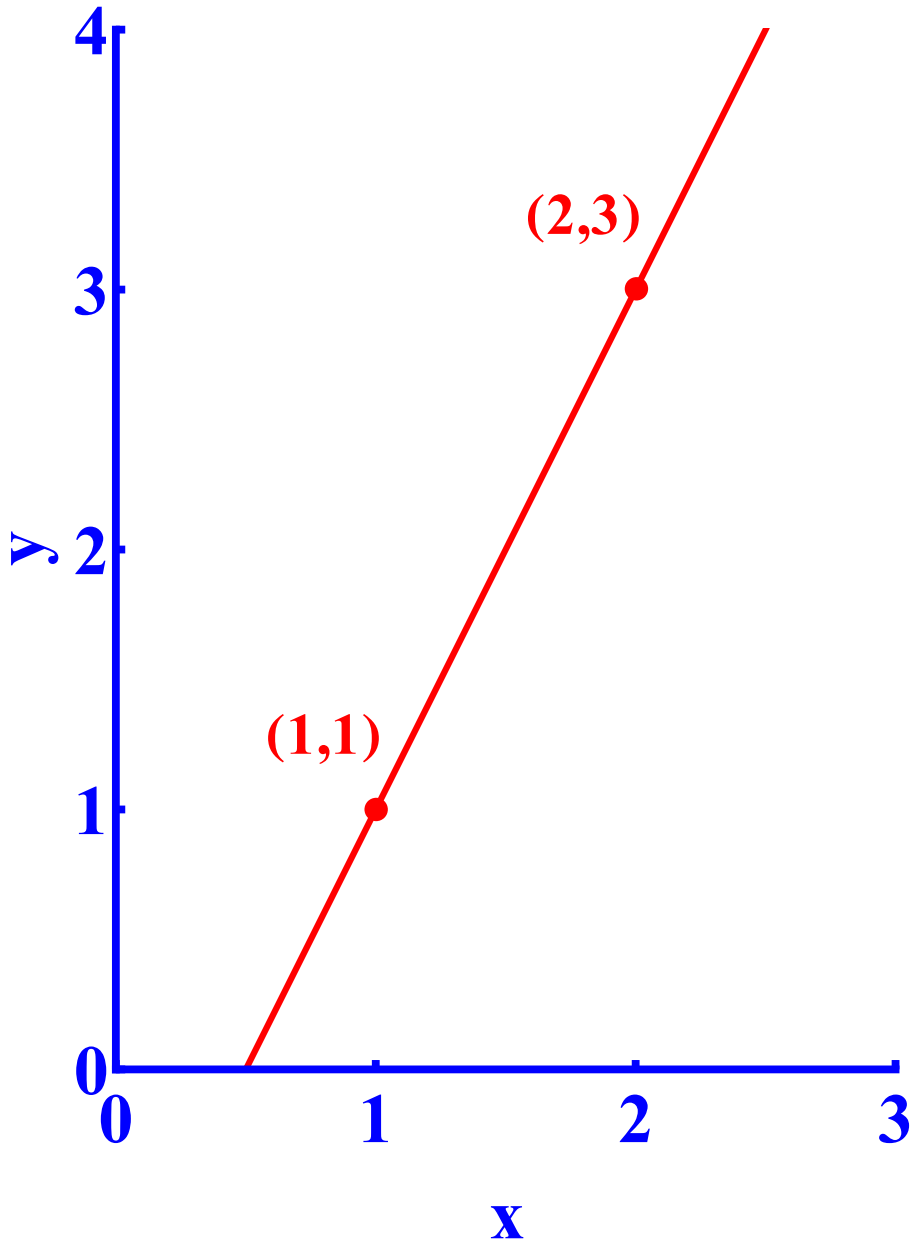
Odredimo dvije točke na pravcu:

$$x = 1 \quad \Longrightarrow \quad y = 2 \cdot 1 - 1 = 1$$

$$x = 2 \quad \Longrightarrow \quad y = 2 \cdot 2 - 1 = 3$$

Točke: $(1, 1)$ i $(2, 3)$.

Nacrtamo točke i provučemo pravac kroz dvije točke.


4
3
2
1
0
y
0
1
2
3
x
(2,3)
(1,1)

Primjer. Nacrtajte pravac $y = -3 \cdot x + 8$.

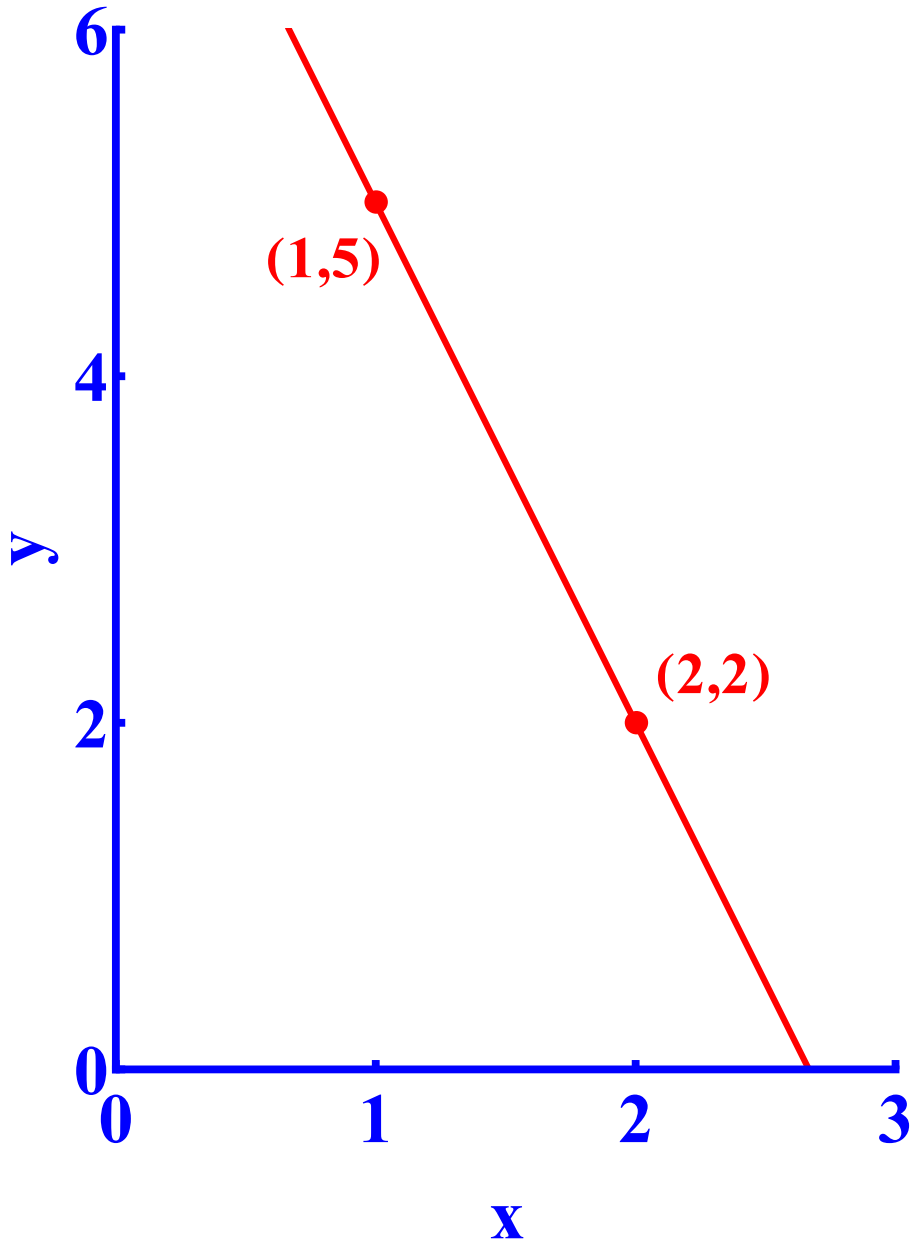
Odredimo dvije točke na pravcu:

$$x = 1 \quad \Longrightarrow \quad y = -3 \cdot 1 + 8 = 5$$

$$x = 2 \quad \Longrightarrow \quad y = -3 \cdot 2 + 8 = 2$$

Točke: $(1, 5)$ i $(2, 2)$.

Nacrtamo točke i provučemo pravac kroz dvije točke.


6
4
2
0
y
(1,5)
(2,2)
0
1
2
3
x

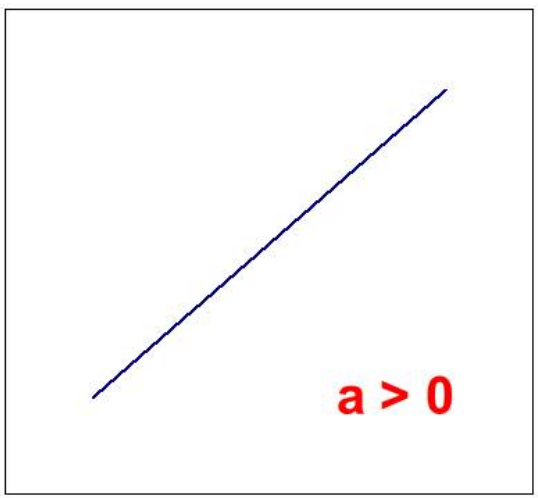
Jednadžba pravca:

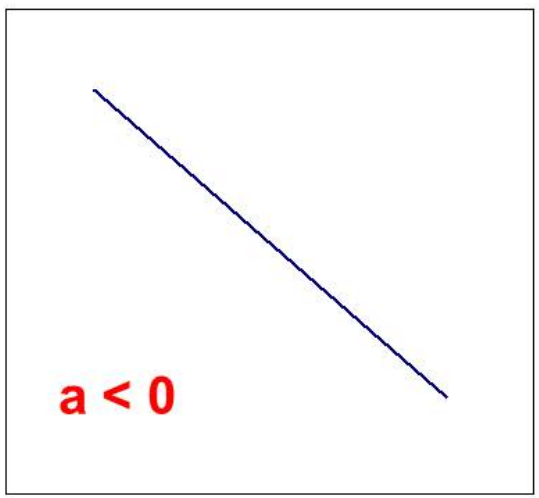
$$y = a \cdot x + b.$$

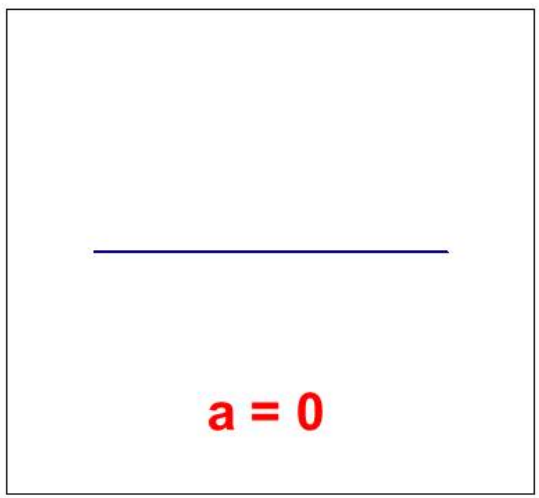
a - koeficijent smjera (*engl.* 'slope')

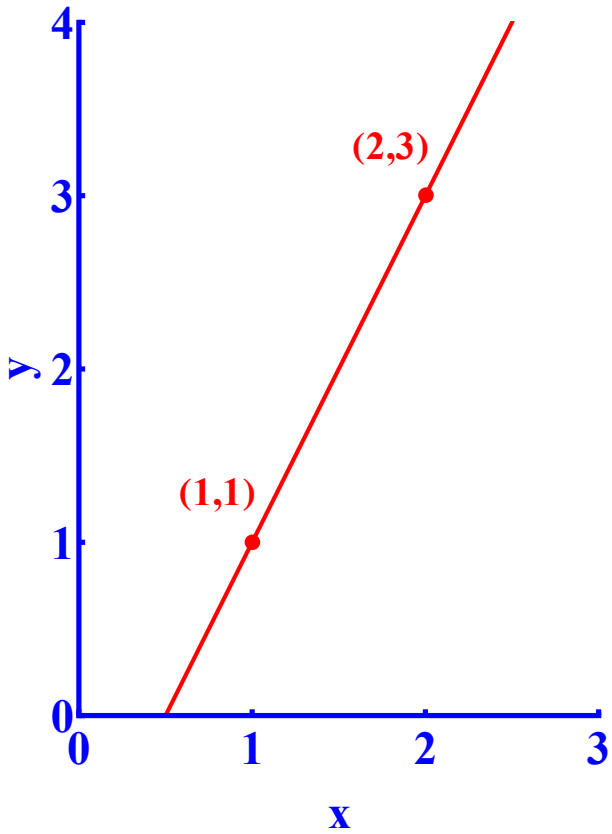
b - slobodni koeficijent (*engl.* 'intercept')

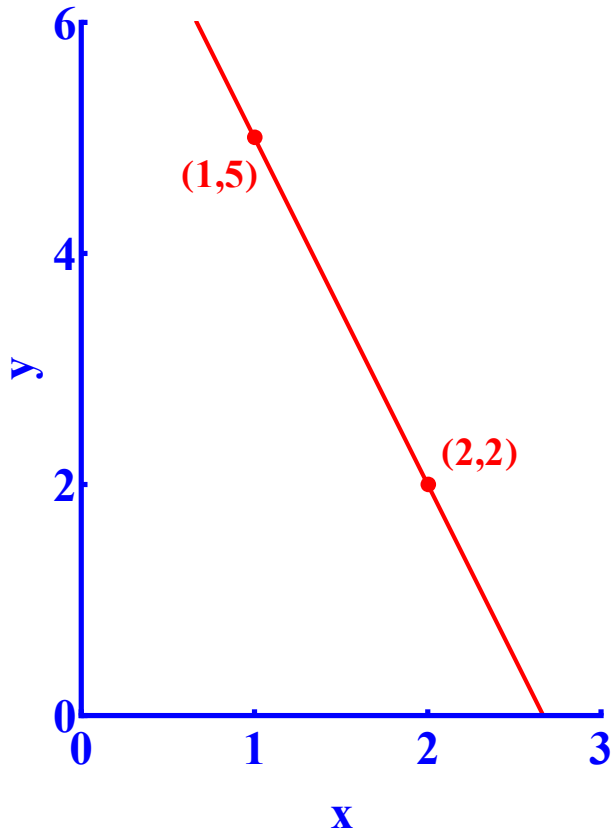
Koeficijent smjera











$$y = 2 \cdot x - 1$$

$$y = -3 \cdot x + 8$$

Jednadžba pravca:

$$y = a \cdot x + b.$$

Interpretacija koeficijenata:

Koeficijent smjera (a) - ukoliko veličinu x povećamo za 1, y će se povećati za a

Slobodni koeficijent (b) - za $x = 0$ je $y = b$.

Regresijska analiza

- primjena metoda kojima se analitički (jednadžbom) objašnjava statistička ovisnost jedne varijable o drugoj ili o više drugih
- iz podataka jedne varijable 'prognoziramo' rezultat druge varijable
- zavisna varijabla - varijabla čiju ovisnost objašnjavamo
- nezavisne varijable - objašnjavaju ponašanje zavisne
- zasniva se na modelu
- model je pojednostavljena slika stvarne pojave
- oblik modela ovisi o primjeru kojeg rješavamo
- ako je odnos između dvije pojave oblikom linearan - model jednostavne linearne regresije
- jedna nezavisna varijabla $\longrightarrow$ jednostavna linearna regresija
- više nezavisnih varijabli $\longrightarrow$ multivarijatna regresija

Dijagram rasipanja

- prvi korak u regresijskoj analizi
- uočiti odnos među pojavama
- pravokutni koordinatni sustav (XY-graf)
- što više vrijednosti (parova) - kvalitetniji zaključak o pojavi

Jednostavna linearna regresija

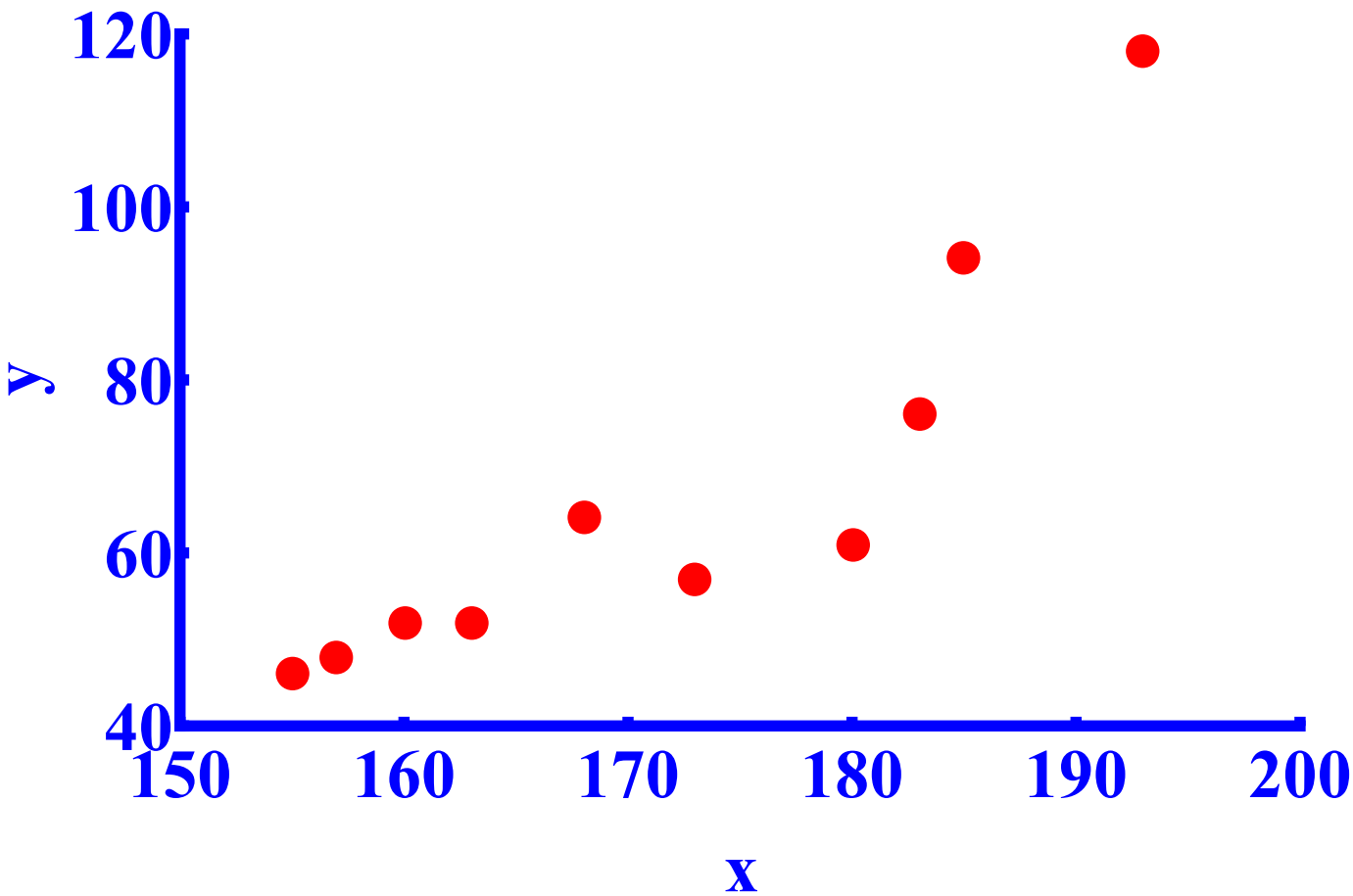
Odrediti oblik linearne veze znači odrediti vezu oblika

$$Y = a \cdot X + b.$$

Odrediti vezu $\longleftrightarrow$ Odrediti koeficijente a i b.

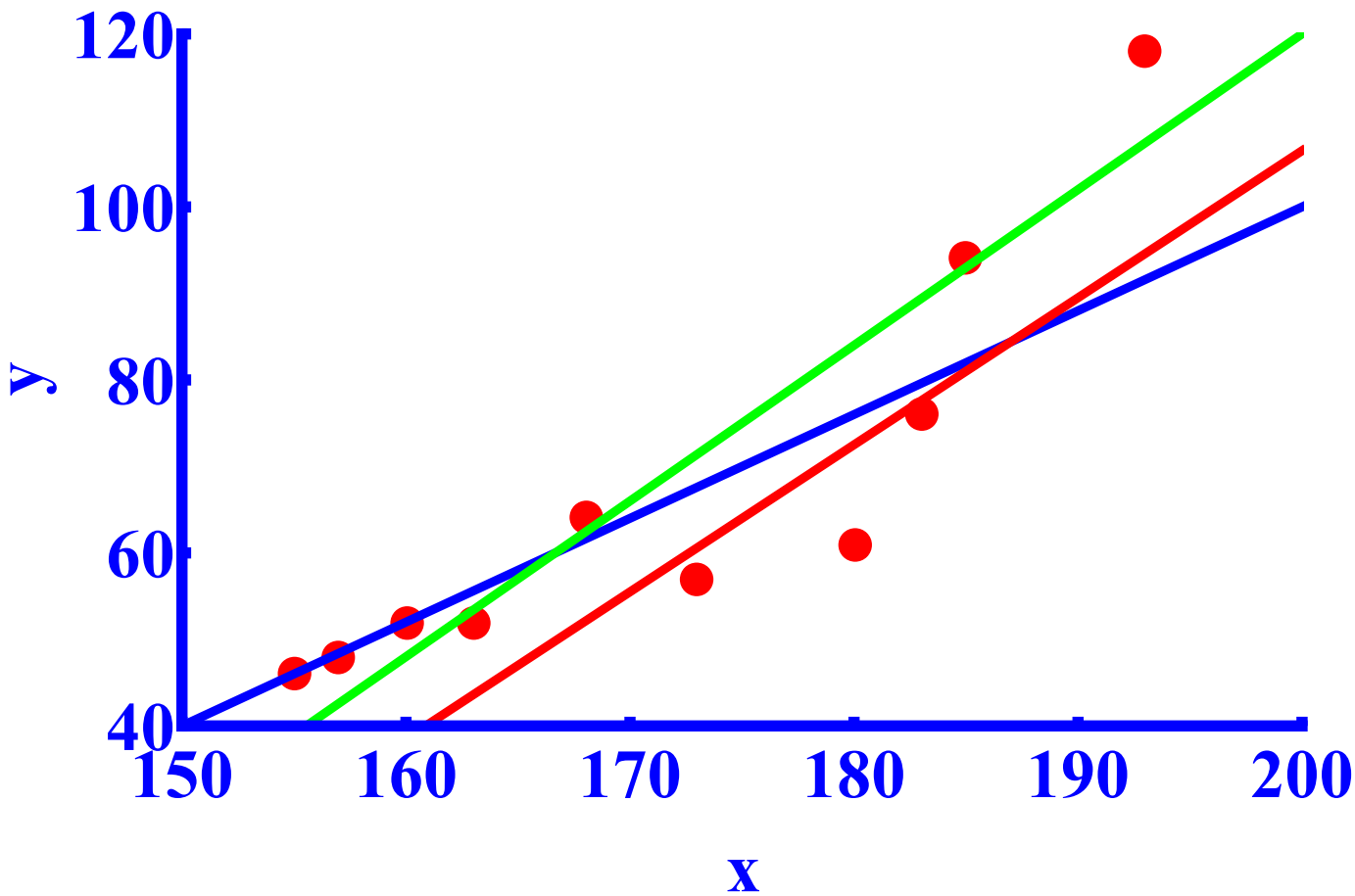
Kako odrediti koeficijente a i b?

Podaci za visinu i težinu:



Kako odrediti pravac koji najbolje opisuje podatke?

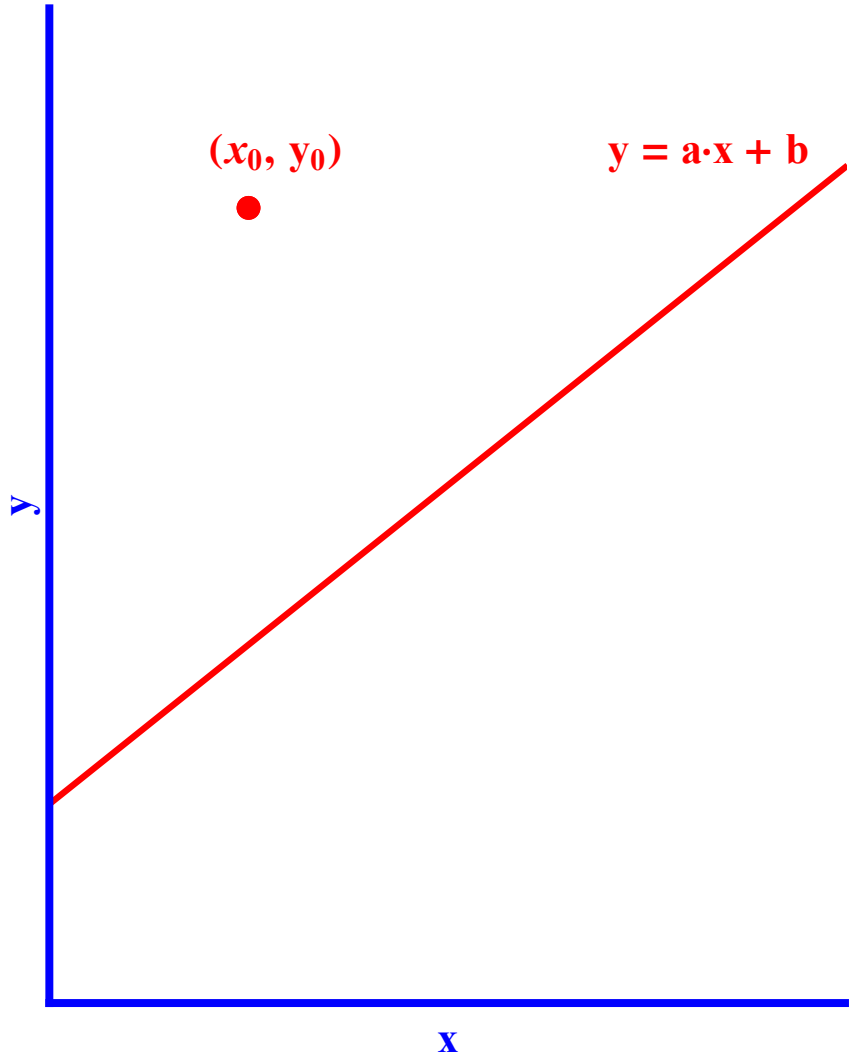
Podaci za visinu i težinu:



Koji pravac bolje opisuje podatke?

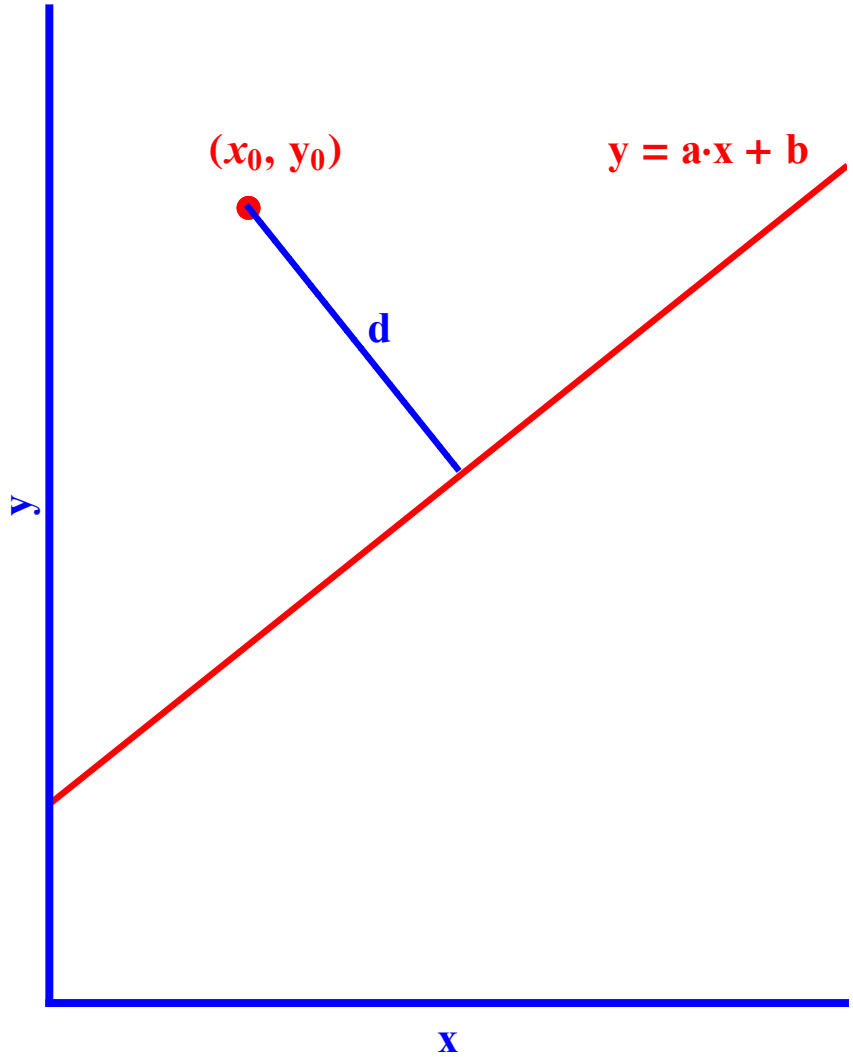
Kvadratno odstupanje

Udaljenost pravca od točke (podatka)



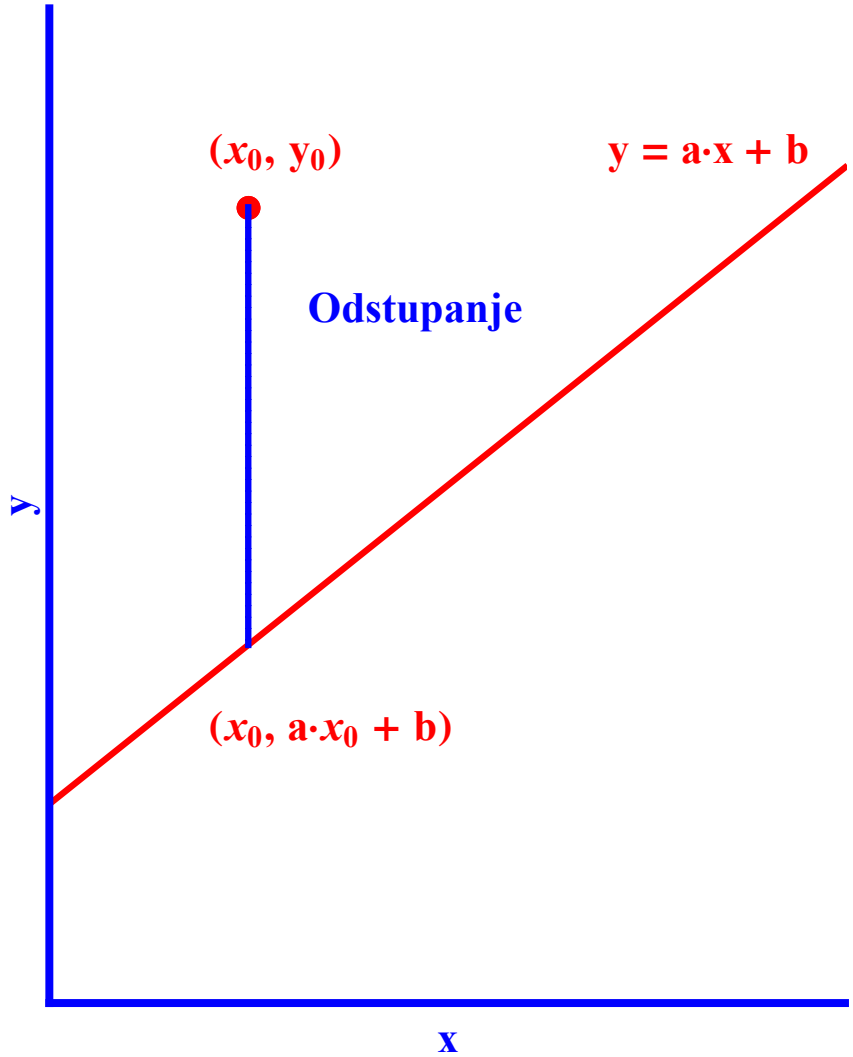
Kvadratno odstupanje

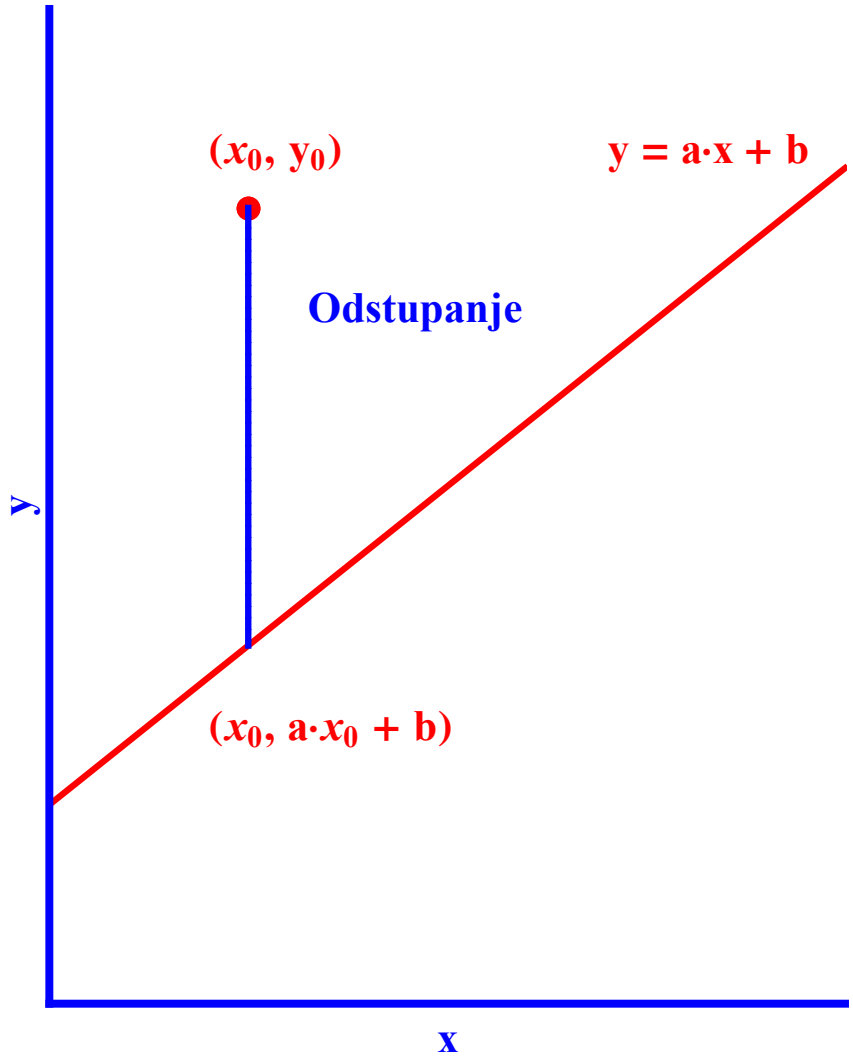
Udaljenost pravca od točke (podatka)



Kvadratno odstupanje

Odstupanje pravca od točke (podatka)





Odstupanje = $a \cdot x_0 + b - y_0$

Odstupanje = $a \cdot x_0 + b - y_0$

Apsolutno odstupanje = $|a \cdot x_0 + b - y_0|$

Kvadratno odstupanje = $(a \cdot x_0 + b - y_0)^2$

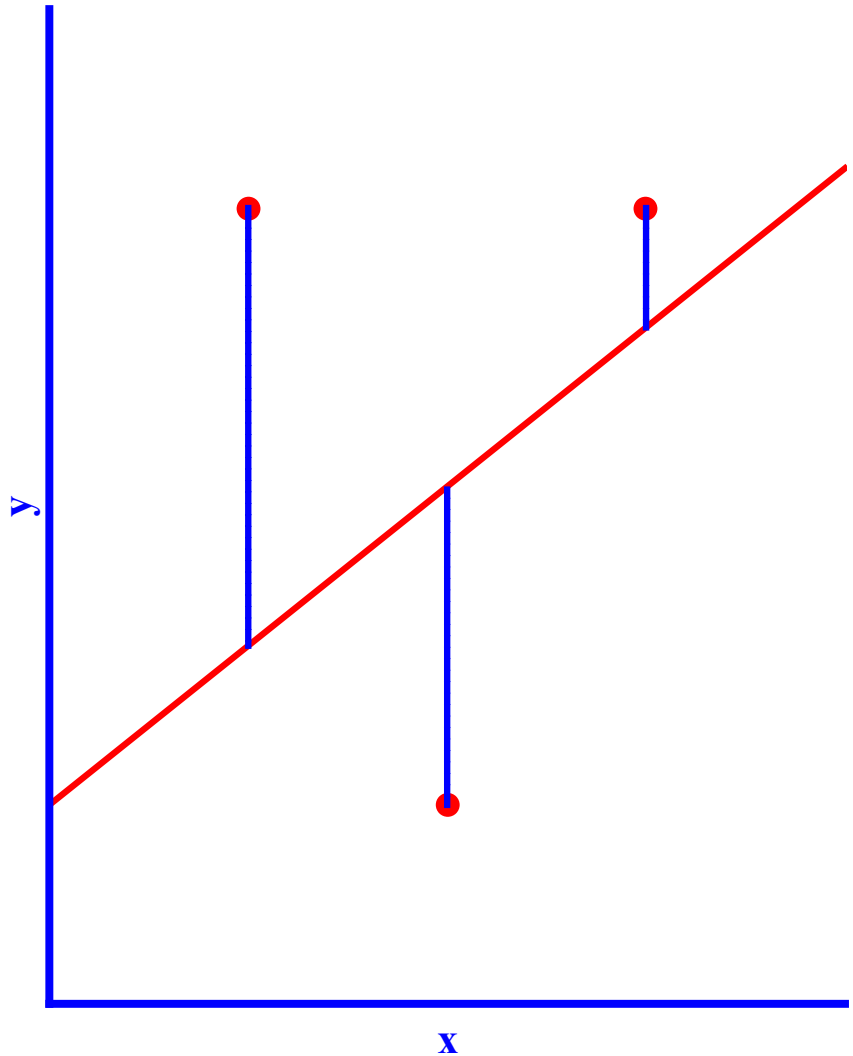
U regresiji se najčešće koristi kvadratno odstupanje.

Kako definirati udaljenost pravca od skupa točaka?

Srednje kvadratno odstupanje: aritmetička sredina kvadratnih odstupanja od pojedine točke.

Srednje kvadratno odstupanje

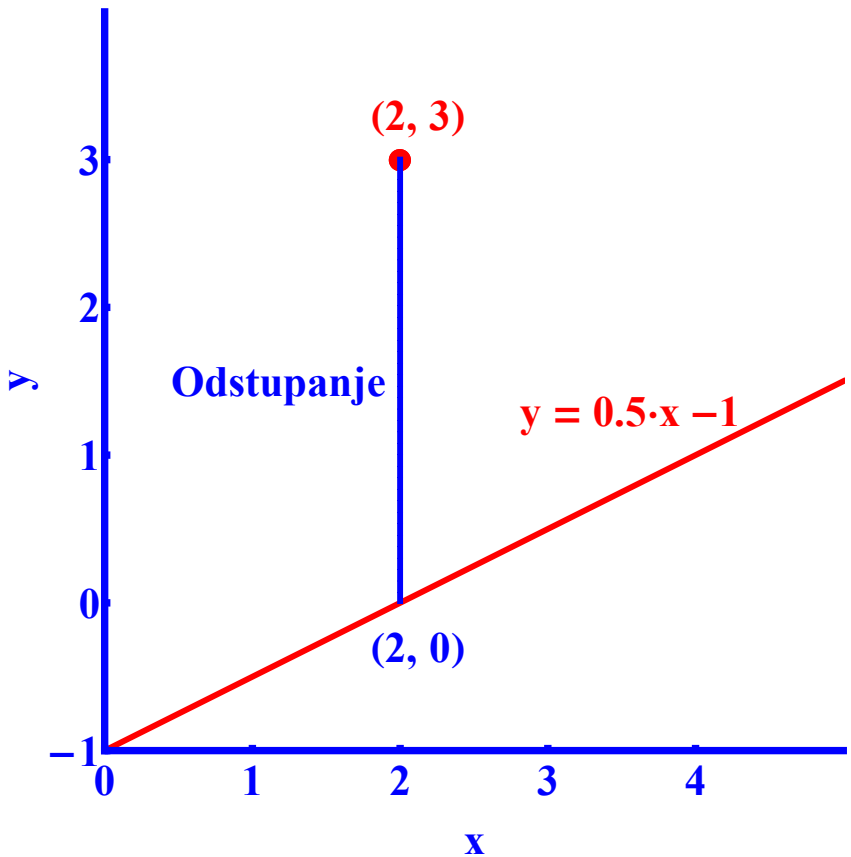
:



Primjer. Izračunajte kvadratno odstupanje točke $(2, 3)$ od pravca $y = 0.5x - 1$.

$(x_0, y_0) = (2, 3)$

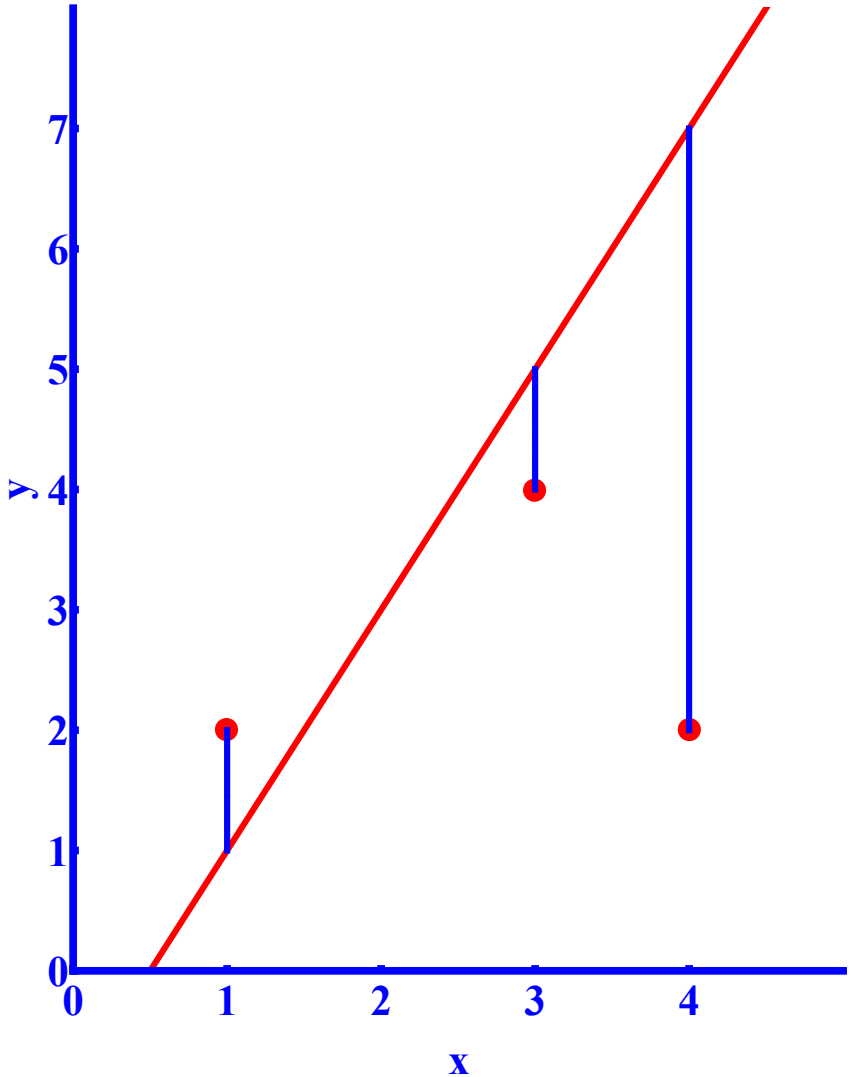
$(a \cdot x_0 + b - y_0)^2 = (0.5x_0 - 1 - y_0)^2 = (0.5 \cdot 2 - 1 - 3)^2 = (-3)^2 = 9$



Primjer. Izračunajte srednje kvadratno odstupanje podataka iz tablice od pravca $y = 2x - 1$.

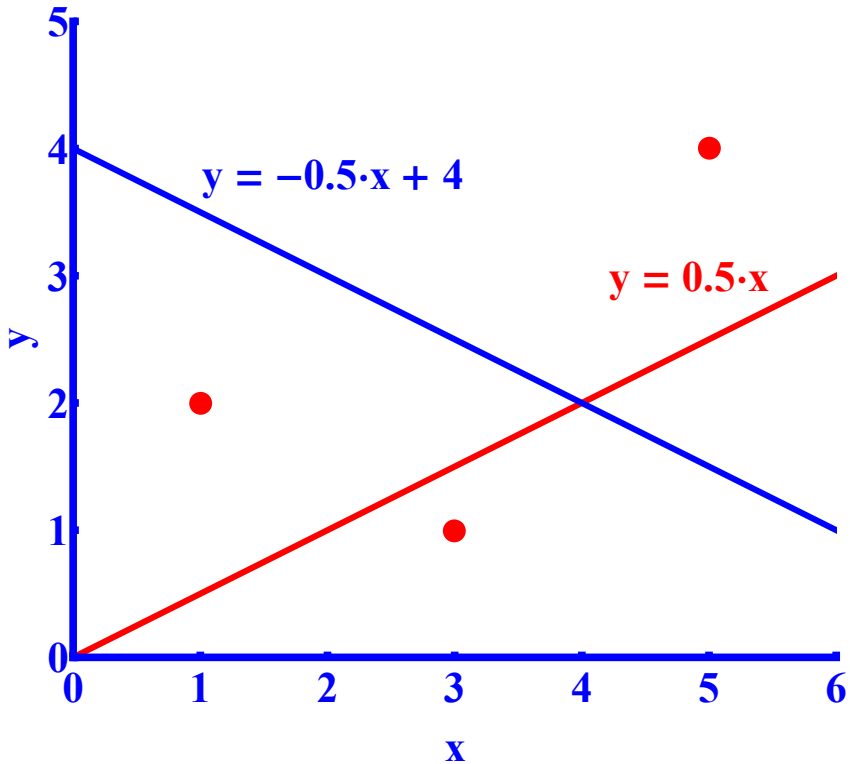
x	y
1	2
4	2
3	4

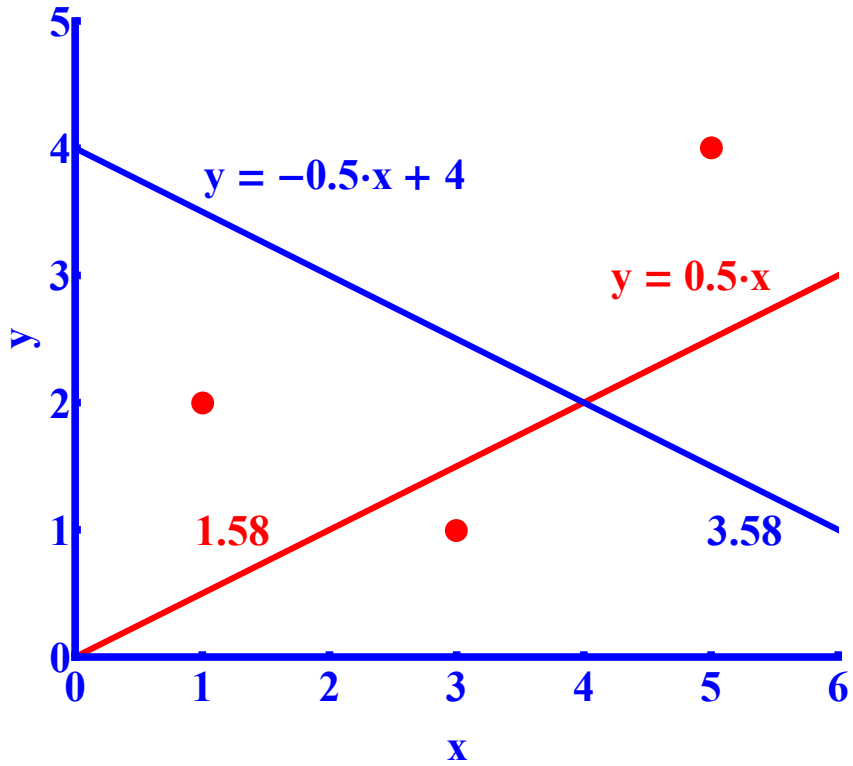
x	y	$2x-1$	$2x-1-y$	$(2x-1-y)^2$
1	2	1	-1	1
4	2	7	5	25
3	4	5	1	1
$\sum$				27
$\sum/n$				9


7
6
5
y 4
3
2
1
0
0
1
2
3
4
x

Primjer. Koji od pravaca $y = 0.5x$ i $y = -0.5x + 4$ bolje opisuje podatke iz tablice?

x	y
1	2
3	1
5	4





Pravac	Srednje kvadratno odstupanje
$y = 0.5x$	1.58
$y = -0.5x + 4$	3.58

Manje srednje kvadratno odstupanje
$\longrightarrow$ **Pravac bolje opisuje podatke.**

Koji pravac najbolje opisuje podatke?

Pravac s **najmanjim** srednjim kvadratnim odstupanjem.

Za podatke $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ tražimo pravac za koji je

$$\frac{1}{n} \sum_i (a \cdot x_i + b - y_i)^2$$

najmanje.

Tražimo a i b za koje je

$$\frac{1}{n} \sum_i (a \cdot x_i + b - y_i)^2$$

najmanje.

Pravac koji minimizira srednje kvadratno odstupanje naziva se **regresijski pravac**.

Koeficijenti regresijskog pravca nazivaju se **regresijski koeficijenti**.

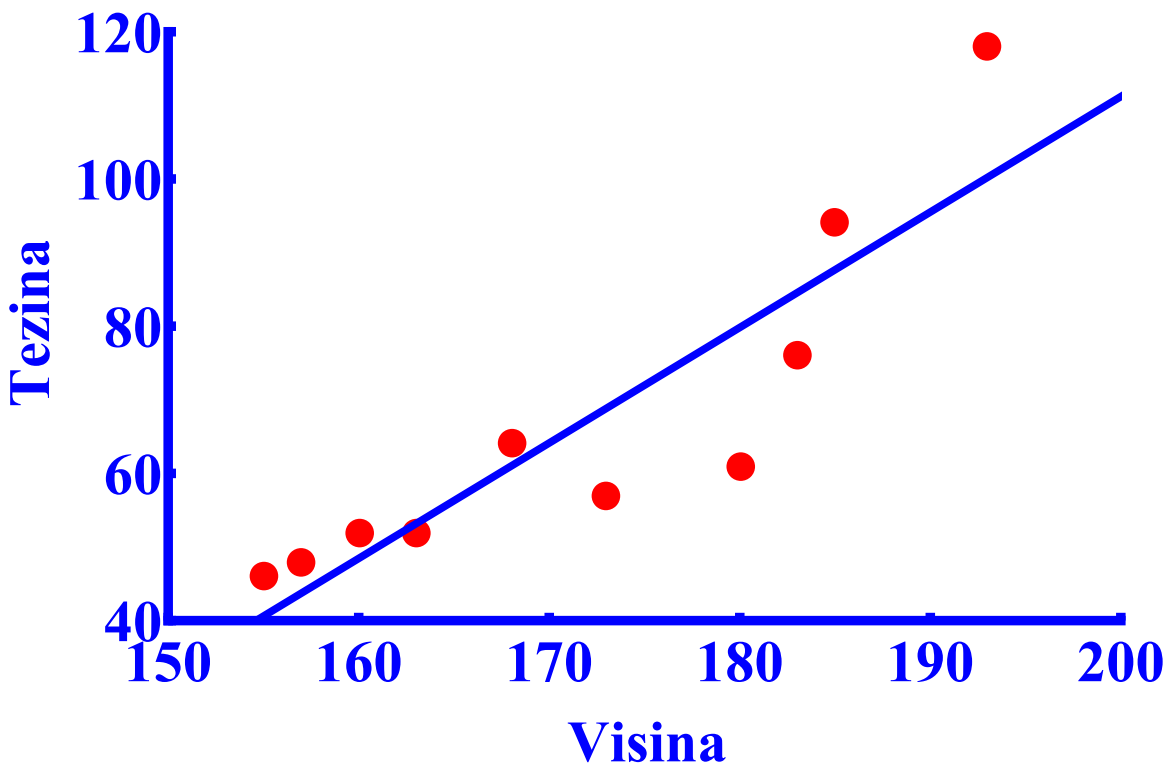
Eksplicitni izraz za regresijske koeficijente:

$$\begin{aligned} a &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \\ &= \frac{Cov(x, Y)}{S_X^2} = \\ &= r_{X,Y} \frac{S_Y}{S_x} \\ b &= \bar{Y} - a \cdot \bar{X} \end{aligned}$$

$r_{X,Y}$ - Pearsonov koeficijent korelacije varijabli X i Y

Primjer. Regresijski pravac za podatke o visini i težini.

$$\text{Težina} = 1.56854 \cdot \text{Visina} - 202.519$$



Standardizirani koeficijenti

Umjesto regresije s varijablama X i Y možemo napraviti regresiju sa standardiziranim varijablama Z_X i Z_Y:

$$Z_Y = \alpha Z_X + \beta.$$

Zbog standardizacije je

$$\bar{Z}_X = \bar{Z}_Y = 0$$

te je slobodni koeficijent

$$\beta = \bar{Z}_Y - a \cdot \bar{Z}_X = 0.$$

Nadalje:

$$\alpha = r_{Z_X,Z_Y} \frac{S_{Z_X}}{S_{Z_Y}} = r_{X,Y}$$

jer su Z_X i Z_Y standardizirane varijable:

$$S_{Z_X} = S_{Z_Y} = 1$$

i

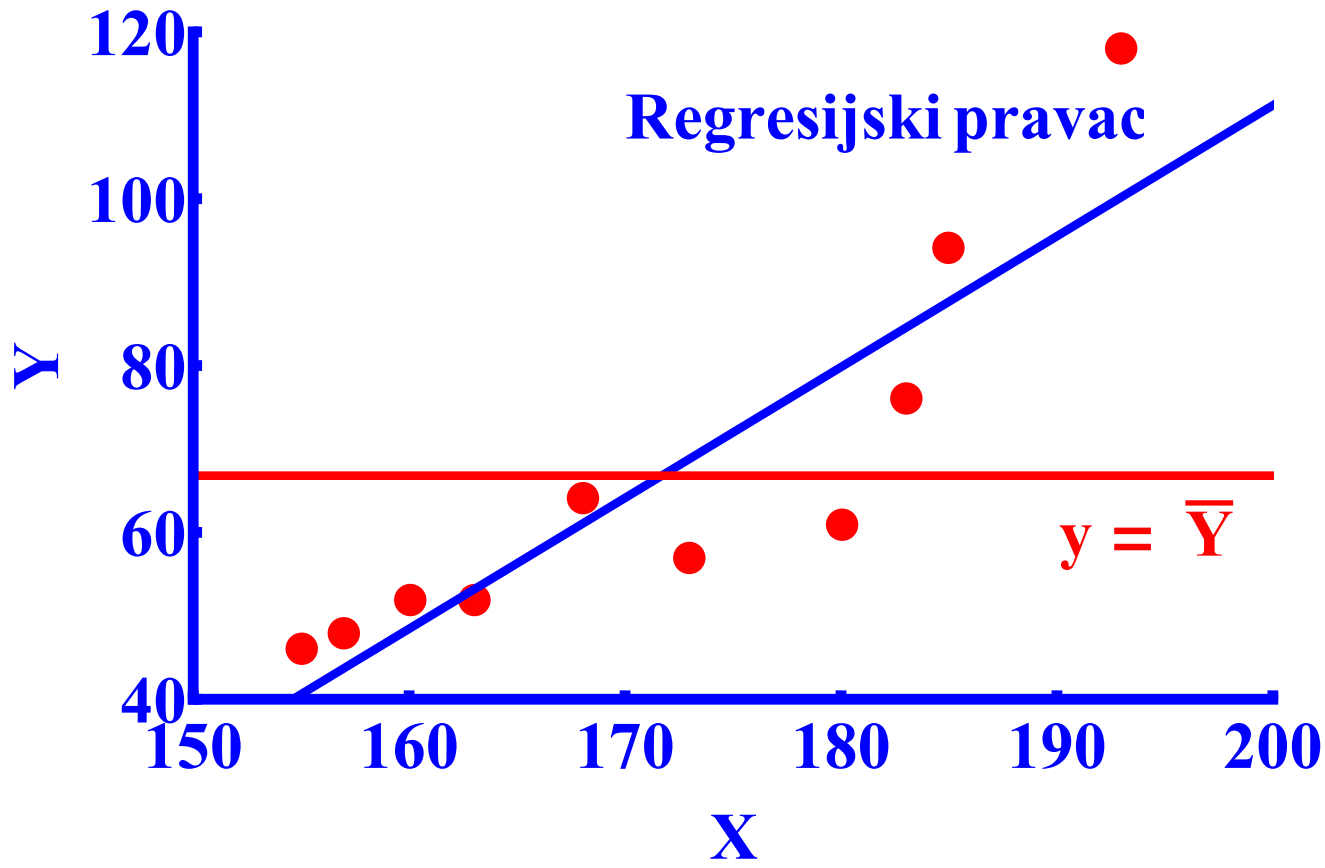
$$r_{Z_X,Z_Y} = r_{X,Y}.$$

Veza između regresijskih koeficijenata i standardiziranih regresijskih koeficijenata:

$$\alpha = a \cdot \frac{S_Y}{S_X}.$$

Koeficijent determinacije

Koliko dobro regresijski pravac opisuje podatke?



Srednje kvadratno odstupanje regresijskog pravca je manje nego za pravac $y = \bar{Y}$.

$$\sum_i (a \cdot x_i + b - y_i)^2 \leq \sum_i (y_i - \bar{Y})^2$$

Desna strana je proporcionalna $Var(Y)$.

Član

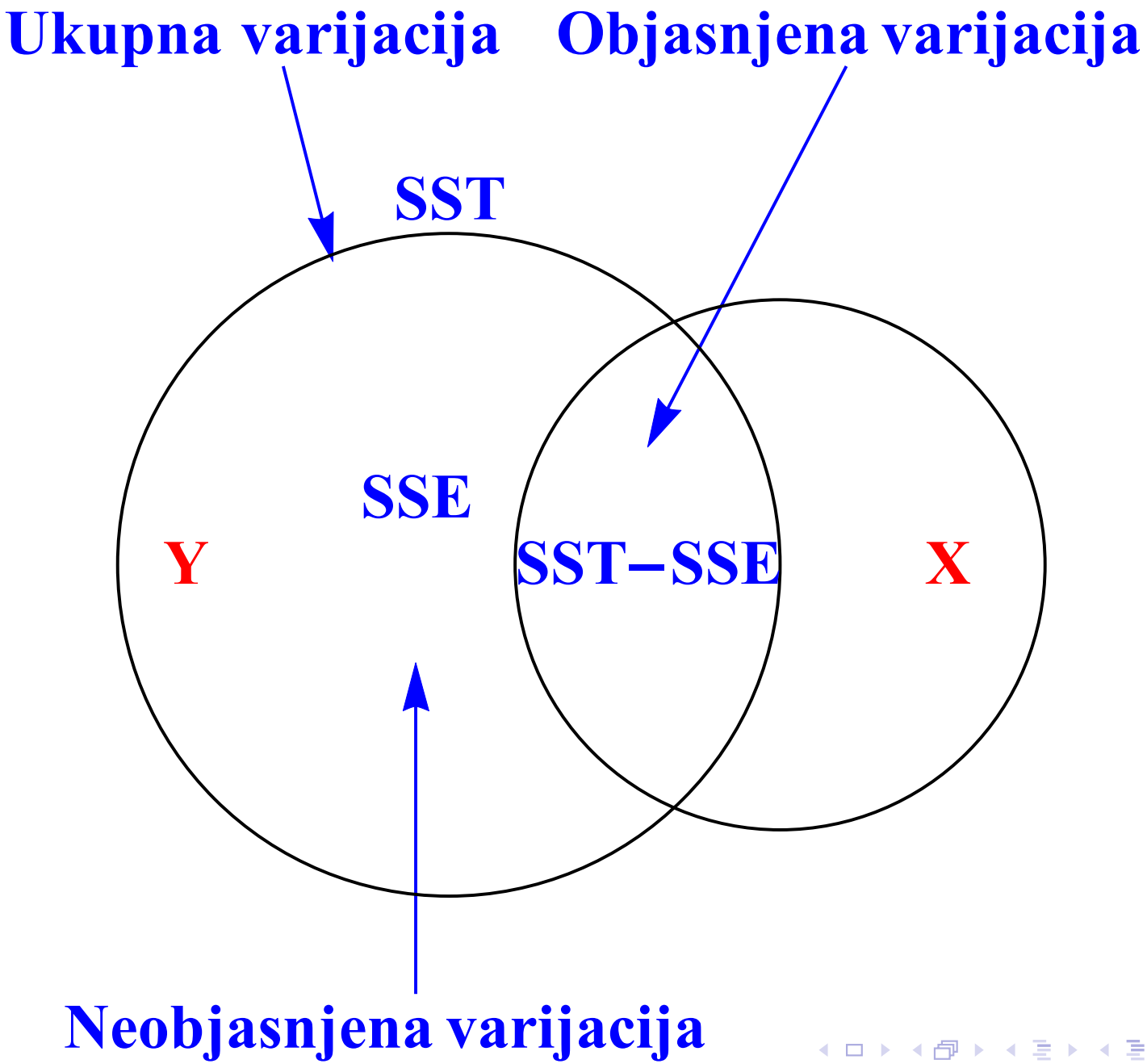
$$\sum_i (a \cdot x_i + b - y_i)^2$$

je **neobjašnjena varijanca** od Y.

Oznake:

SSE $= \sum_i (a \cdot x_i + b - y_i)^2$

SST $= \sum_i (y_i - \bar{Y})^2$


Ukupna varijacija
Objasnjena varijacija
SST
SSE
Y
SST−SSE
X
Neobjasnjena varijacija

Objašnjena varijanca: $\text{SST} - \text{SSE}$

SSE ovisi o mjernim jedinicama.

$0 \leq \text{SSE} \leq \text{SST}$

$\text{SSE} = 0 \quad \longrightarrow \quad$ pravac idealno opisuje podatke

$\text{SSE} = \text{SST} \quad \longrightarrow \quad$ nema utjecaja obilježja X na obilježje Y.

Mjera kvalitete regresije

$$\frac{\text{objašnjena varijanca}}{\text{ukupna varijanca}} = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

Koeficijent determinacije:

$$r^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

r^2 - udio objašnjene varijacije u ukupnoj varijaciji

$r^2 = 1$ - pravac idealno opisuje podatke

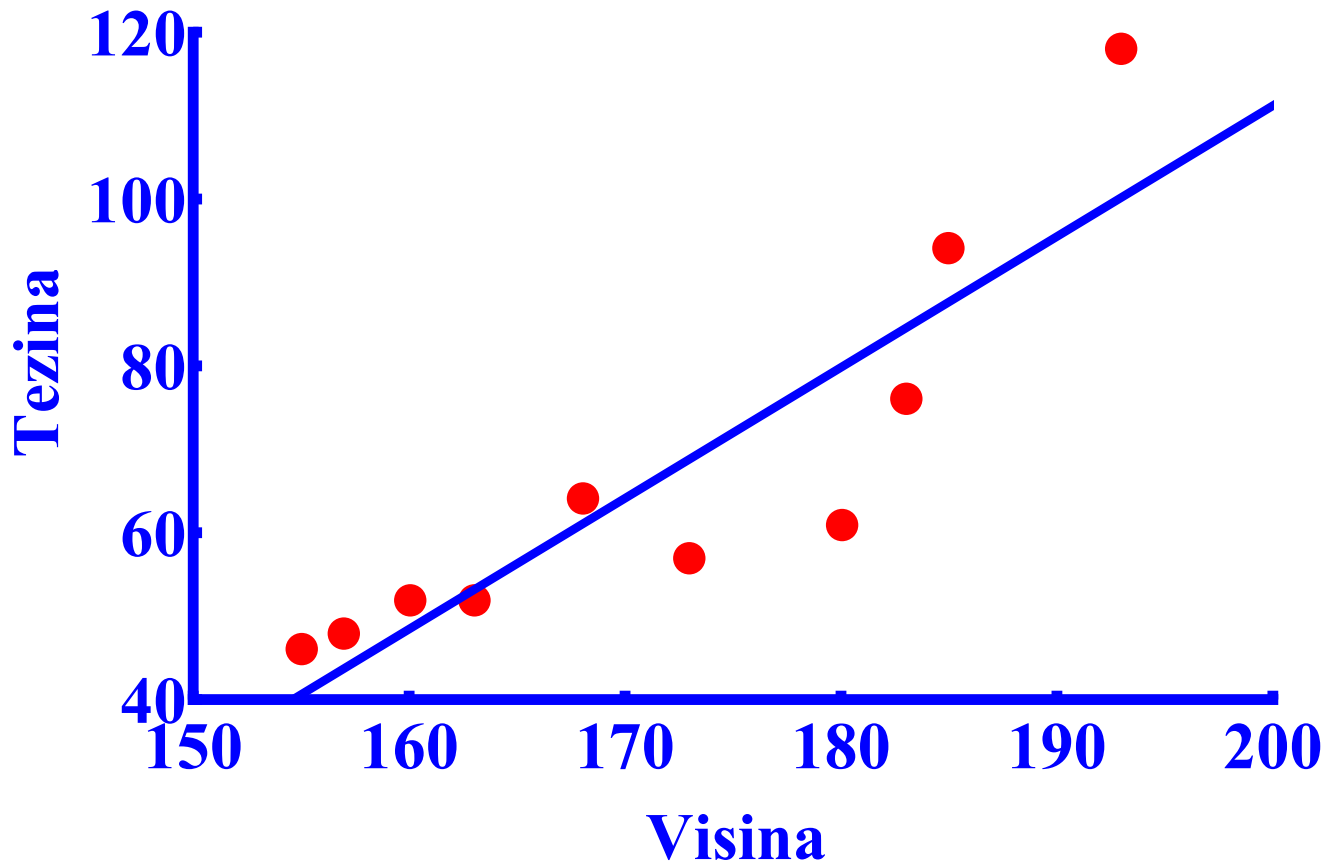
$r^2 = 0$ - nema utjecaja obilježja X na obilježje Y

Veza koeficijenta determinacije i koeficijenta korelacije:

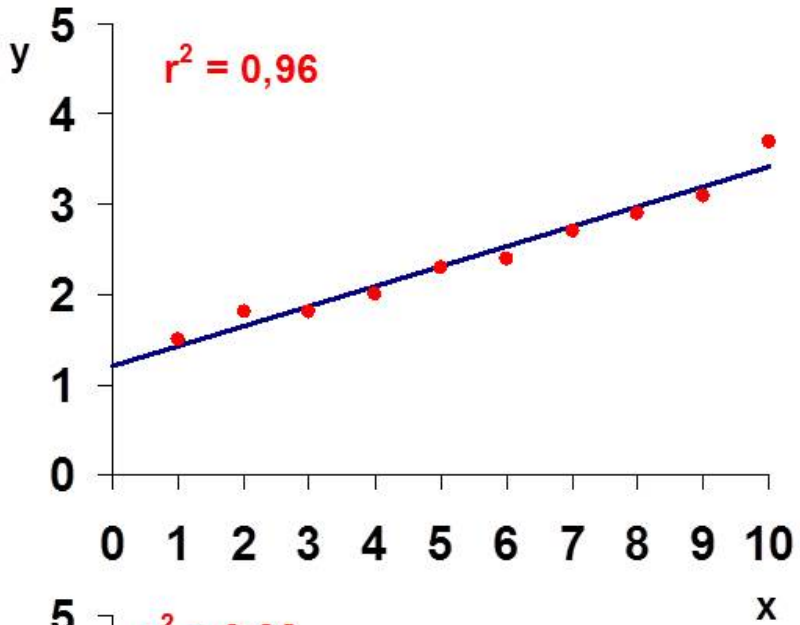
$$r^2 = r_{XY}^2$$

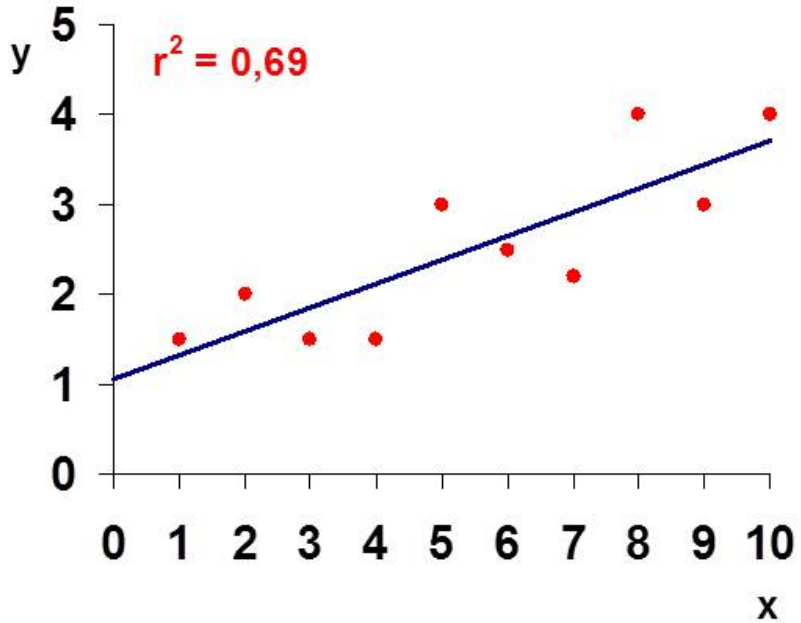
r^2 - koeficijent determinacije

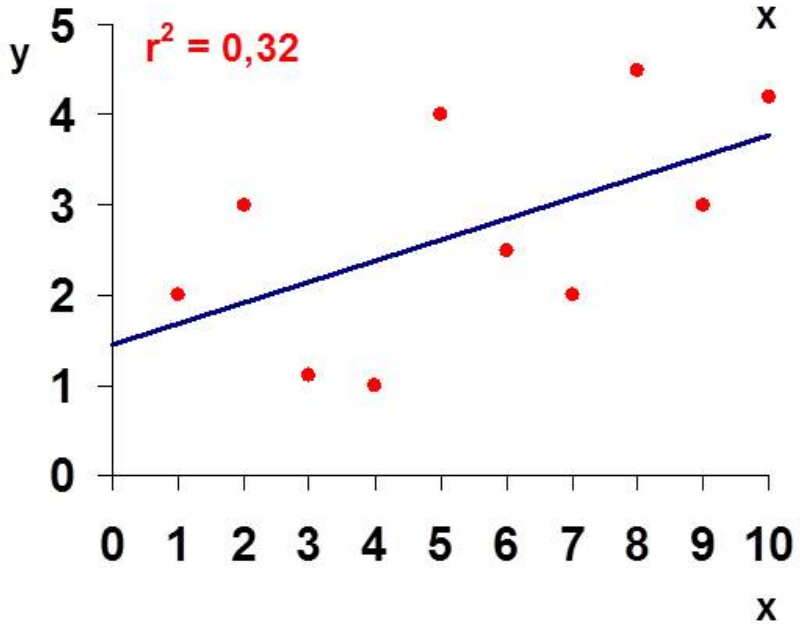
r_{XY} - koeficijent korelacije

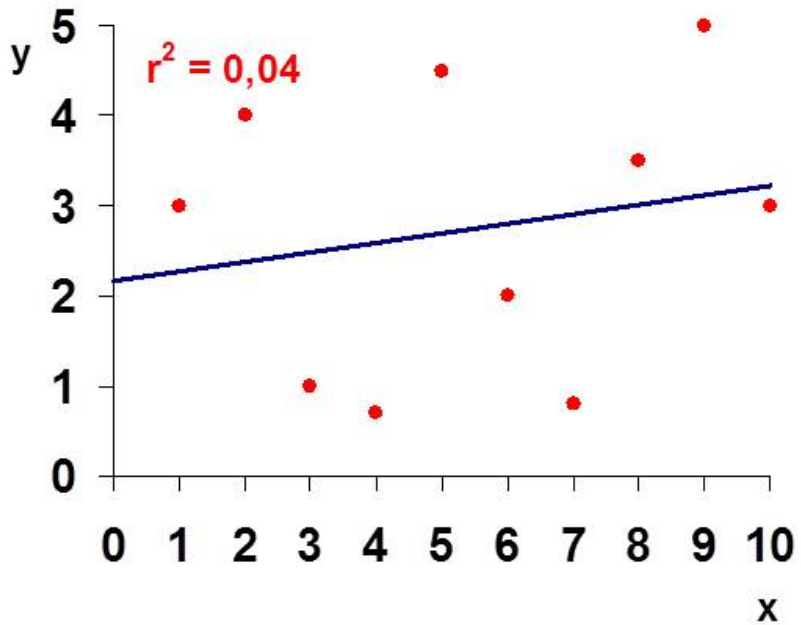


$$r^2 = 0.79$$


y
x
$r^2 = 0,96$
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5


y
x
$r^2 = 0,69$
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5


y
x
$r^2 = 0,32$
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5


y
x
$r^2 = 0,04$
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5

Parcijalna korelacija

Zanima nas korelacija varijabli X i Y ali bez dijela varijacije koja je opisana obilježjem Z.

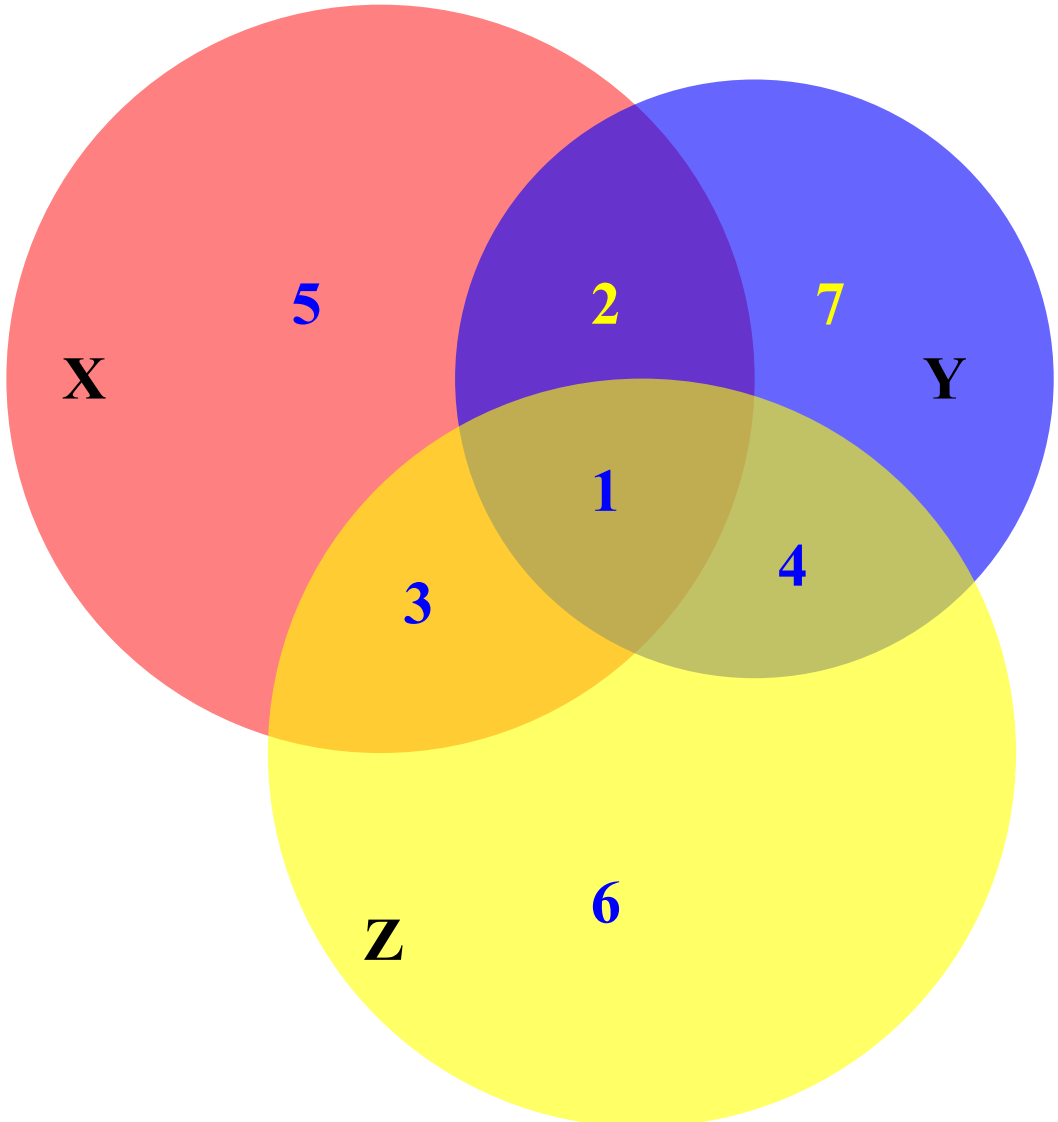
Od varijabli X i Y oduzmemo dio koji opisuje Z (dobiven regresijom, za svaku varijablu posebno):

$$R_{X.Z} = X - a_1 Z - b_1$$

$$R_{Y.Z} = Y - a_2 Z - b_2$$

Parcijalna korelacija od X i Y:

$$r_{XY.Z} = \text{Corr}\left(R_{X.Z}, R_{Y.Z}\right)$$


X
Y
Z
5
2
7
1
4
3
6

Testiranje hipoteza o koeficijentima

Pretpostavka: X i Y imaju **bivarijatnu normalnu** razdiobu.

Regresijski model:

$$Y = a \cdot X + b$$

Možemo testirati hipoteze

$$a = 0 \quad \text{i / ili} \quad b = 0.$$

Znamo:

$$\text{SST} = S_Y^2 = \sum_i (Y_i - \bar{Y})^2 \sim \chi^2(n-1)$$

Vrijedi

$$\text{SSE} = \sum_i (a \cdot X_i + b - Y_i)^2 \sim \chi^2(n-2)$$

Može se pokazati da je

$$\text{SST} - \text{SSE} \sim \chi^2(1)$$

Testiranje hipoteze $a = 0$

Promatramo regresijski pravac za koji je $a = 0$:

$$Y = b$$

Suma kvadratnih odstupanja

$$\sum_i (y_i - b)^2$$

je najmanja za $b = \bar{Y}$.

Suma kvadratnih odstupanja je

$$\sum_i (y_i - \bar{Y})^2 = \text{SST}$$

Statistika:

$$F = \frac{\text{SST} - \text{SSE}}{\text{SSE}} \sim F(1, n-2)$$

Testiranje hipoteze $b = 0$ je analogno jedino promatramo model za koji je $b = 0$:

$$Y = a \cdot X.$$

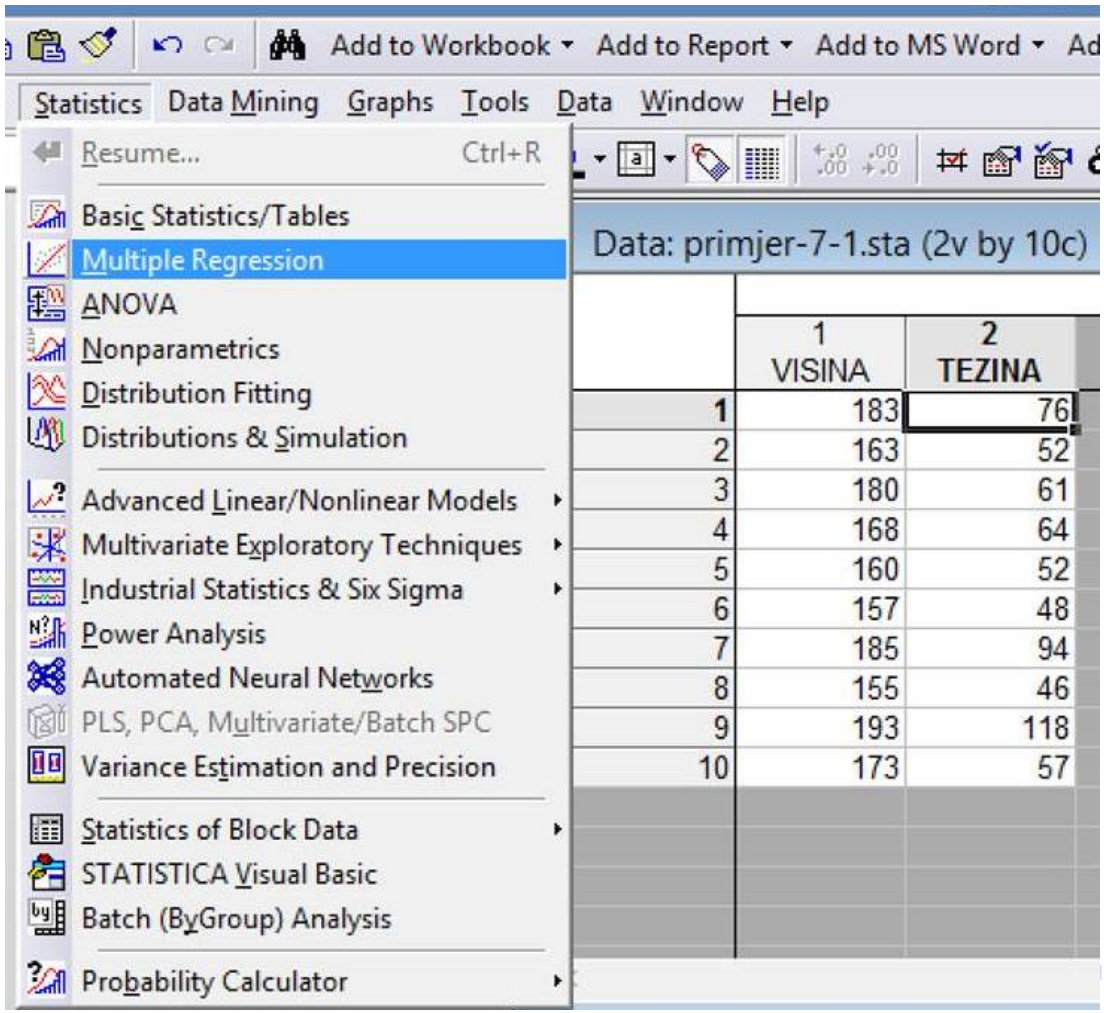
Statistiku dobijemo analogno kao kod testiranja hipoteze $a = 0$.

Drugi pristup je konstrukcija testa na osnovu distribucije regresijskih koeficijenata i uporaba t-statistike.

Statistica

Primjer. Podaci o visini i težini.

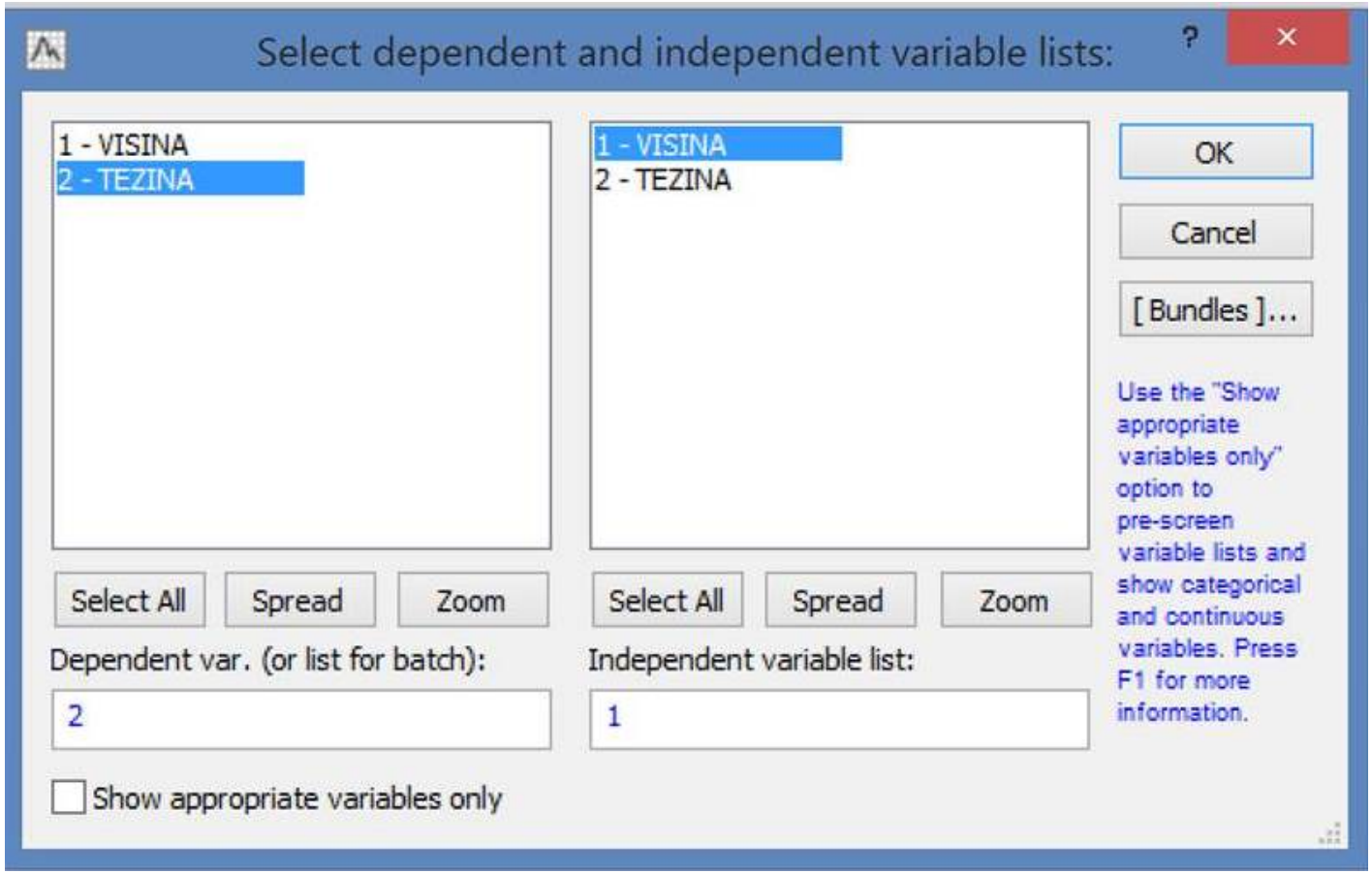
Izbor regresijske analize.

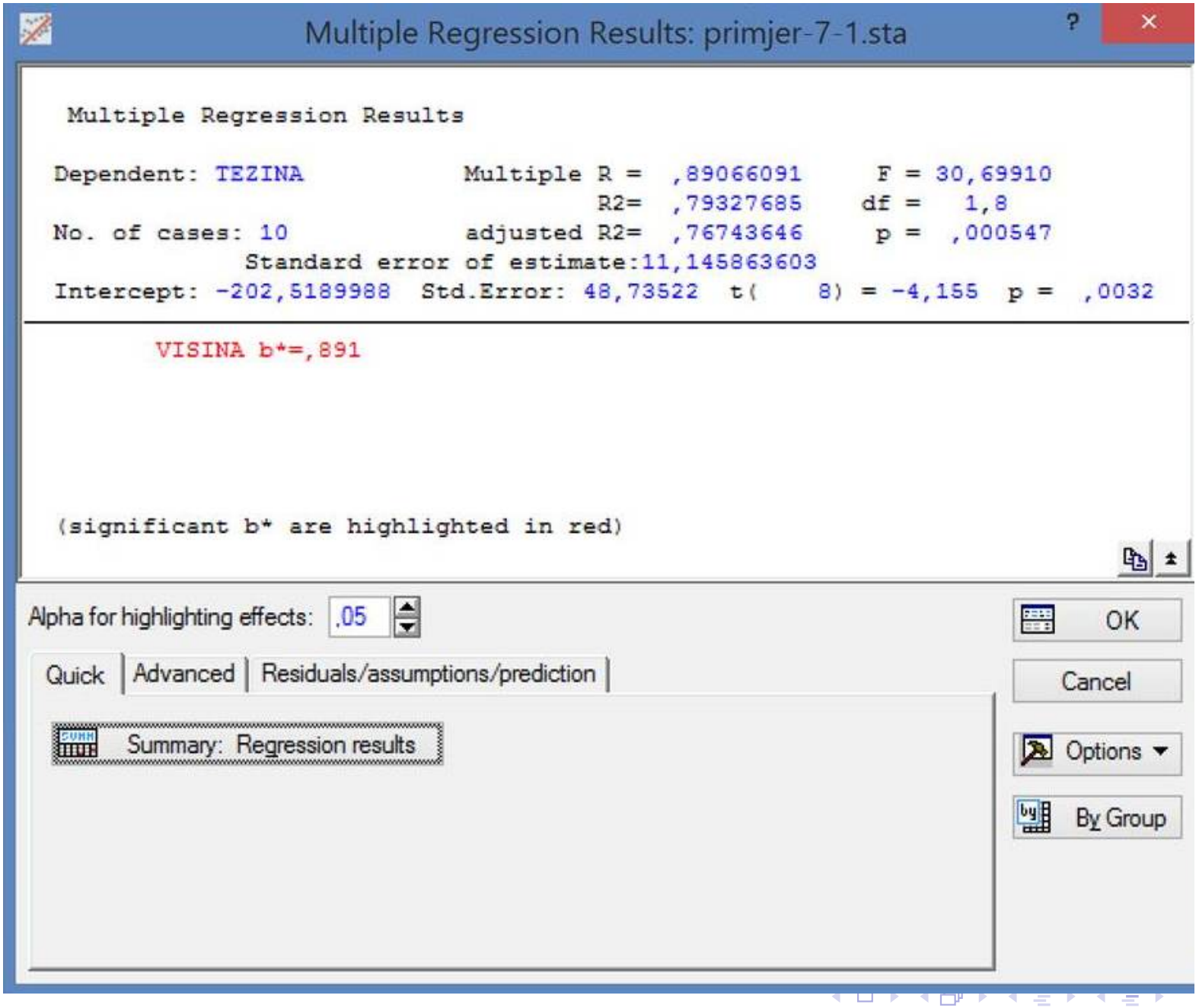


Definiranje varijabli:

Zavisna varijabla - Težina

Nezavisna varijabla - Visina




Multiple Regression Results: primjer-7-1.sta
Multiple Regression Results
Dependent: TEZINA
Multiple R = ,89066091
F = 30,69910
R2= ,79327685
df = 1,8
No. of cases: 10
adjusted R2= ,76743646
p = ,000547
Standard error of estimate:11,145863603
Intercept: -202,5189988 Std.Error: 48,73522 t(8) = -4,155 p = ,0032
VISINA b*=,891
(significant b* are highlighted in red)
Alpha for highlighting effects: .05
OK
Cancel
Options
By Group
Quick
Advanced
Residuals/assumptions/prediction
Summary: Regression results

Regression Summary for Dependent Variable: TEZINA (primjer-7-1.sta)
R= ,89066091 R2= ,79327685 Adjusted R2= ,76743646
F(1,8)=30,699 p<,00055 Std.Error of estimate: 11,146

N=10	b*	Std.Err. of b*	b	Std.Err. of b	t(8)	p-value
Intercept			-202,519	48,73522	-4,15550	0,003185
VISINA	0,890661	0,160749	1,569	0,28310	5,54068	0,000547

Višestruka linearna regresija

Promatra se ovisnost **jedne** varijable o **više** varijabli.

Model:

$$Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + \ldots + a_k \cdot X_k$$

zavisna varijabla: Y

Nezavisne varijable: $X_1, X_2, \ldots, X_k$

Još se koristi naziv **multivarijatna regresija**.

Zavisna varijabla = **varijabla odziva** ('response')

nezavisne varijable = **prediktorske varijable**

Uzorak:

$$(y_1, x_{11}, x_{21}, \ldots, x_{k1})$$
$$(y_2, x_{12}, x_{22}, \ldots, x_{k2})$$
$$\vdots$$
$$(y_n, x_{1n}, x_{2n}, \ldots, x_{kn})$$

Regresijske koeficijente dobijemo minimiziranjem sume kvadratnih odstupanja:

$$\sum_i (a_0 + a_1 \cdot x_{1i} + a_2 \cdot x_{2i} + \ldots + a_k \cdot x_{ki} - y_i)^2$$

Koeficijent determinacije

Isti princip kao u jednostavnoj linearnoj regresiji.

Koeficijent determinacije je udio objašnjene varijance u ukupnoj varijanci.

Ukupna varijanca: **SST** $= \sum_i (y_i - \bar{Y})^2$

Neobjašnjena varijanca od Y:
SSE $= \sum_i (a_0 + a_1 \cdot x_{1i} + a_2 \cdot x_{2i} + \ldots + a_k \cdot x_{ki} - y_i)^2$

Objašnjena varijanca od Y: **SST** $-$ **SSE**

Koeficijent determinacije

$$r^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Značajnost modela

Testiramo hipotezu da nezavisne varijable nisu korelirane s zavisnom.

$\leftrightarrow$ Model opisuje zavisnu varijablu jednako dobro kao model sa slobodnim koeficijentom.

Distribucije suma:

$$\textbf{SST} \sim \chi^2(n-1)$$

$$\textbf{SSE} \sim \chi^2(n-k-1)$$

$$\textbf{SST - SSE} \sim \chi^2(k)$$

Statistika:

$$F = \frac{(\text{SST - SSE})/k}{\text{SSE}/(n-k-1)} \sim F(k, n-k-1)$$

k - broj nezavisnih varijabli

Značajnost regresijskih koeficijenata

Želimo provjeriti da li varijabla X_m značajno doprinosi opisu zavisne varijable Y u modelu

$$Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + \ldots + a_k \cdot X_k$$

Npr., za $m = k$, gornji model uspoređujemo s modelom

$$Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + \ldots + a_{k-1} \cdot X_{k-1}$$

(polazni model bez varijable X_k)

Ukoliko varijabla X_k nije značajna tada oba modela podjednako opisuju Y.

Uspoređujemo sume kvadratnih odstupanja za oba modela.

Ukupna varijanca: **SST** $= \sum_i (y_i - \bar{Y})^2$

Neobjašnjena varijanca od Y (puni model):
SSE $= \sum_i (a_0 + a_1 \cdot x_{1i} + a_2 \cdot x_{2i} + \ldots + a_k \cdot x_{ki} - y_i)^2$

Neobjašnjena varijanca od Y (model bez X_k):
SSE$_k = \sum_i (b_0 + b_1 \cdot x_{1i} + a_2 \cdot x_{2i} + \ldots + b_{k-1} \cdot x_{(k-1)i} - y_i)^2$

Distribucija:

$$\text{SSE}_k \sim \chi^2(n-k) \quad \text{i} \quad \text{SSE} - \text{SSE}_k \sim \chi^2(1)$$

Statistika:

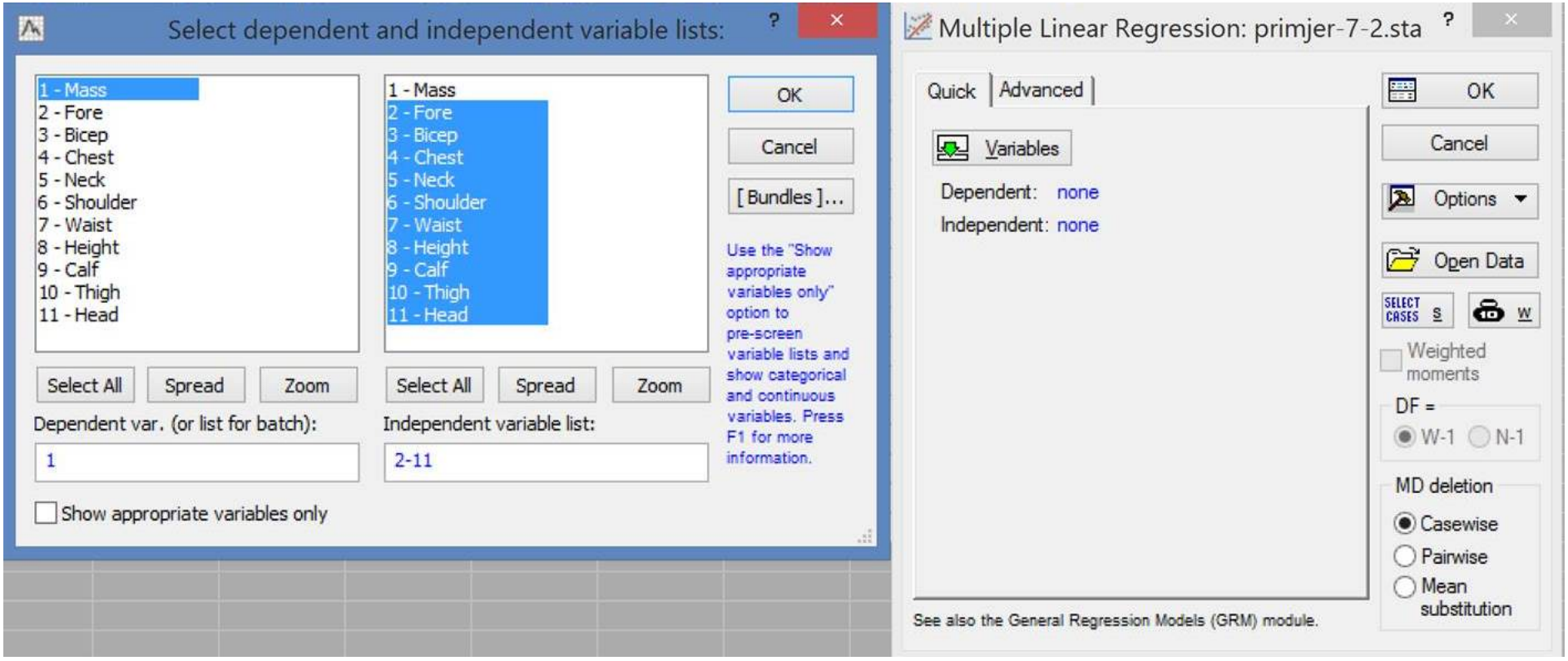
$$F = \frac{\text{SSE}_k - \text{SSE}}{\text{SSE}/(n-k-1)} \sim F(k, n-k-1)$$

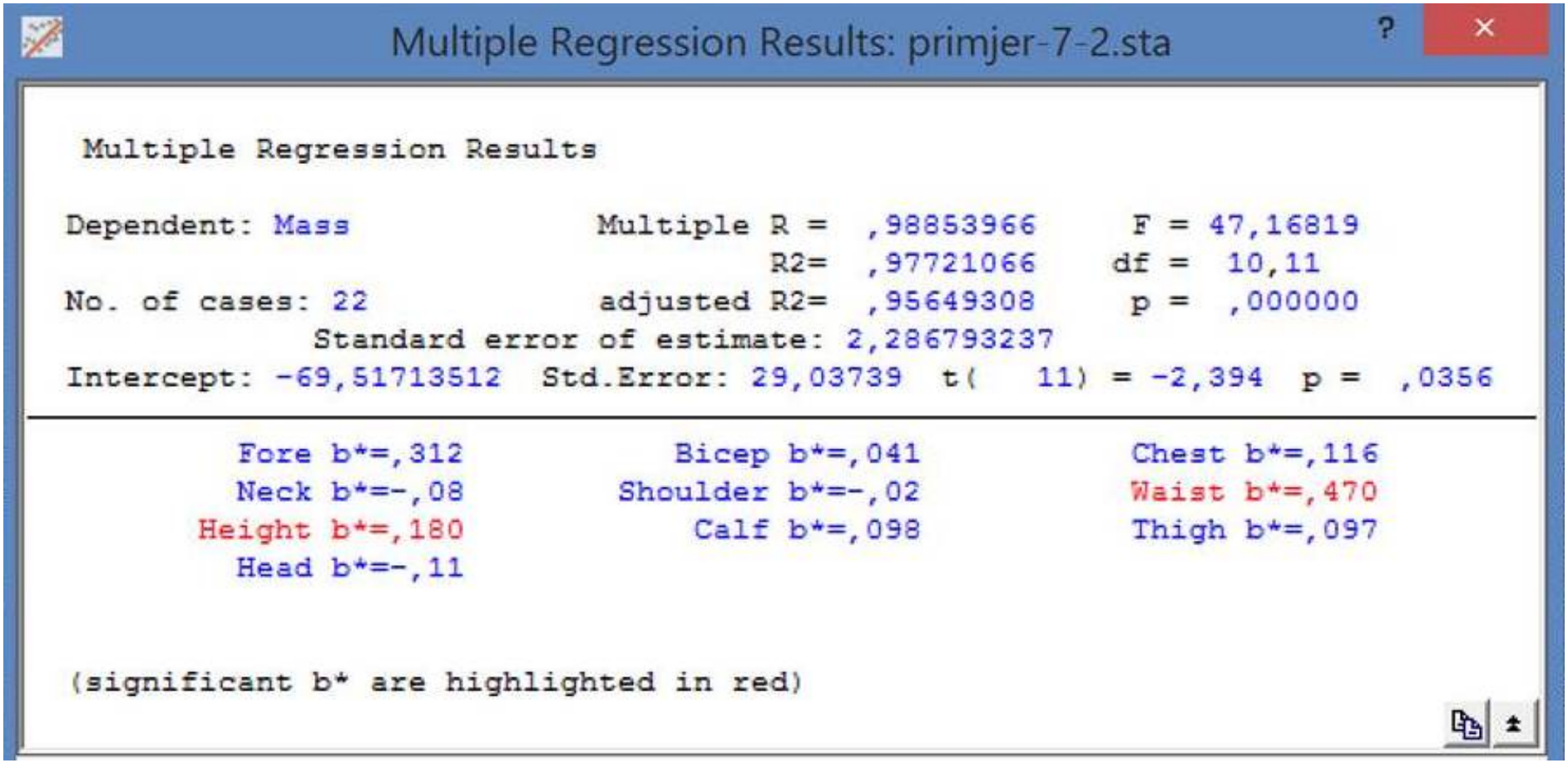
Statistica. Primjer.

Na 22 slučajno izabranih muških osoba starih između 16 i 30 godina izmjereno je:

1. mass - masa osobe u kg
2. fore - maksimalni opseg podlaktice
3. bicep - maksimalni opseg bicepsa
4. chest - opseg grudi
5. neck - opseg vrata
6. waist - opseg struka
7. thigh - opseg bedra
8. calf - maksimalni opseg potkoljenice
9. height - visina
10. shoulders - opseg ramena
11. head - opseg glave

Može li se masa osobe procijeniti na osnovu izmjerenih veličina?


Select dependent and independent variable lists:
1 - Mass
2 - Fore
3 - Bicep
4 - Chest
5 - Neck
6 - Shoulder
7 - Waist
8 - Height
9 - Calf
10 - Thigh
11 - Head
OK
Cancel
[Bundles]...
Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.
Select All
Spread
Zoom
Dependent var. (or list for batch):
1
Independent variable list:
2-11
Show appropriate variables only
Multiple Linear Regression: primjer-7-2.sta
Quick
Advanced
Variables
Dependent: none
Independent: none
OK
Cancel
Options
Open Data
Weighted moments
DF =
W-1
N-1
MD deletion
Casewise
Pairwise
Mean substitution
See also the General Regression Models (GRM) module.



$r^2 = 0.97721066$ - varijable dobro opisuju masu

$p = 0.000000$ - model je značajan

Height, waist su značajne

Jesu li druge varijable značajne u opisu mase?

ANOVA tablica za regresiju:

Effect	Analysis of Variance; DV: Mass (primjer-7-2.sta)				
	Sums of Squares	df	**Mean Squares**	F	p-value
Regress.	2466,624	10	246,6624	47,16819	0,000000
Residual	57,524	11	5,2294		
Total	2524,148				

Regression Summary for Dependent Variable: Mass (primjer-7-2.sta)
R= ,98853966 R2= ,97721066 Adjusted R2= ,95649308
F(10,11)=47,168 p<,00000 Std.Error of estimate: 2,2868

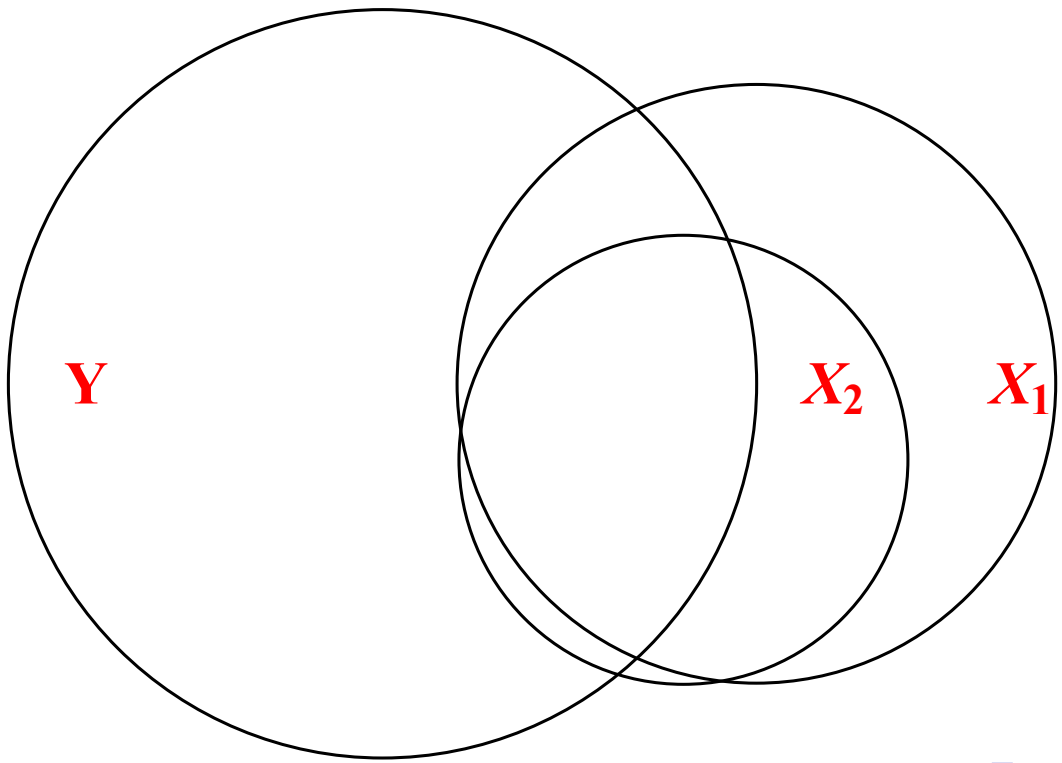
N=22	b*	Std.Err. of b*	b	Std.Err. of b	t(11)	p-value
Intercept			-69,5171	29,03739	-2,39406	0,035605
Fore	0,311943	0,149637	1,7818	0,85473	2,08466	0,061204
Bicep	0,041107	0,128629	0,1551	0,48530	0,31958	0,755275
Chest	0,116160	0,138697	0,1891	0,22583	0,83751	0,420132
Neck	-0,080708	0,120712	-0,4818	0,72067	-0,66860	0,517537
Shoulder	-0,017068	0,139410	-0,0293	0,23943	-0,12243	0,904769
Waist	0,470310	0,082822	0,6614	0,11648	5,67854	0,000143
Height	0,179747	0,073725	0,3178	0,13037	2,43807	0,032935
Calf	0,098306	0,090947	0,4459	0,41251	1,08091	0,302865
Thigh	0,097451	0,100037	0,2972	0,30510	0,97415	0,350917
Head	-0,105378	0,059600	-0,9196	0,52009	-1,76808	0,104735

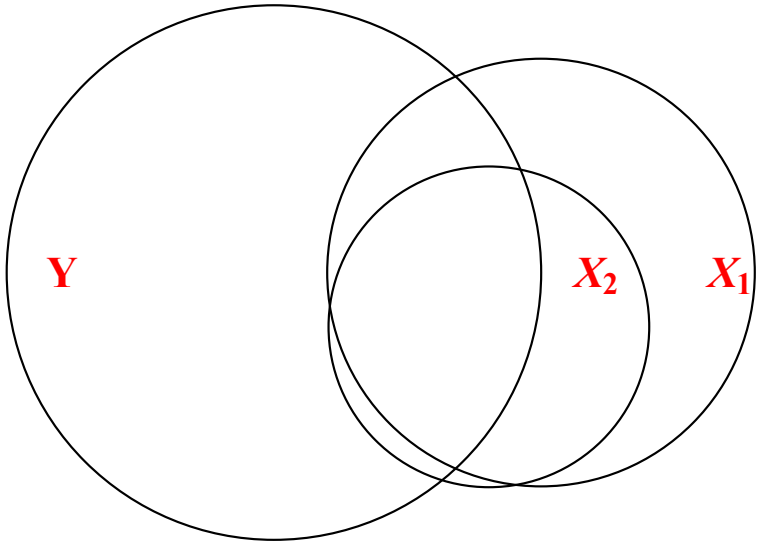
Izgradnja modela u višestrukoj regresiji

Samo su dvije varijable značajne.

Znači li to da druge varijable ne sudjeluju značajno u opisu zavisne varijable Masa?

Prediktorske varijable mogu biti korelirane.





Modeli

$$Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2$$

i

$$Y = a_0 + a_1 \cdot X_1$$

jednako dobro opisuju Y.

U prvom modelu X_1 i X_2 nisu značajne.

Međutim, X_1 je značajna u drugom modelu.

Primjer. Analiziramo tri varijable:

Hcm - visina u centimetrima

Hm - visina u metrima

Hinch - visina u inchima

Promatramo model

$$\text{Hcm} = a_0 + a_1 \cdot \text{Hm} + a_2 \cdot \text{Hinch}$$

Jer je

$$\text{Hcm} = 100 \cdot \text{Hm} \quad \text{i} \quad \text{Hcm} = 2.54 \cdot \text{Hinch}$$

modeli

$$\text{Hcm} = a_0 + a_1 \cdot \text{Hm}$$

i

$$\text{Hcm} = a_0 + a_1 \cdot \text{Hinch}$$

jednako dobro opisuju Hcm ($r^2 = 1$ za sva tri modela).

Varijable Hm i Hinch će biti nesignifikantne u modelu

$$\text{Hcm} = a_0 + a_1 \cdot \text{Hm} + a_2 \cdot \text{Hinch}$$

iako je svaka od njih jako povezana s zavisnom varijablom Hcm.

Prediktorska varijabla može biti nesignifikantna jer je linearno zavisna s jednom ili više drugih prediktorskih varijabli.

Interpretacija koeficijenata.

Regression Summary for Dependent Variable: Mass (primjer-7-2.sta)
R= ,98853966 R2= ,97721066 Adjusted R2= ,95649308
F(10,11)=47,168 p<,00000 Std.Error of estimate: 2,2868

N=22	b*	Std.Err. of b*	b	Std.Err. of b	t(11)	p-value
Intercept			-69,5171	29,03739	-2,39406	0,035605
Fore	0,311943	0,149637	1,7818	0,85473	2,08466	0,061204
Bicep	0,041107	0,128629	0,1551	0,48530	0,31958	0,755275
Chest	0,116160	0,138697	0,1891	0,22583	0,83751	0,420132
Neck	-0,080708	0,120712	-0,4818	0,72067	-0,66860	0,517537
Shoulder	-0,017068	0,139410	-0,0293	0,23943	-0,12243	0,904769
Waist	0,470310	0,082822	0,6614	0,11648	5,67854	0,000143
Height	0,179747	0,073725	0,3178	0,13037	2,43807	0,032935
Calf	0,098306	0,090947	0,4459	0,41251	1,08091	0,302865
Thigh	0,097451	0,100037	0,2972	0,30510	0,97415	0,350917
Head	-0,105378	0,059600	-0,9196	0,52009	-1,76808	0,104735

Širina leđa i ramena te opseg glave negativno utječu na masu!

(Koeficijenti su negativni.)

Treba odrediti model u kojem su sve varijable značajne.

Strategija: Izbacujemo jednu po jednu varijablu iz modela.

Izbacujemo varijablu koja najmanje doprinosi objašnjenoj varijanci.

$\longrightarrow$ Izbacujemo varijablu s najvećom p-vrijednosti za regresijski koeficijent.

Ovaj postupak se naziva **eliminacija unatrag** ('backward elimination').

Izbacivanje prekinemo kada su sve varijable značajne.

Često se za izbacivanje koristi veća razina značajnosti od standardnih $\alpha = 0.05$ (npr. 0.10).

1.korak

Regression Summary for Dependent Variable: Mass (primjer-7-2.sta)
R= ,98853966 R2= ,97721066 Adjusted R2= ,95649308
F(10,11)=47,168 p<,00000 Std.Error of estimate: 2,2868

N=22	b*	Std.Err. of b*	b	Std.Err. of b	t(11)	p-value
Intercept			-69,5171	29,03739	-2,39406	0,035605
Fore	0,311943	0,149637	1,7818	0,85473	2,08466	0,061204
Bicep	0,041107	0,128629	0,1551	0,48530	0,31958	0,755275
Chest	0,116160	0,138697	0,1891	0,22583	0,83751	0,420132
Neck	-0,080708	0,120712	-0,4818	0,72067	-0,66860	0,517537
Shoulder	-0,017068	0,139410	-0,0293	0,23943	-0,12243	0,904769
Waist	0,470310	0,082822	0,6614	0,11648	5,67854	0,000143
Height	0,179747	0,073725	0,3178	0,13037	2,43807	0,032935
Calf	0,098306	0,090947	0,4459	0,41251	1,08091	0,302865
Thigh	0,097451	0,100037	0,2972	0,30510	0,97415	0,350917
Head	-0,105378	0,059600	-0,9196	0,52009	-1,76808	0,104735

Izbacujemo varijablu Shoulder.

(Regresiju ponovimo sa preostalih 9 varijabli.)

2.korak

Regression Summary for Dependent Variable: Mass (primjer-7-2.sta)
R= ,98852395 R2= ,97717961 Adjusted R2= ,96006432
F(9,12)=57,094 p<,00000 Std.Error of estimate: 2,1909

N=22	b*	Std.Err. of b*	b	Std.Err. of b	t(12)	p-value
Intercept			-70,5386	26,64703	-2,64715	0,021291
Fore	0,300753	0,113512	1,7179	0,64838	2,64952	0,021198
Bicep	0,042819	0,122507	0,1615	0,46220	0,34953	0,732752
Chest	0,106164	0,107421	0,1729	0,17491	0,98830	0,342515
Neck	-0,081168	0,115596	-0,4846	0,69012	-0,70217	0,495968
Waist	0,468188	0,077594	0,6585	0,10913	6,03385	0,000059
Height	0,175775	0,063430	0,3108	0,11216	2,77118	0,016925
Calf	0,099848	0,086295	0,4529	0,39141	1,15706	0,269761
Thigh	0,102383	0,087727	0,3123	0,26756	1,16706	0,265855
Head	-0,102361	0,051992	-0,8932	0,45370	-1,96878	0,072513

Izbacujemo varijablu Bicep.

3.korak

Regression Summary for Dependent Variable: Mass (primjer-7-2.sta)
R= ,98840644 R2= ,97694728 Adjusted R2= ,96276100
F(8,13)=68,866 p<,00000 Std.Error of estimate: 2,1157

N=22	b*	Std.Err. of b*	b	Std.Err. of b	t(13)	p-value
Intercept			-71,9503	25,43433	-2,82886	0,014222
Fore	0,314561	0,102760	1,7968	0,58696	3,06114	0,009103
Chest	0,118421	0,098048	0,1928	0,15965	1,20779	0,248644
Neck	-0,062698	0,099279	-0,3743	0,59271	-0,63153	0,538638
Waist	0,464972	0,074399	0,6539	0,10463	6,24968	0,000030
Height	0,163146	0,050340	0,2885	0,08902	3,24085	0,006441
Calf	0,104696	0,082247	0,4749	0,37305	1,27295	0,225330
Thigh	0,100032	0,084464	0,3051	0,25761	1,18431	0,257491
Head	-0,097703	0,048529	-0,8526	0,42348	-2,01330	0,065266

I varijabla Fore je značajna!

Izbacujemo varijablu Neck.

4.korak

Regression Summary for Dependent Variable: Mass (primjer-7-2.sta)
R= ,98804860 R2= ,97624003 Adjusted R2= ,96436005
F(7,14)=82,175 p<,00000 Std.Error of estimate: 2,0697

N=22	b*	Std.Err. of b*	b	Std.Err. of b	t(14)	p-value
Intercept			-76,0501	24,05809	-3,16110	0,006937
Fore	0,284642	0,089208	1,6259	0,50955	3,19078	0,006539
Chest	0,084727	0,080476	0,1380	0,13103	1,05283	0,310254
Waist	0,452564	0,070200	0,6365	0,09873	6,44673	0,000015
Height	0,151979	0,046110	0,2687	0,08154	3,29601	0,005304
Calf	0,120562	0,076617	0,5468	0,34751	1,57357	0,137908
Thigh	0,105318	0,082224	0,3212	0,25077	1,28087	0,221053
Head	-0,094210	0,047166	-0,8221	0,41159	-1,99740	0,065598

Izbacujemo varijablu Chest.

5.korak

Regression Summary for Dependent Variable: Mass (primjer-7-2.sta)
R= ,98709617 R2= ,97435885 Adjusted R2= ,96410239
F(6,15)=95,000 p<,00000 Std.Error of estimate: 2,0772

N=22	b*	Std.Err. of b*	b	Std.Err. of b	t(15)	p-value
Intercept			-79,7262	23,88925	-3,33733	0,004501
Fore	0,314224	0,084972	1,7949	0,48536	3,69796	0,002148
Waist	0,466943	0,069107	0,6567	0,09719	6,75677	0,000006
Height	0,143570	0,045577	0,2539	0,08059	3,15007	0,006605
Calf	0,111820	0,076440	0,5072	0,34671	1,46284	0,164148
Thigh	0,141967	0,074761	0,4330	0,22801	1,89896	0,076977
Head	-0,075315	0,043776	-0,6572	0,38200	-1,72048	0,105902

Izbacujemo varijablu Calf.

6.korak

Regression Summary for Dependent Variable: Mass (primjer-7-2.sta)
R= ,98524154 R2= ,97070089 Adjusted R2= ,96154492
F(5,16)=106,02 p<,00000 Std.Error of estimate: 2,1499

N=22	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			-80,4533	24,72024	-3,25455	0,004973
Fore	0,371707	0,077978	2,1232	0,44541	4,76680	0,000210
Waist	0,473276	0,071386	0,6656	0,10040	6,62978	0,000006
Height	0,156669	0,046253	0,2770	0,08179	3,38720	0,003760
Thigh	0,171539	0,074496	0,5232	0,22720	2,30267	0,035061
Head	-0,073014	0,045279	-0,6371	0,39512	-1,61254	0,126392

I varijabla Thigh je značajna!

Izbacujemo varijablu Head.

N=22	Regression Summary for Dependent Variable: Mass (primjer-7-2.sta) R= ,98282210 R2= ,96593928 Adjusted R2= ,95792499 F(4,17)=120,53 p<,00000 Std.Error of estimate: 2,2488 b*	Std.Err. of b*	b	Std.Err. of b	t(17)	p-value
Intercept			-113,312	14,63911	-7,74036	0,000001
Fore	0,356369	0,080957	2,036	0,46243	4,40196	0,000390
Waist	0,459959	0,074169	0,647	0,10431	6,20147	0,000010
Height	0,153677	0,048342	0,272	0,08548	3,17893	0,005491
Thigh	0,177084	0,077840	0,540	0,23740	2,27498	0,036145

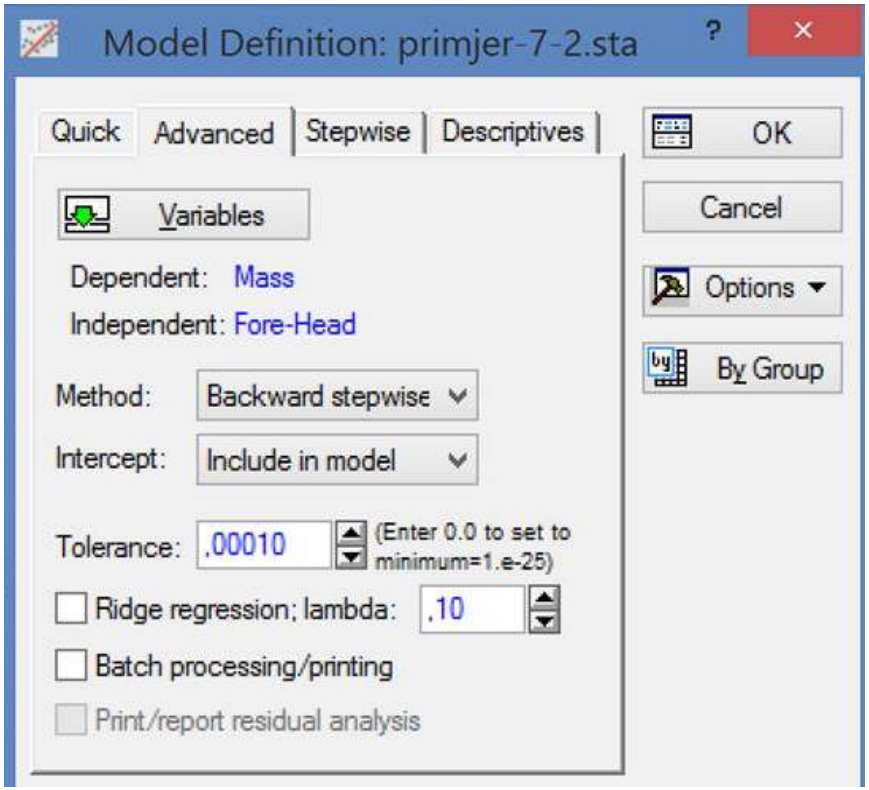
Sve su varijable značajne!

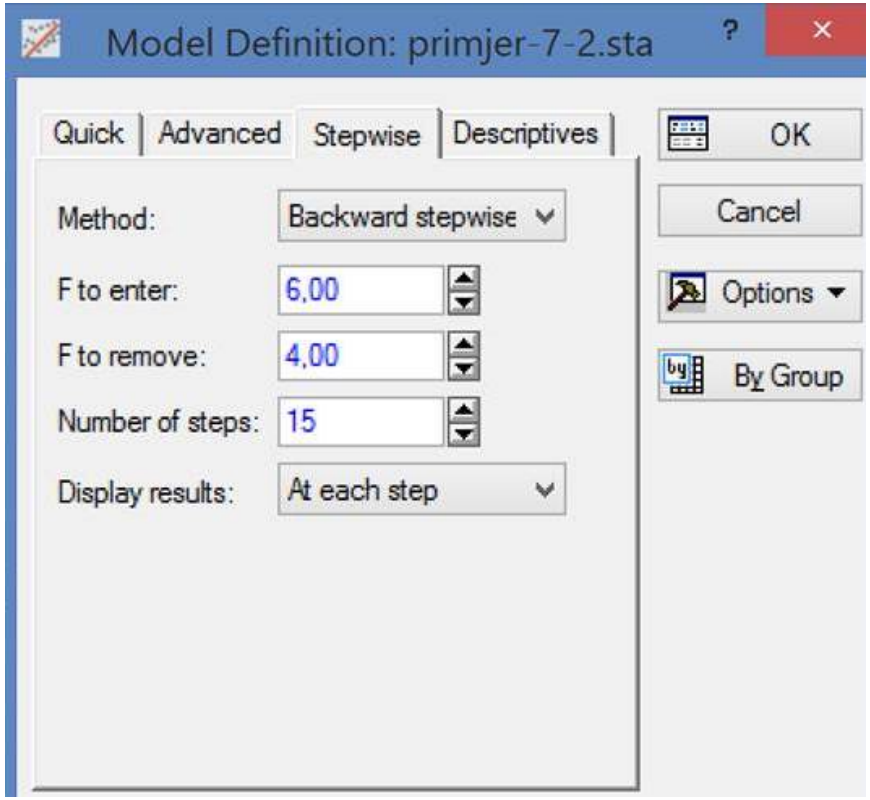
Model:

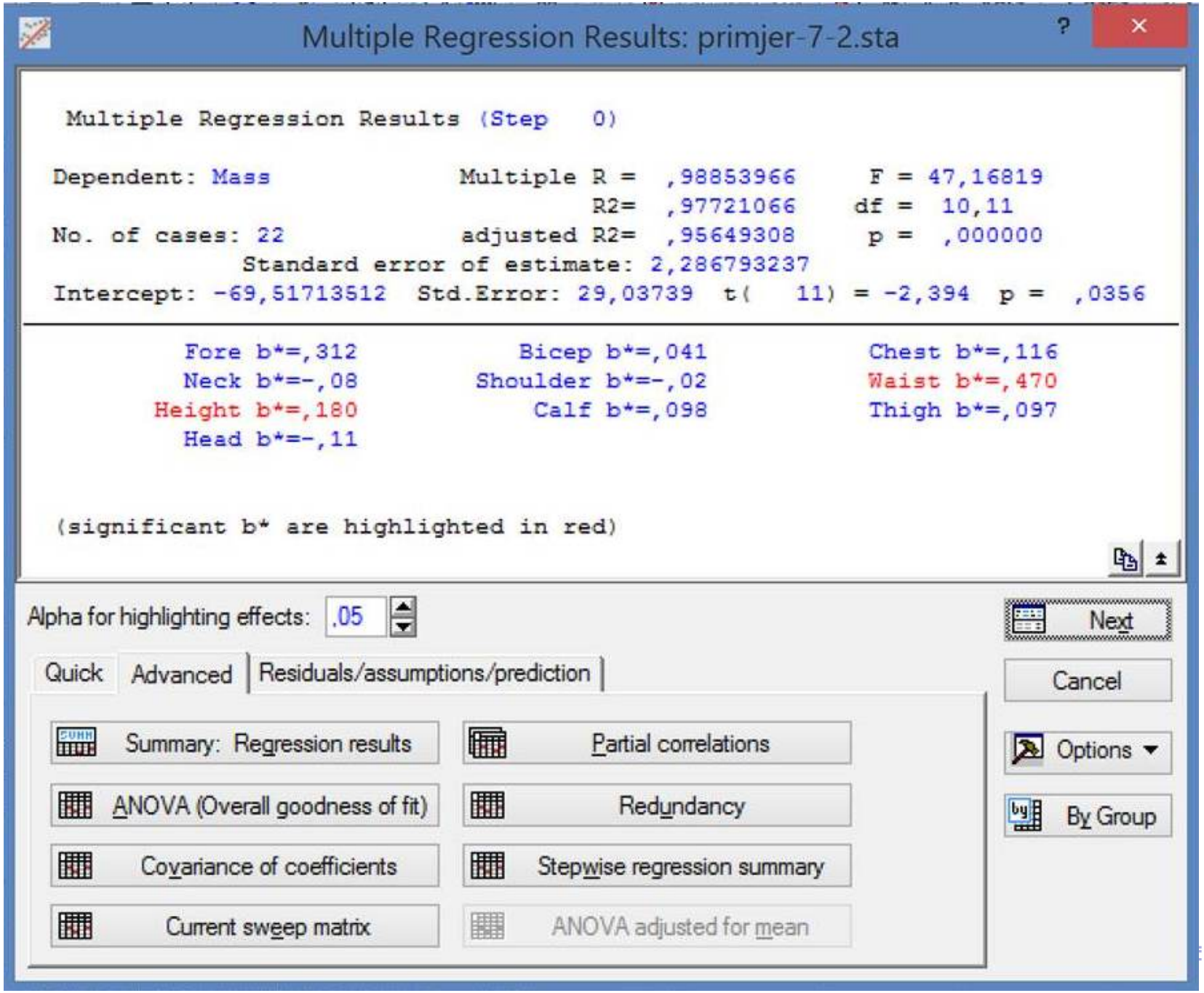
$$\text{Mass} = -113.312 + 2.036 \cdot \text{Fore} + 0.647 \cdot \text{Waist} +$$

$$+0.272 \cdot \text{Height} + 0.540 \cdot \text{Thigh}$$

Izbacivanje unazad u Statistici






Multiple Regression Results: primjer-7-2.sta
Multiple Regression Results (Step 0)
Dependent: Mass
Multiple R = ,98853966
F = 47,16819
R2= ,97721066
df = 10,11
No. of cases: 22
adjusted R2= ,95649308
p = ,000000
Standard error of estimate: 2,286793237
Intercept: -69,51713512 Std.Error: 29,03739 t(11) = -2,394 p = ,0356
Fore b*=,312
Bicep b*=,041
Chest b*=,116
Neck b*=-,08
Shoulder b*=-,02
Waist b*=,470
Height b*=,180
Calf b*=,098
Thigh b*=,097
Head b*=-,11
(significant b* are highlighted in red)
Alpha for highlighting effects: .05
Next
Quick
Advanced
Residuals/assumptions/prediction
Cancel
Summary: Regression results
Partial correlations
Options
ANOVA (Overall goodness of fit)
Redundancy
By Group
Covariance of coefficients
Stepwise regression summary
Current sweep matrix
ANOVA adjusted for mean

Multiple Regression Results (Step 1)

Dependent: Mass	Multiple R = ,98852395	F = 57,09394	
	R2= ,97717961	df = 9,12	
No. of cases: 22	adjusted R2= ,96006432	p = ,000000	
Standard error of estimate: 2,190928966			
Intercept: -70,53856768	Std.Error: 26,64703	t(12) = -2,647	p = ,0213

Fore b*=,301	Bicep b*=,043	Chest b*=,106
Neck b*=-,08	Waist b*=,468	Height b*=,176
Calf b*=,100	Thigh b*=,102	Head b*=-,10

Multiple Regression Results (Step 2)

Dependent: Mass	Multiple R = ,98840644	F = 68,86561	
	R2= ,97694728	df = 8,13	
No. of cases: 22	adjusted R2= ,96276100	p = ,000000	
Standard error of estimate: 2,115664350			
Intercept: -71,95026952	Std.Error: 25,43433	t(13) = -2,829	p = ,0142

Fore b*=,315	Chest b*=,118	Neck b*=-,06
Waist b*=,465	Height b*=,163	Calf b*=,105
Thigh b*=,100	Head b*=-,10	

Multiple Regression Results (Step 3)

Dependent: Mass — Multiple R = ,98804860 — F = 82,17520
R2= ,97624003 — df = 7,14
No. of cases: 22 — adjusted R2= ,96436005 — p = ,000000
Standard error of estimate: 2,069742330
Intercept: -76,05013222 Std.Error: 24,05809 t(14) = -3,161 p = ,0069

Fore b*=,285	Chest b*=,085	Waist b*=,453
Height b*=,152	Calf b*=,121	Thigh b*=,105
Head b*=-,09		

Multiple Regression Results (Step 4)

Dependent: Mass — Multiple R = ,98709617 — F = 94,99953
R2= ,97435885 — df = 6,15
No. of cases: 22 — adjusted R2= ,96410239 — p = ,000000
Standard error of estimate: 2,077210452
Intercept: -79,72624203 Std.Error: 23,88925 t(15) = -3,337 p = ,0045

Fore b*=,314	Waist b*=,467	Height b*=,144
Calf b*=,112	Thigh b*=,142	Head b*=-,08

Multiple Regression Results (Step 5)

Dependent: Mass	Multiple R = ,98524154	F = 106,0183	
	R2= ,97070089	df = 5,16	
No. of cases: 22	adjusted R2= ,96154492	p = ,000000	
Standard error of estimate: 2,149931472			
Intercept: -80,45329530	Std.Error: 24,72024	t(16) = -3,255	p = ,0050

Fore b*=,372	Waist b*=,473	Height b*=,157
Thigh b*=,172	Head b*=-,07	

Multiple Regression Results (step 6, final solution)
no other F to remove is less than specified limit

Dependent: Mass	Multiple R = ,98282210	F = 120,5272	
	R2= ,96593928	df = 4,17	
No. of cases: 22	adjusted R2= ,95792499	p = ,000000	
Standard error of estimate: 2,248846644			
Intercept: -113,3120436	Std.Error: 14,63911	t(17) = -7,740	p = ,0000

Fore b*=,356	Waist b*=,460	Height b*=,154
Thigh b*=,177		

Varijable izbačene u pojedinom koraku

Variable	Summary of Stepwise Regression; DV: Mass (primjer-7-2.sta)						
	Step +in/-out	Multiple R	Multiple R-square	R-square change	F - to entr/rem	p-value	Variables included
Shoulder	-1	0,988524	0,977180	-0,000031	0,014988	0,904769	9
Bicep	-2	0,988406	0,976947	-0,000232	0,122168	0,732752	8
Neck	-3	0,988049	0,976240	-0,000707	0,398836	0,538638	7
Chest	-4	0,987096	0,974359	-0,001881	1,108442	0,310254	6
Calf	-5	0,985242	0,970701	-0,003658	2,139897	0,164148	5
Head	-6	0,982822	0,965939	-0,004762	2,600275	0,126392	4

Može se koristiti i strategija dodavanja najznačajnije varijable (**dodavanje unaprijed**).

Prvo u model stavimo varijablu koja ima najveću objašnjenu varijancu (najmanju p-vrijednost u modelima s jednom varijablom).

Zatim, korak po korak, dodajemo varijablu koja najviše povećava objašnjenu varijancu.

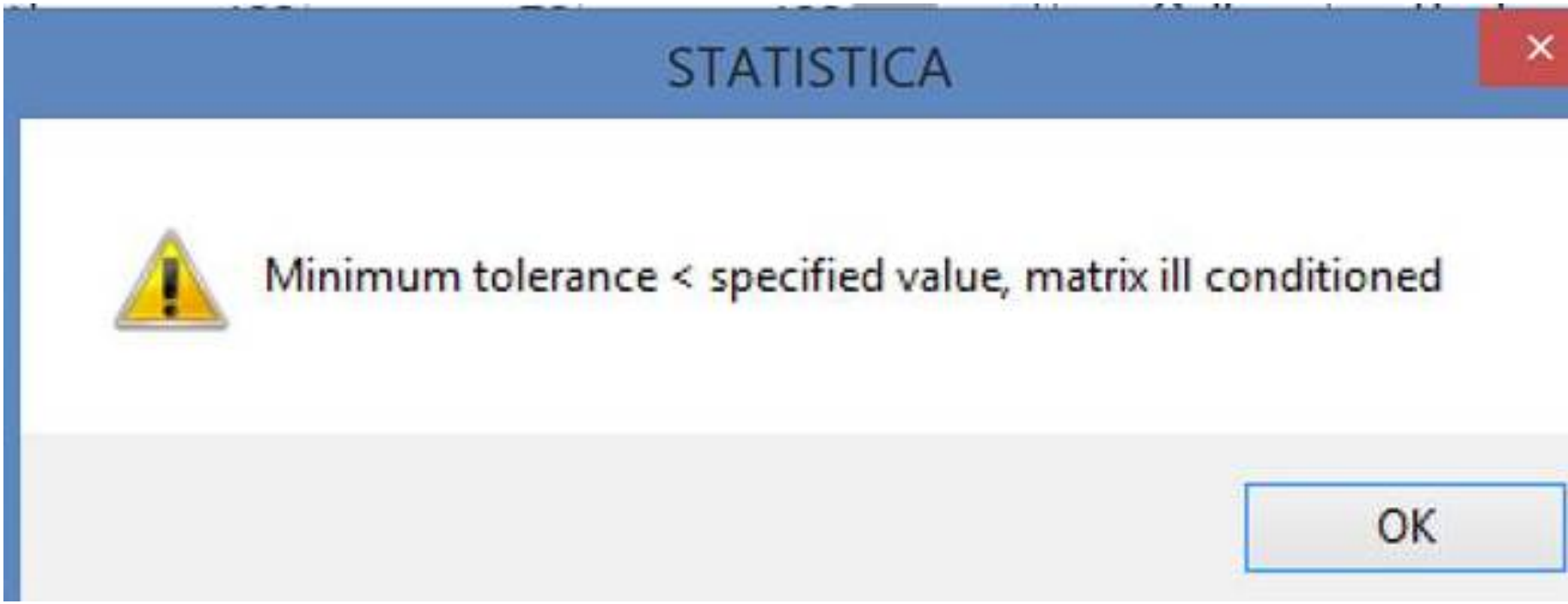
Može se koristiti i kombinacija ova dva pristupa, u svakom koraku ubacimo ili izbacimo po jednu varijablu.

Napomene o regresiji

- Broj podataka treba biti barem 5 puta veći od od broja parametara (broj varijabli + slobodni koeficijent).
- Preveliki broj podatak može rezultirati zaključkom da su sve varijable značajne.
- Preporučljivo je između 20 i 40 podataka po parametru.
- Normalnost. Pretpostavka je da je zavisna varijabla Y normalna za svaku moguću vrijednost varijabli $X_1, X_2, \ldots, X_k$.
- Varijanca slučajne varijable Y treba biti ista za svaku moguću vrijednost varijabli $X_1, X_2, \ldots, X_k$. (homoskedastičnost)
- postojanje ekstremnih vrijednosti (outliera) može znatno utjecati na rezultat regresije.
- Regresijskom analizom ispitujemo povezanost **neprekidnih** varijabli.

Napomene o regresiji

- Ukoliko je jedna nezavisna varijabla linearna kombinacija nekoliko preostalih varijabli tada se ne mogu odrediti regresijski koeficijenti.



Ovo se najčešće događkada jednu varijablu u regresiji definiramo preko nekoliko drugih varijabli (zbroj ili aritmetička sredina)

Napomene o regresiji

- U regresiju je moguće uključiti i kategorijske varijable upotrebom tzv. praznih ('dummy') varijabli.
- Ukoliko želimo u regresiji kao zavisnu varijablu koristiti kategorijsku (dihotomnu) varijablu koristi se **logistička regresija**.
- Povezanost varijabli ne znači uzročno posljedičnu povezanost!