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

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

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

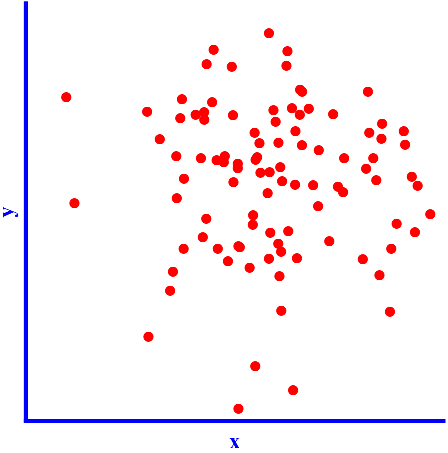
σ_Y - standardna devijacija obilježja Y

σ_X - standardna devijacija obilježja X

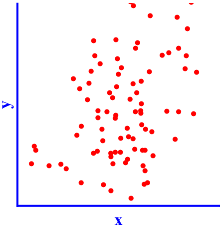
σ_Y - standardna devijacija obilježja Y

Dijagram raspršenja

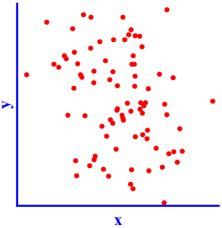
Za verijable X i Y promatramo (X, Y)-graf:



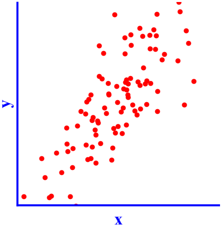
$\rho = 0$



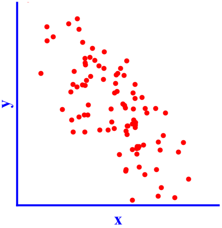
$\rho = 0.47$



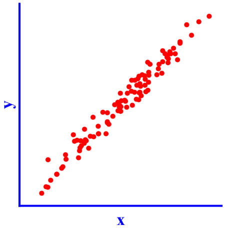
$\rho = -0.46$



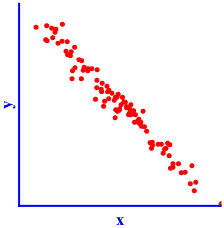
$\rho = 0.80$



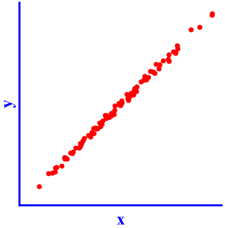
$\rho = -0.82$



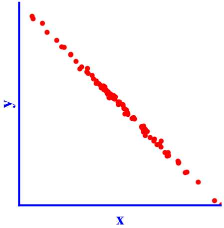
$\rho = 0.98$



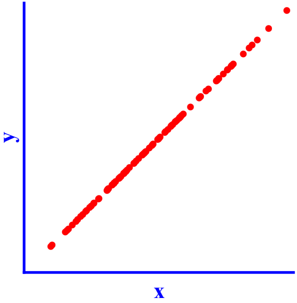
$\rho = -0.98$



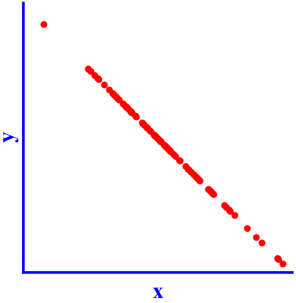
$\rho = 0.999$



$\rho = -0.999$



$\rho = 1$



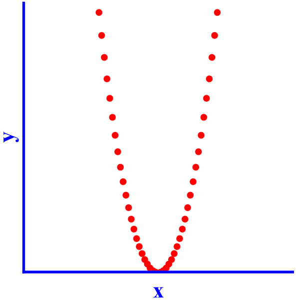
$\rho = -1$

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

$S_X = \sqrt{S_X^2}$ - 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

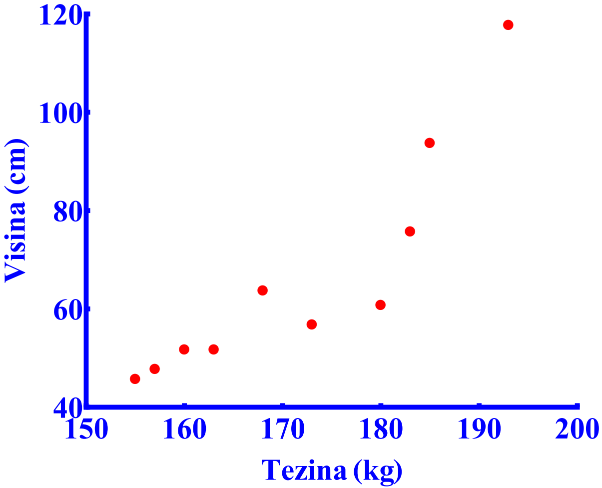
$$t = \frac{\sqrt{n-2} \cdot r}{\sqrt{1 - r^2}}$$

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:



$$r = 0.89$$

R

Podaci:

```
> visina <-
c(183,163,180,168,160,157,185,155,193,173)
> tezina <- c(76,52,61,64,52,48,94,46,118,57)
```

Korelacija:

```
> cor(visina,tezina)
[1] 0.8906609
```

Testiranje hipoteze $\rho = 0$:

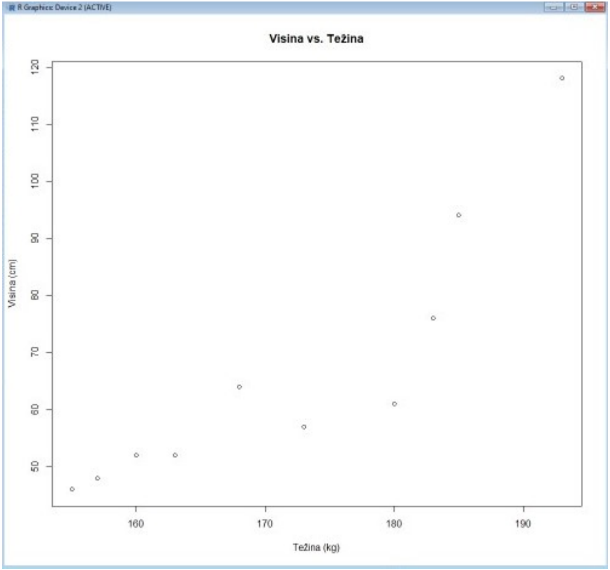
```
> t <- cor.test(visina,tezina)
> t

    Pearson's product-moment correlation

data:  visina and tezina
t = 5.5407, df = 8, p-value = 0.0005469
alternative hypothesis:  true correlation is not
equal to 0
95 percent confidence interval:
0.5943188 0.9740538
sample estimates:
cor
0.8906609
```

Dijagram raspršenja:

```
> plot(tezina ~ visina ,
  main = "Visina vs. Težina",
  xlab = "Težina (kg)",
  ylab = "Visina (cm)")
```


R Graphics Device 2 (ACTIVE)
Visina vs. Težina
Visina (cm)
50
60
70
80
90
100
110
120
160
170
180
190
Težina (kg)

Interpretacija:

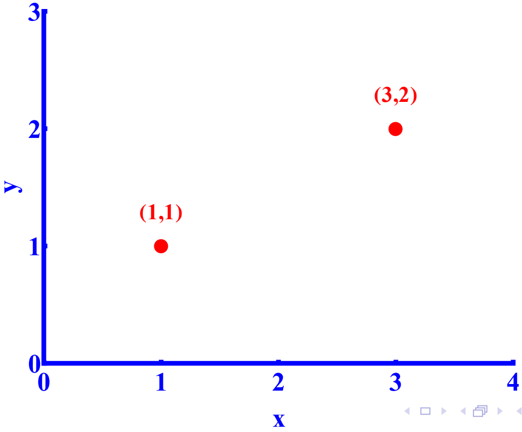
- Kod osobe s većom visinom očekujemo i veću težinu (pozitivna koreliranost)
- Kod osobe s većom težinom očekujemo i veću visinu (pozitivna koreliranost)
- Korelacija ne pokazuje uzročno-posljedičnu povezanost!
- Povećanjem težine nećemo povećati visinu.

LINEARNA REGRESIJA

Jednostavna linearna regresija
Pravac

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

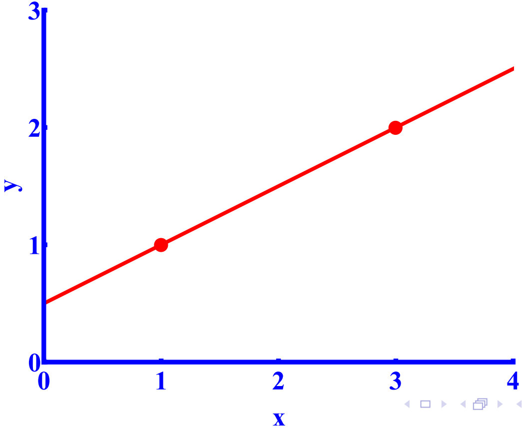
Nacrtamo točke:



Jednostavna linearna regresija
Pravac

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

Nacrtamo točke i provučemo pravac kroz njih:



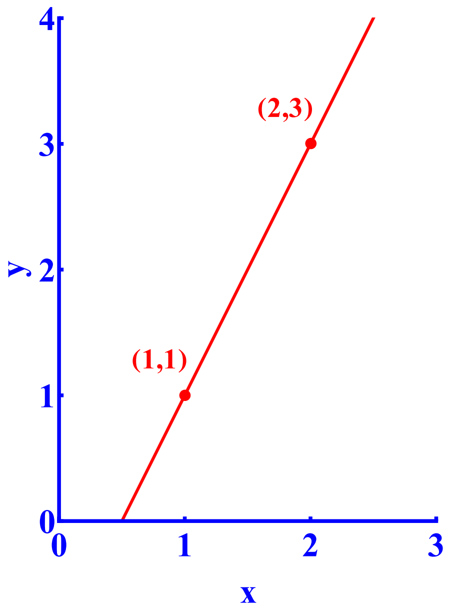
Primjer. Nacrtajte pravac $y = 2 \cdot x - 1$.

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.



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

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

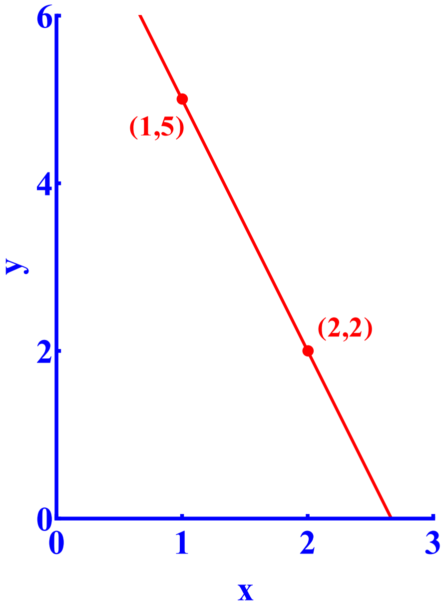
Odredimo dvije točke na pravcu:

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

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

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

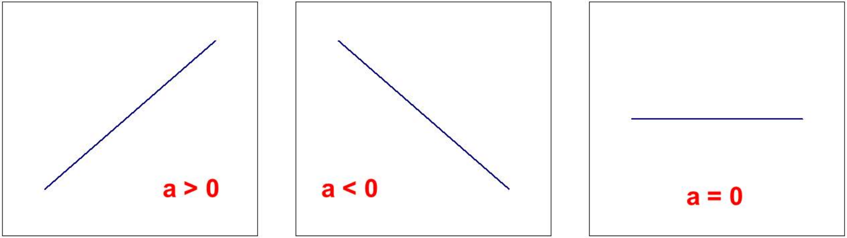
Jednadžba pravca:

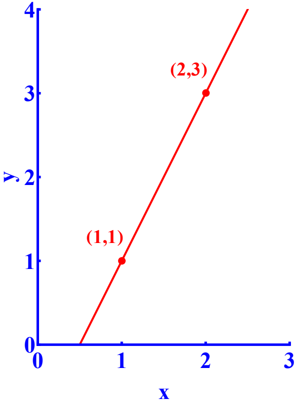
$$y = a \cdot x + b.$$

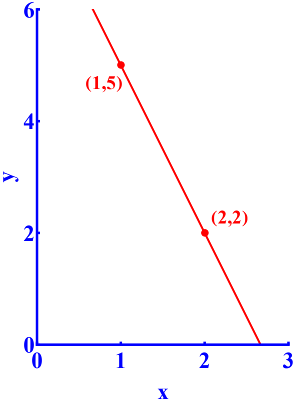
a - koeficijent smjera (*engl.* 'slope')

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

Koeficijent smjera







$y = \mathbf{2} \cdot x - 1$ $\qquad$ $y = \mathbf{-3} \cdot x + 8$

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

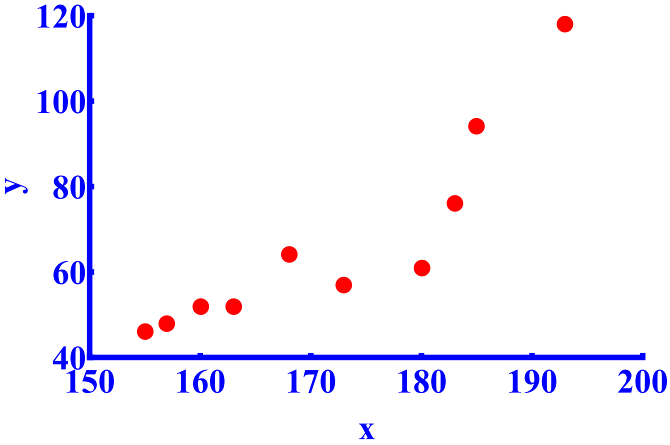
Odrediti oblik linearne veze znači odrediti vezu oblika

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

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

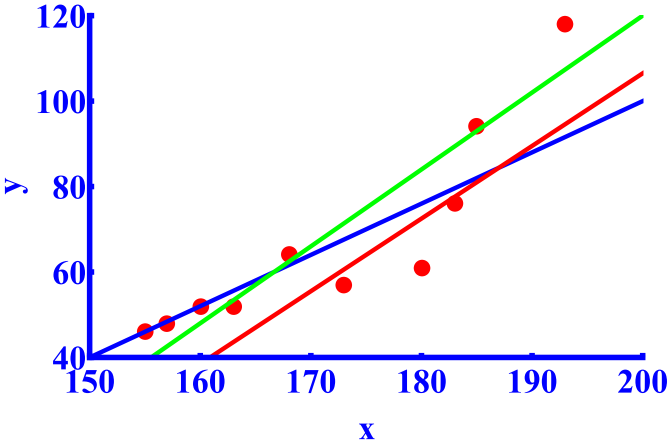
Kako odrediti koeficijente a i b?

Podaci za visinu i težinu:



Kako odrediti pravac koji najbolje opisuje podatke?

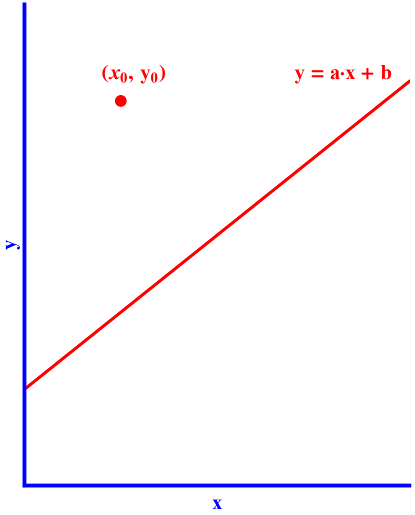
Podaci za visinu i težinu:



Koji pravac bolje opisuje podatke?

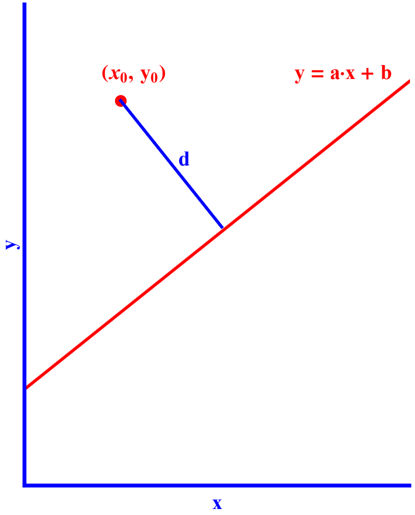
Kvadratno odstupanje

Udaljenost pravca od točke (podatka)



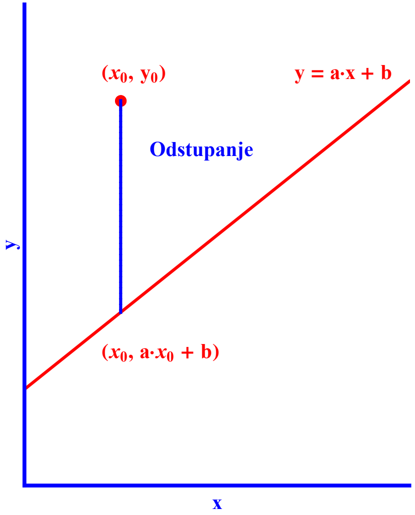
Kvadratno odstupanje

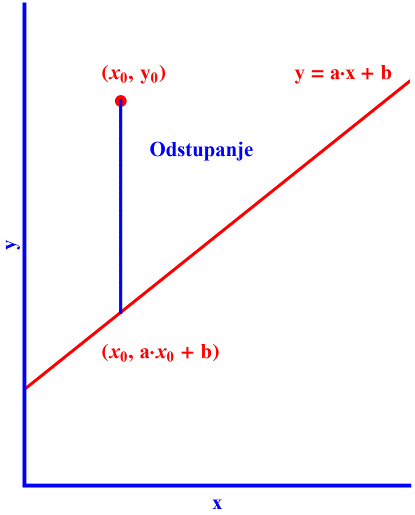
Udaljenost pravca od točke (podatka)



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

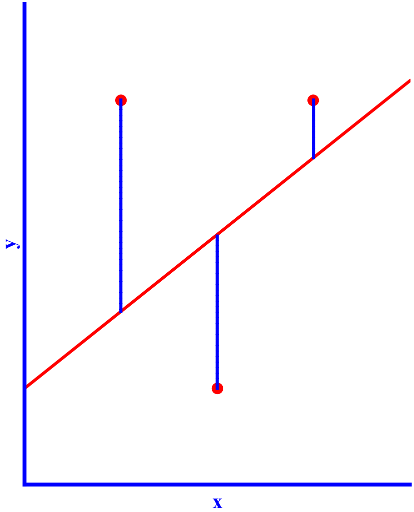
U regresiji se najčešće koristi kvadratno odstupanje.

Kako definirati udaljenost pravca od skupa točaka?

Srednje kvadratno odstupanje aritmetička sredina kvadratnih odstupanja od pojedine točke.

Srednje kvadratno odstupanje

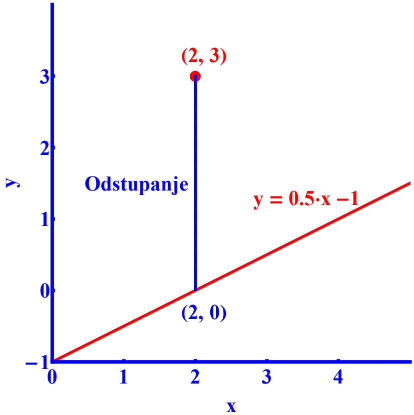
:



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

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

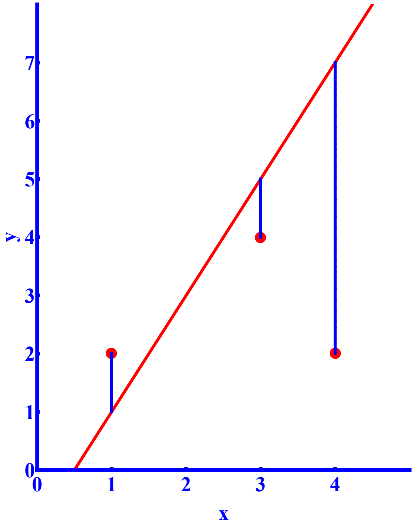
$(a \cdot x_0 + b - y_0)^2 = (0.5x_0 - 1 - y_0)^2 = (0.5 \cdot 2 - 1 - 3)^2 = (-3)^2 = 9$



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

x	y
1	2
4	2
3	4

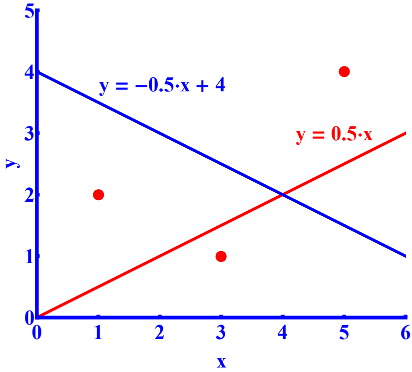
x	y	$2x-1$	$2x-1-y$	$(2x-1-y)^2$
1	2	1	-1	1
4	2	7	5	25
3	4	5	1	1
$\sum$				27
$\sum/n$				9

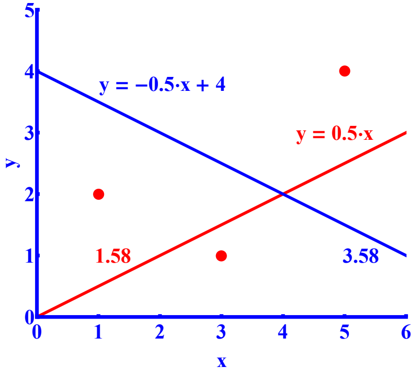


0
1
2
3
4
5
6
7
y
0
1
2
3
4
x

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**.

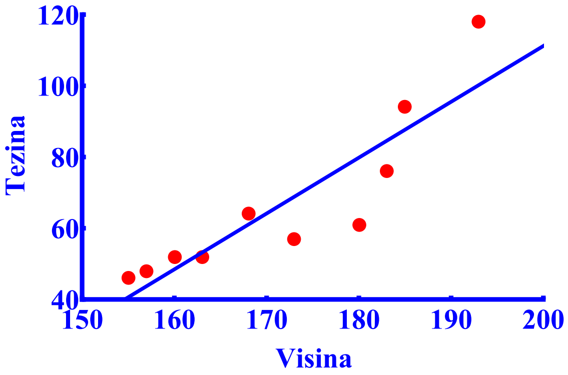
Eksplicitni izraz za regresijske koeficijente:

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

$r_{X,Y}$ - 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

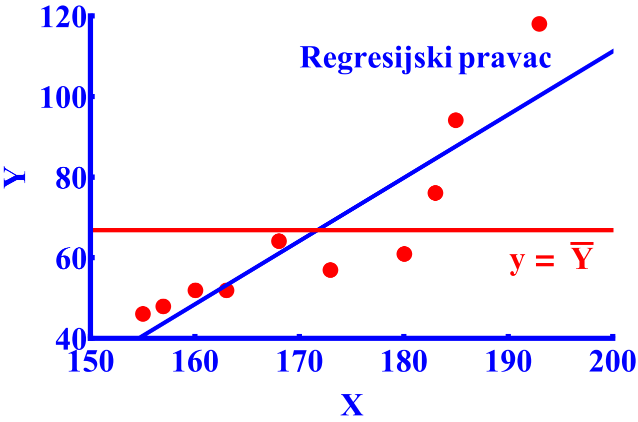
$$r_{Z_X,Z_Y} = r_{X,Y}.$$

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

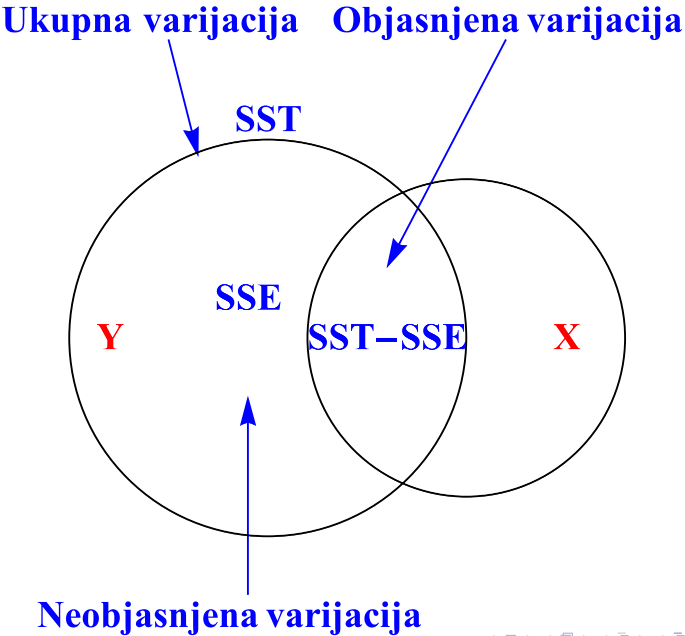
$$\sum_i (a \cdot x_i + b - y_i)^2$$

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

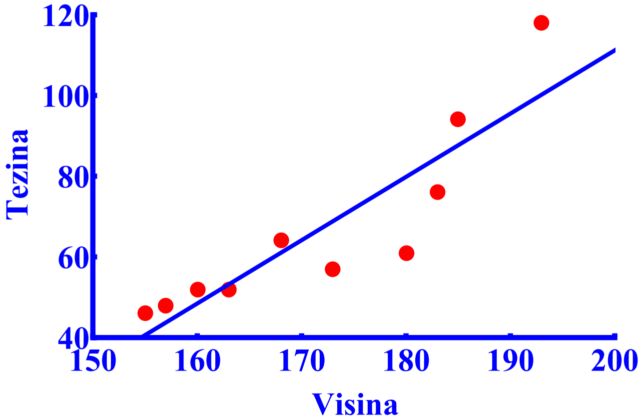
$r^2 = 0$ - nema utjecaja obilježja X na obilježje Y

Veza koeficijenta determinacije i koeficijenta korelacije:

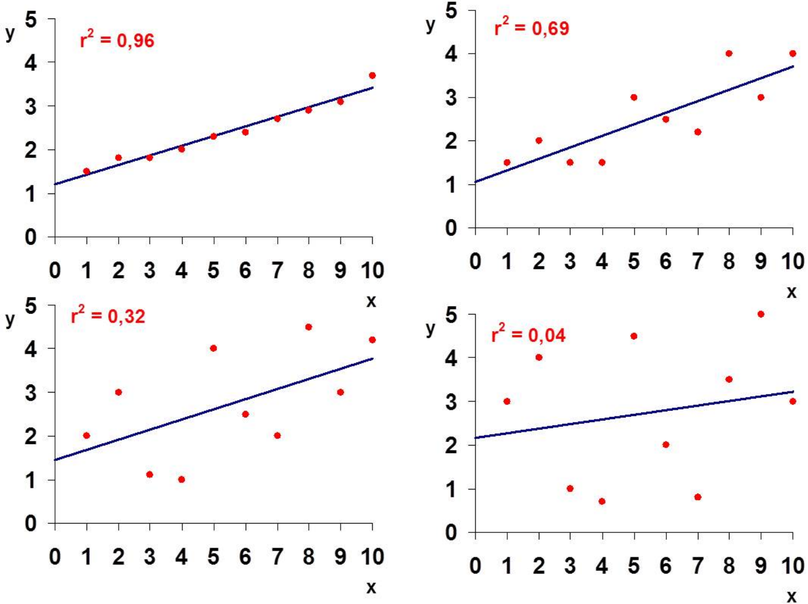
$$r^2 = r_{XY}^2$$

r^2 - koeficijent determinacije

r_{XY} - koeficijent korelacije



$$r^2 = 0.79$$


y
$r^2 = 0{,}96$
x
y
$r^2 = 0{,}69$
x
y
$r^2 = 0{,}32$
x
y
$r^2 = 0{,}04$
x

Parcijalna korelacija

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

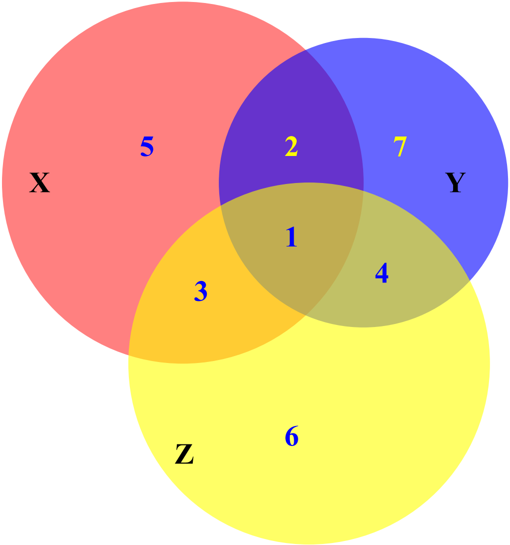
Od varijabli X i Y oduzmemo dio koji opisuje Z (dobiven regresijom, za svaku varijablu posebno):

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

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

Parcijalna korelacija od X i Y:

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


5
2
7
X
Y
1
4
3
6
Z

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:

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

Vrijedi

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

Može se pokazati da je

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

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

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

Statistika:

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

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.

R

Primjer. Podaci o visini i težini.

```
> lm(tezina ~ visina)
Call:
lm(formula = tezina   visina)
 Coefficients:
 (Intercept)     visina
 -202.519        1.569
```

Prikazani su samo koeficijent.

Drugi pristup:

```
> regresija=lm(tezina   visina)
> regresija
Call:
lm(formula = tezina   visina)

Residuals:
     Min       1Q   Median      3Q      Max
 -18.819   -6.682    3.278   5.110   17.790  Coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)  -202.5190     48.7352   -4.155  0.003185  **
 visina          1.5685      0.2831    5.541  0.000547  ***
--
Signif. codes: 0 ´***´ 0.001 ´**´ 0.01 ´*´ 0.05 ´.´ 0.1
´ ´ 1
```

```
Residual standard error: 11.15 on 8 degrees of
freedom
Multiple R-squared: 0.7933, Adjusted R-squared:
0.7674
F-statistic: 30.7 on 1 and 8 DF, p-value:
0.0005469
```

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:

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

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

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

Objašnjena varijanca od Y: $\mathbf{SST} - \mathbf{SSE}$

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$_k = \sum_i \left(b_0 + b_1 \cdot x_{1i} + a_2 \cdot x_{2i} + \ldots + b_{k-1} \cdot x_{(k-1)i} - y_i\right)^2$

Distribucija:

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

Statistika:

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

R. 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?

Regresija:

```
> reg=lm(Mass ~ Fore + Bicep + Chest + Neck +
Shoulder + Waist + Height + Calf + Thigh + Head,
data = podaci)
> summary(reg)
```

```
R Console
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -69.51714   29.03739  -2.394 0.035605 *  
Fore          1.78182    0.85473   2.085 0.061204 .  
Bicep         0.15509    0.48530   0.320 0.755275    
Chest         0.18914    0.22583   0.838 0.420132    
Neck         -0.48184    0.72067  -0.669 0.517537    
Shoulder     -0.02931    0.23943  -0.122 0.904769    
Waist         0.66144    0.11648   5.679 0.000143 ***
Height        0.31785    0.13037   2.438 0.032935 *  
Calf          0.44589    0.41251   1.081 0.302865    
Thigh         0.29721    0.30510   0.974 0.350917    
Head         -0.91956    0.52009  -1.768 0.104735    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.287 on 11 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9565 
F-statistic: 47.17 on 10 and 11 DF,  p-value: 1.408e-07

> 
```

$r^2 = 0.9772$ - varijable dobro opisuju masu

$p = 1.408 \cdot 10^{-7}$ - model je značajan

Height, waist su značajne

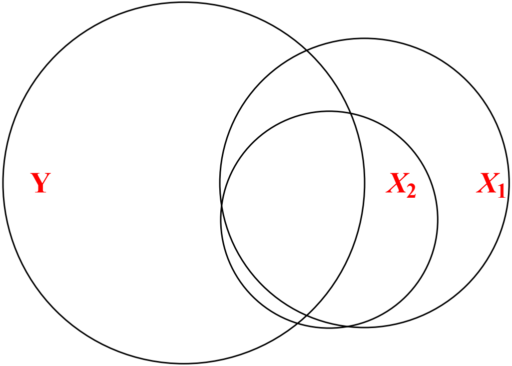
Jesu li druge varijable značajne u opisu mase?

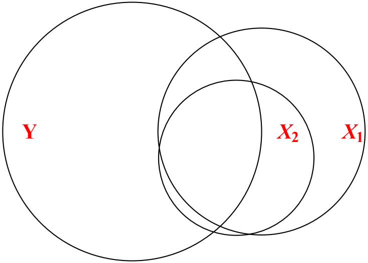
Izgradnja modela u višestrukoj regresiji

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

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

Jer je

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

modeli

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

i

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

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

Varijable Hm i Hinch će biti nesignifikantne u modelu

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

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.

R Console

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -69.51714   29.03739  -2.394 0.035605 *
Fore          1.78182    0.85473   2.085 0.061204 .
Bicep         0.15509    0.48530   0.320 0.755275
Chest         0.18914    0.22583   0.838 0.420132
Neck         -0.48184    0.72067  -0.669 0.517537
Shoulder     -0.02931    0.23943  -0.122 0.904769
Waist         0.66144    0.11648   5.679 0.000143 ***
Height        0.31785    0.13037   2.438 0.032935 *
Calf          0.44589    0.41251   1.081 0.302865
Thigh         0.29721    0.30510   0.974 0.350917
Head         -0.91956    0.52009  -1.768 0.104735
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.287 on 11 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9565
F-statistic: 47.17 on 10 and 11 DF,  p-value: 1.408e-07

>
```

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

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

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -69.51714   29.03739  -2.394 0.035605 *
Fore          1.78182    0.85473   2.085 0.061204 .
Bicep         0.15509    0.48530   0.320 0.755275
Chest         0.18914    0.22583   0.838 0.420132
Neck         -0.48184    0.72067  -0.669 0.517537
Shoulder     -0.02931    0.23943  -0.122 0.904769
Waist         0.66144    0.11648   5.679 0.000143 ***
Height        0.31785    0.13037   2.438 0.032935 *
Calf          0.44589    0.41251   1.081 0.302865
Thigh         0.29721    0.30510   0.974 0.350917
Head         -0.91956    0.52009  -1.768 0.104735
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.287 on 11 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9565
F-statistic: 47.17 on 10 and 11 DF,  p-value: 1.408e-07

>
```

Izbacujemo varijablu Shoulder.

2.korak

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -70.5386    26.6470  -2.647   0.0213 *
Fore          1.7179     0.6484   2.650   0.0212 *
Bicep         0.1615     0.4622   0.350   0.7328
Chest         0.1729     0.1749   0.988   0.3425
Neck         -0.4846     0.6901  -0.702   0.4960
Waist         0.6585     0.1091   6.034  5.9e-05 ***
Height        0.3108     0.1122   2.771   0.0169 *
Calf          0.4529     0.3914   1.157   0.2698
Thigh         0.3123     0.2676   1.167   0.2659
Head         -0.8932     0.4537  -1.969   0.0725 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.191 on 12 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9601
F-statistic: 57.09 on 9 and 12 DF,  p-value: 1.784e-08

>
```

I varijabla Fore je značajna!

Izbacujemo varijablu Bicep.

3.korak

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -71.95027   25.43433  -2.829  0.01422 *
Fore          1.79678    0.58696   3.061  0.00910 **
Chest         0.19282    0.15965   1.208  0.24864
Neck         -0.37432    0.59271  -0.632  0.53864
Waist         0.65393    0.10463   6.250 2.97e-05 ***
Height        0.28849    0.08902   3.241  0.00644 **
Calf          0.47487    0.37305   1.273  0.22533
Thigh         0.30508    0.25761   1.184  0.25749
Head         -0.85259    0.42348  -2.013  0.06527 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.116 on 13 degrees of freedom
Multiple R-squared:  0.9769,    Adjusted R-squared:  0.9628
F-statistic: 68.87 on 8 and 13 DF,  p-value: 2.165e-09

>
```

Izbacujemo varijablu Neck.

4.korak

```
R Console
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -76.05013   24.05809  -3.161  0.00694 **
Fore          1.62588    0.50955   3.191  0.00654 **
Chest         0.13796    0.13103   1.053  0.31025
Waist         0.63648    0.09873   6.447 1.53e-05 ***
Height        0.26875    0.08154   3.296  0.00530 **
Calf          0.54684    0.34751   1.574  0.13791
Thigh         0.32121    0.25077   1.281  0.22105
Head         -0.82210    0.41159  -1.997  0.06560 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.07 on 14 degrees of freedom
Multiple R-squared:  0.9762,    Adjusted R-squared:  0.9644
F-statistic: 82.18 on 7 and 14 DF,  p-value: 2.744e-10

>
```

Izbacujemo varijablu Chest.

5.korak

```
R Console

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -79.72624   23.88925  -3.337  0.00450 **
Fore          1.79485    0.48536   3.698  0.00215 **
Waist         0.65671    0.09719   6.757 6.45e-06 ***
Height        0.25388    0.08059   3.150  0.00661 **
Calf          0.50718    0.34671   1.463  0.16415
Thigh         0.43298    0.22801   1.899  0.07698 .
Head         -0.65722    0.38200  -1.720  0.10590
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.077 on 15 degrees of freedom
Multiple R-squared:  0.9744,    Adjusted R-squared:  0.9641
F-statistic:    95 on 6 and 15 DF,  p-value: 4.501e-11

>
```

Izbacujemo varijablu Calf.

6.korak

R Console

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -80.45330   24.72024  -3.255  0.00497 ** 
Fore          2.12319    0.44541   4.767  0.00021 ***
Waist         0.66561    0.10040   6.630 5.79e-06 ***
Height        0.27704    0.08179   3.387  0.00376 ** 
Thigh         0.52317    0.22720   2.303  0.03506 *  
Head         -0.63714    0.39512  -1.613  0.12639    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.15 on 16 degrees of freedom
Multiple R-squared:  0.9707,    Adjusted R-squared:  0.9615 
F-statistic:   106 on 5 and 16 DF,  p-value: 1.104e-11

> |
```

I varijabla Thigh je značajna!

Izbacujemo varijablu Head.

R Console

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -113.31204   14.63911  -7.740 5.70e-07 ***
Fore           2.03558    0.46243   4.402  0.00039 ***
Waist          0.64688    0.10431   6.201 9.67e-06 ***
Height         0.27175    0.08548   3.179  0.00549 **
Thigh          0.54008    0.23740   2.275  0.03614 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.249 on 17 degrees of freedom
Multiple R-squared:  0.9659,    Adjusted R-squared:  0.9579
F-statistic: 120.5 on 4 and 17 DF,  p-value: 3.079e-12
```

Sve su varijable značajne!

Model:

$$\begin{aligned}\text{Mass} &= -113.312 + 2.036\cdot\text{Fore} + 0.647\cdot\text{Waist} + \\ &\quad +0.272\cdot\text{Height} + 0.540\cdot\text{Thigh}\end{aligned}$$

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 događa kada jednu varijablu u regresiji definiramo preko nekoliko drugih varijabli (zbroj ili aritmetička sredina)
- 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!