Korelacija i Regresija

Nedelja 9/10 - Vežbe

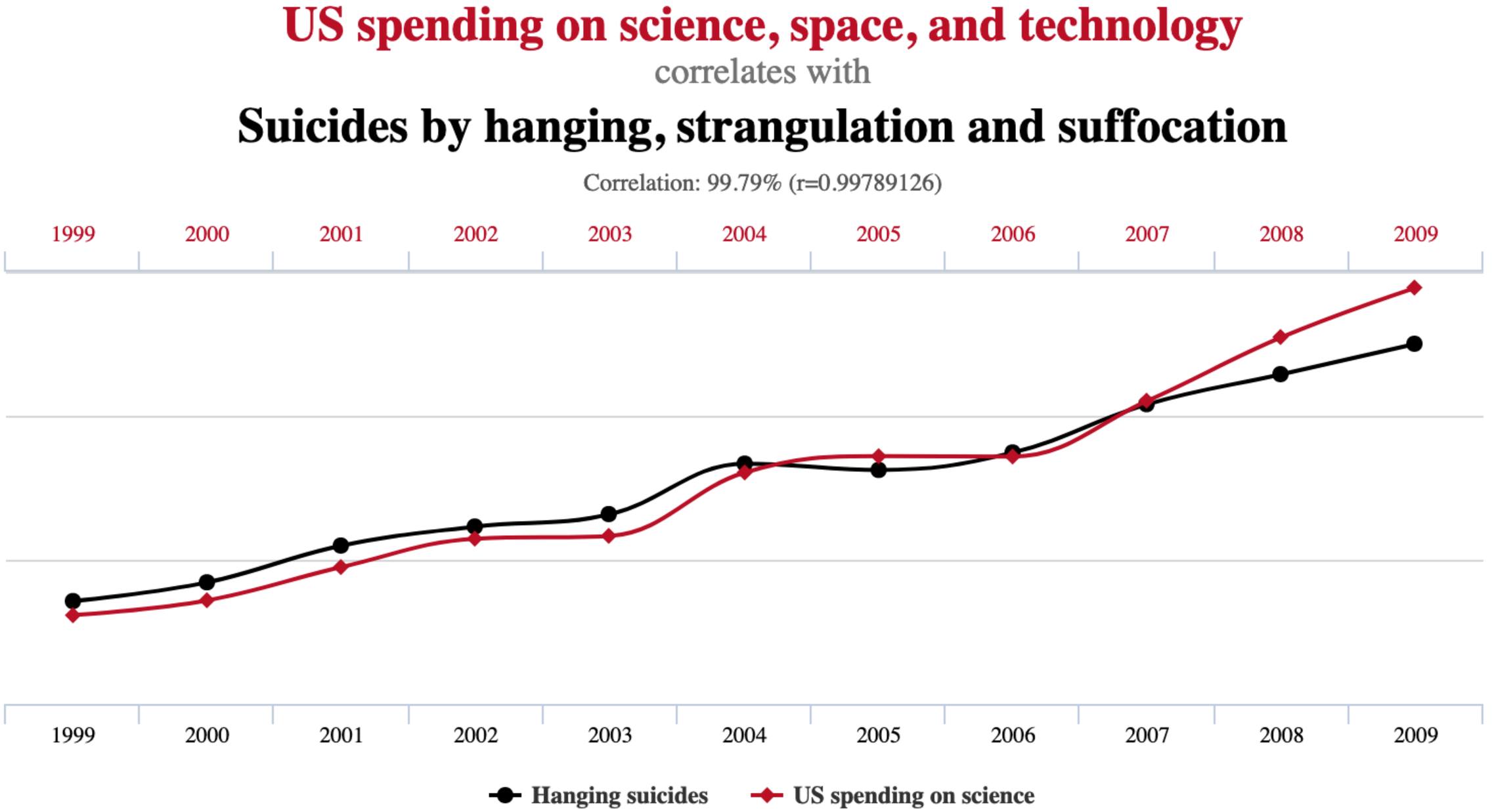
Dr Nikola N. Grubor

Stephen Jay Gould

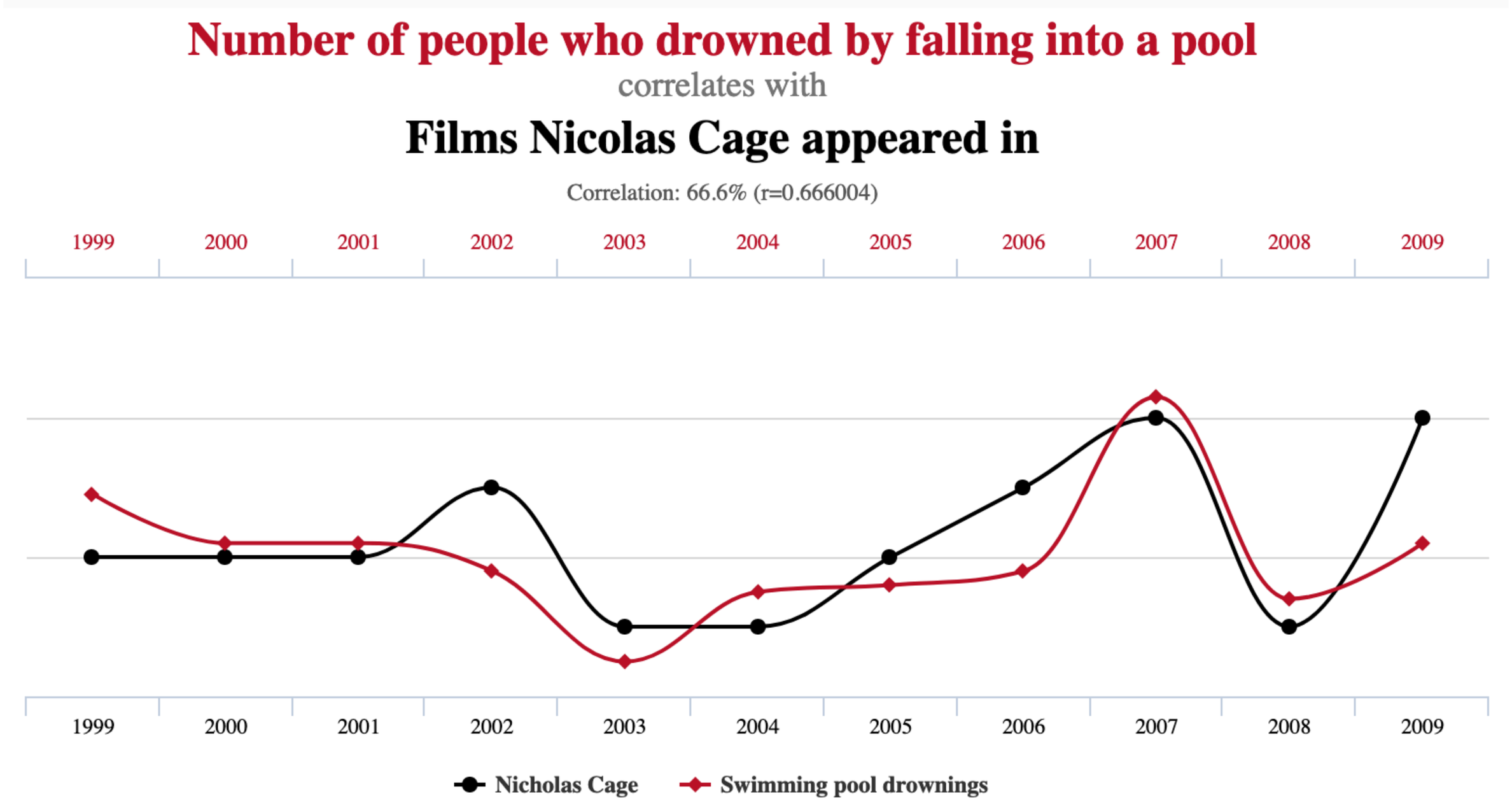
“Netačna pretpostavka da korelacija podrazumeva uzročnu vezu je verovatno jedna od dve ili tri najveće greške u čovekovom zaključivanju.”

Korelacija ≠ Kauzalnost

Korelacija ≠ Kauzalnost



Korelacija ≠ Kauzalnost



When doctors strike, fewer patients die

By **Ryan Hoskins** Globe Correspondent, February 9, 2016, 1:36 p.m.

July effect

2 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

"Killing season" redirects here. For other uses, see Killing Season (disambiguation).

The **July effect**, sometimes referred to as the **July phenomenon**, is a perceived but scientifically unfounded increase in the risk of medical errors and surgical complications that occurs in association with the time of year in which United States medical school graduates begin residencies.[1] A similar period in the United Kingdom is known as the **killing season** or, more specifically, **Black Wednesday,** referring to the first Wednesday in August when postgraduate trainees commence their rotations.

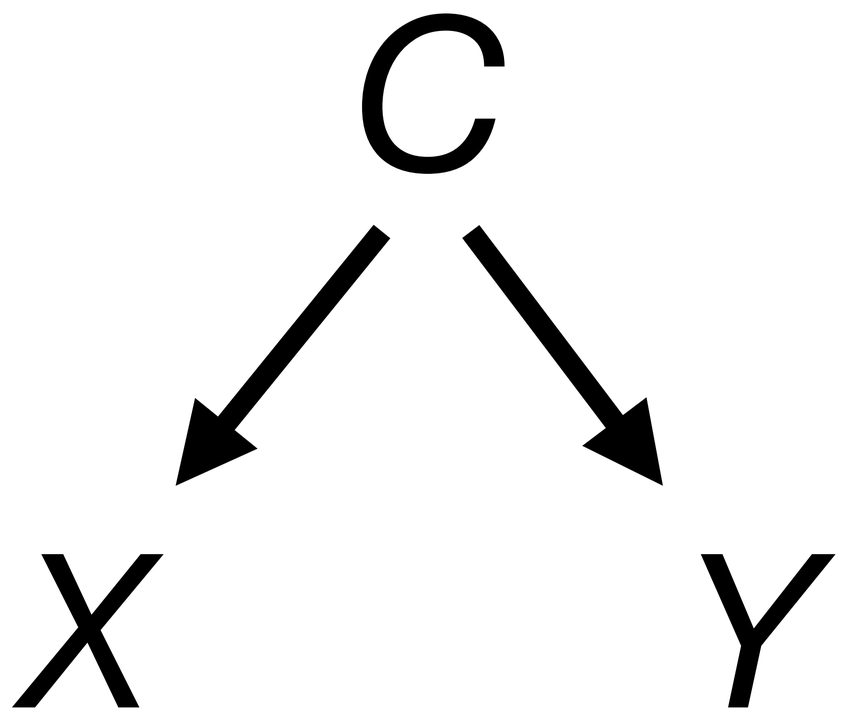
Objašnjenje povezanosti

Objašnjenje povezanosti

Objašnjenje povezanosti

X Y **Slučajno povezane**

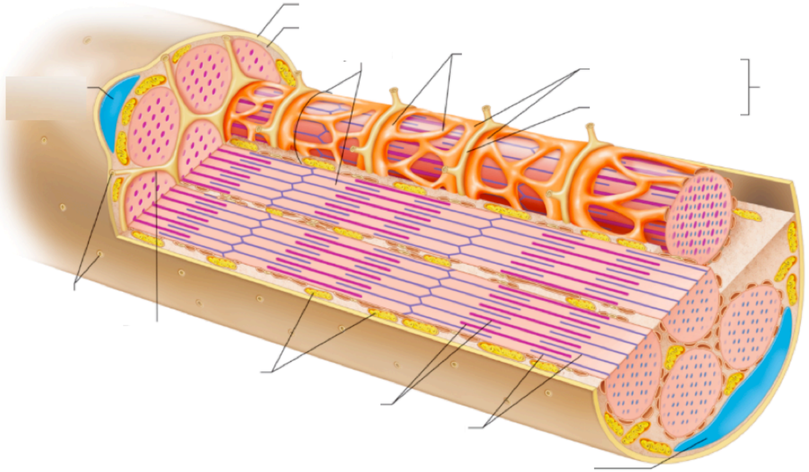
Objašnjenje povezanosti

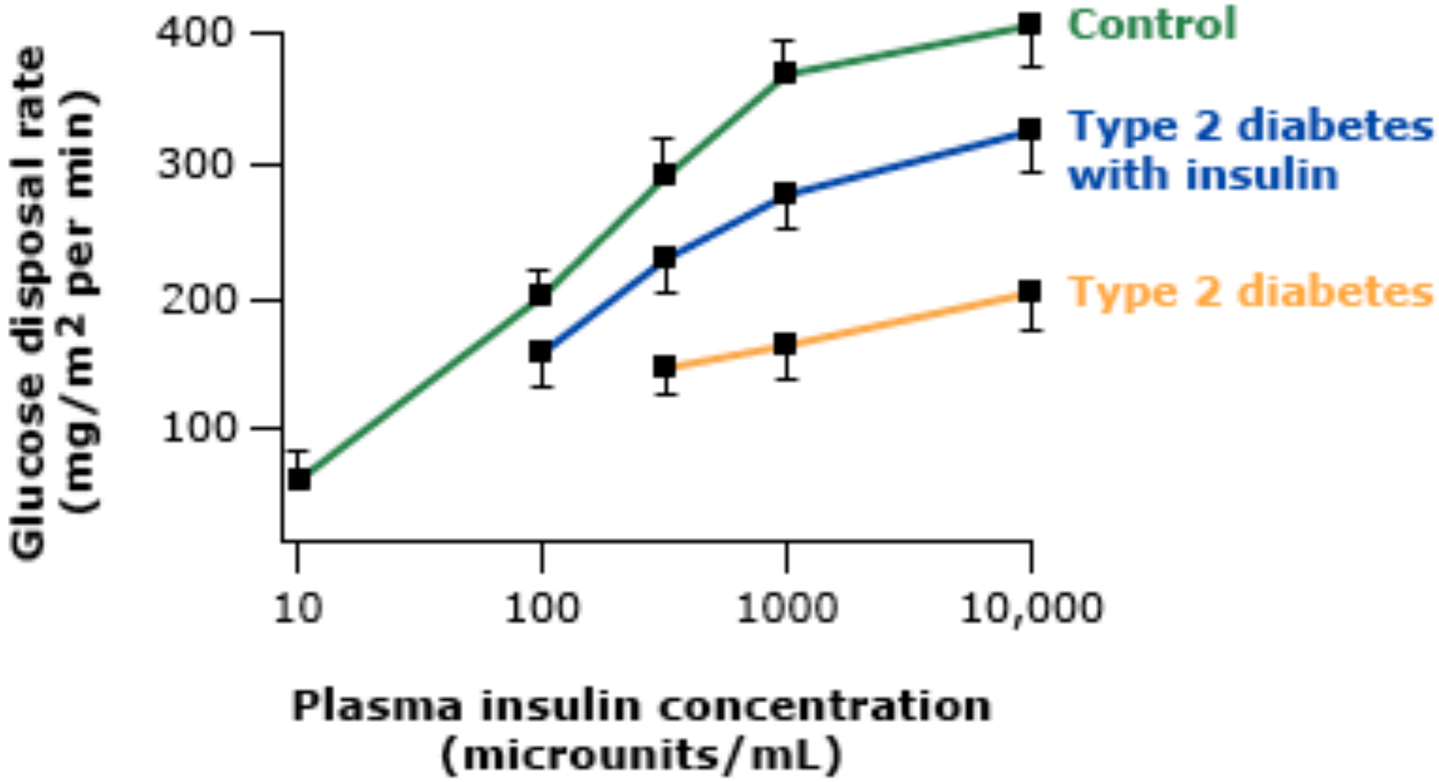


Primer: lipidi i insulinska rezistencija

Borkman et al. (1993)

- **Hipoteza**: lipidni sastav membrane skeletnih mišićih ćelija utiče na insulinsku rezistenciju.

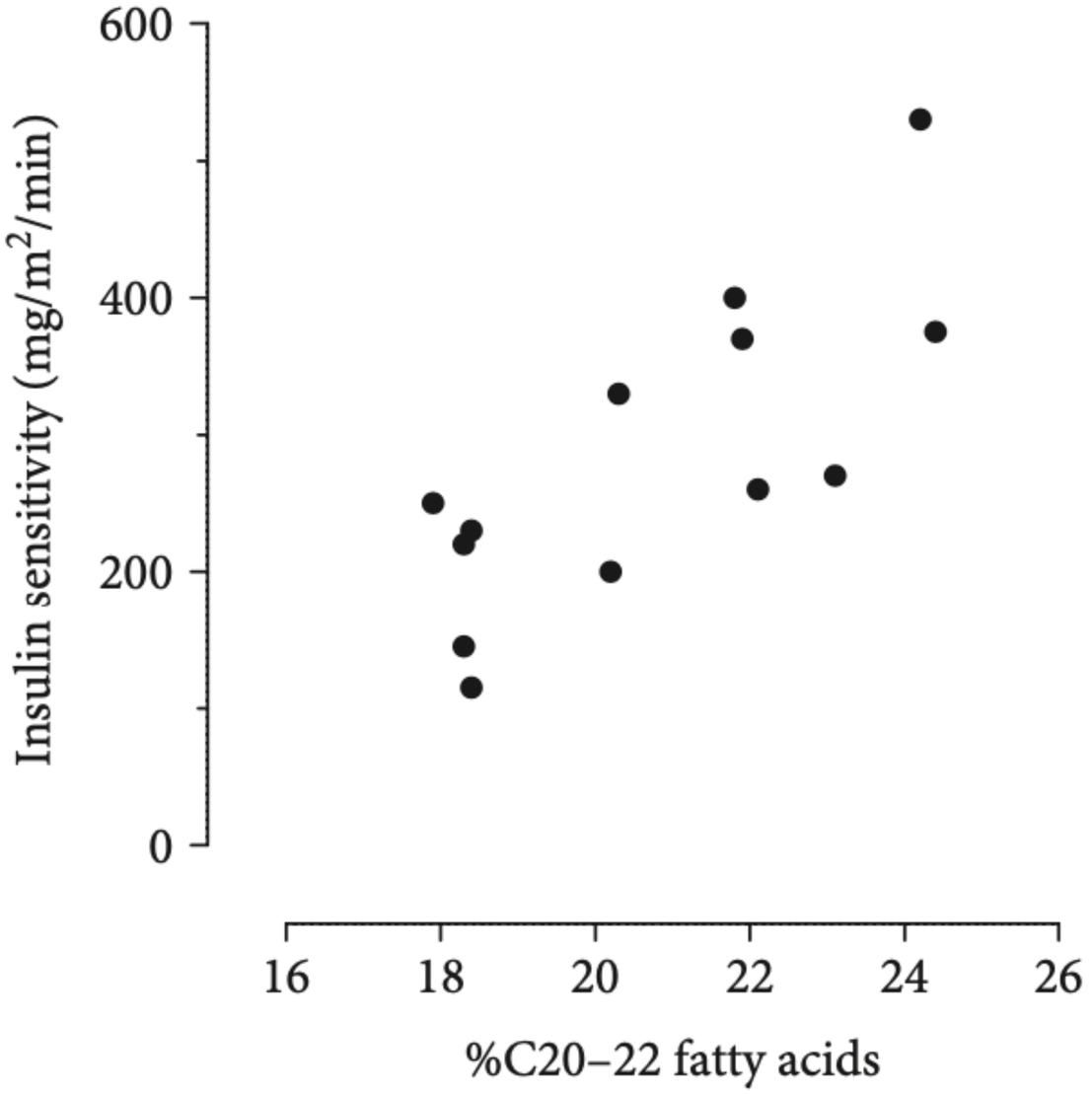




Primer: lipidi i insulinska rezistencija

Borkman et al. (1993)

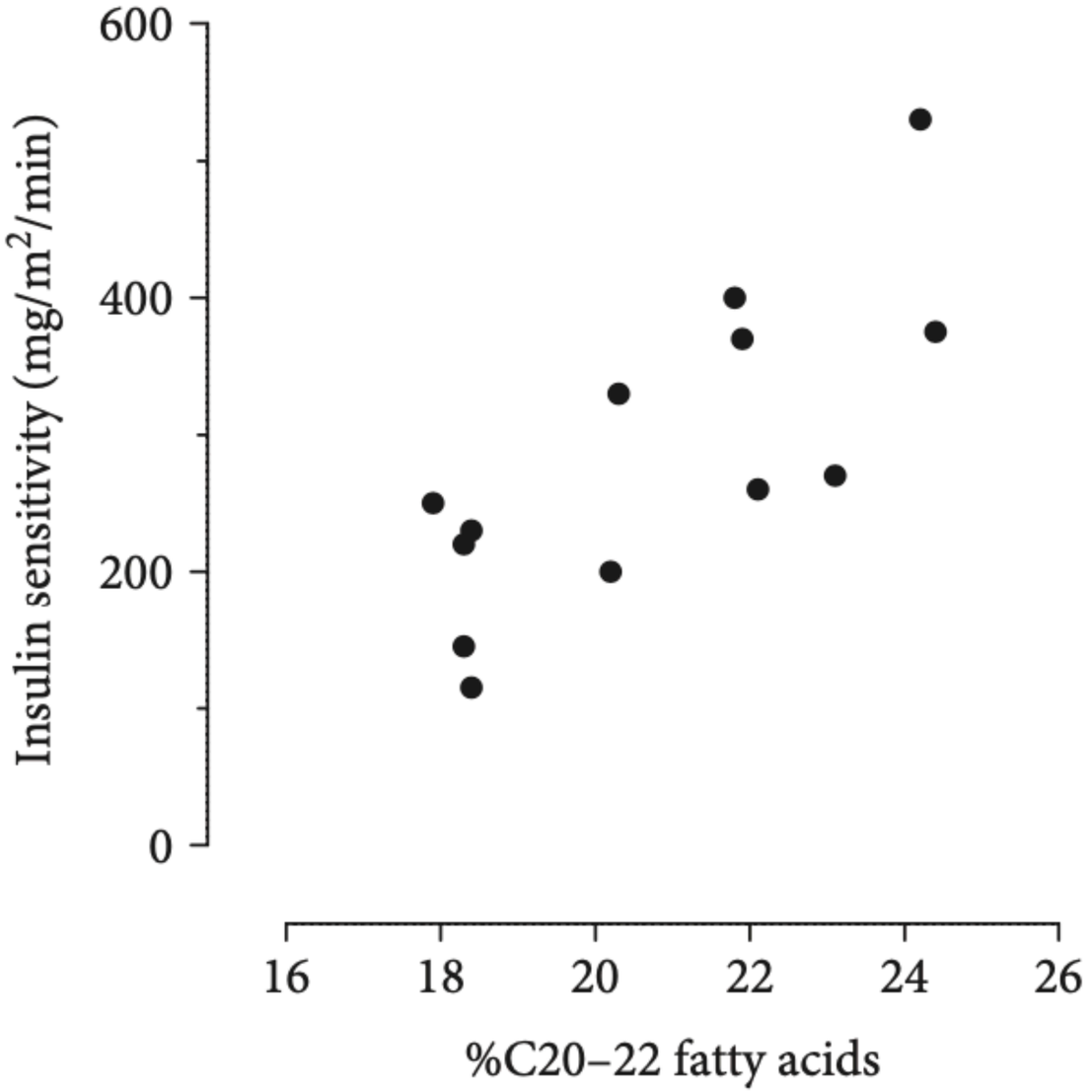
- Scatterplot (dijagram rasturanja)
- Kako da opišemo odnos ove dve varijable?

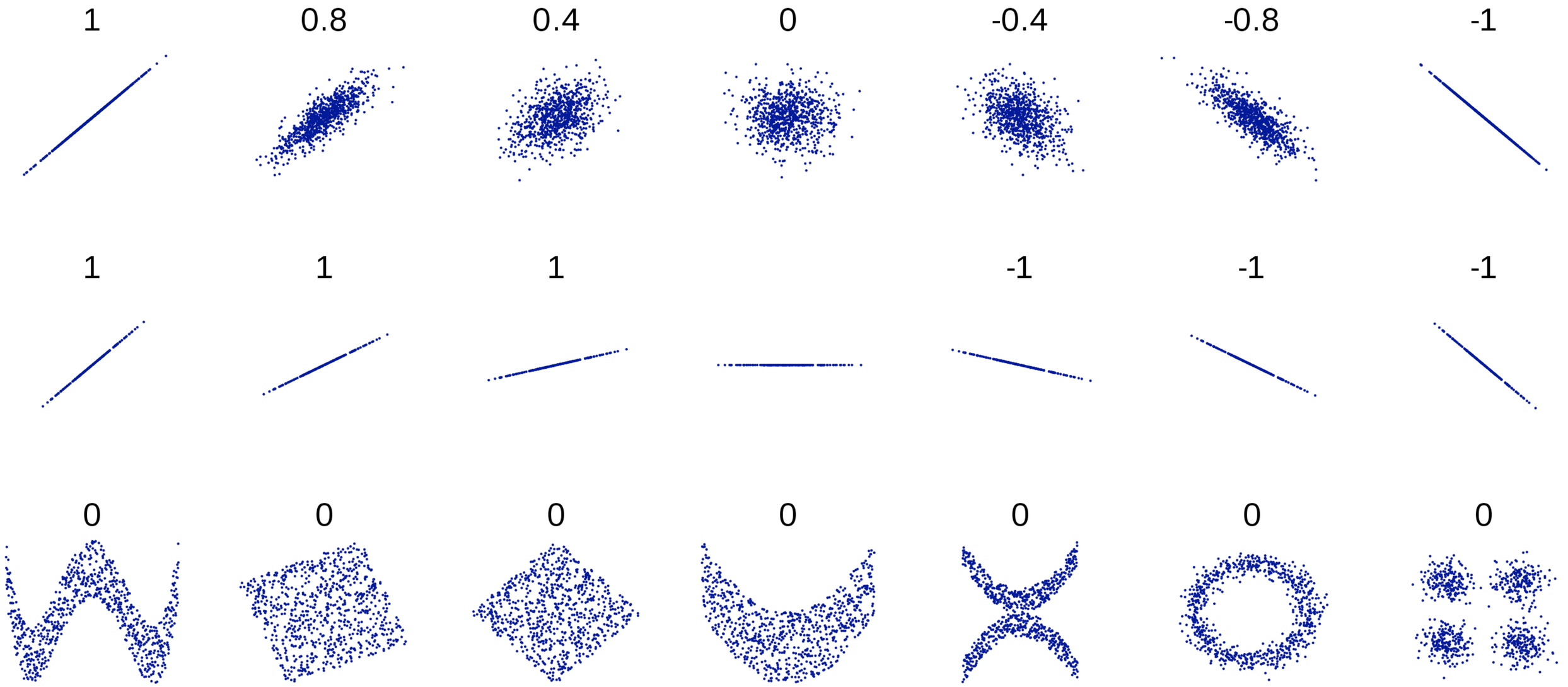


Primer: lipidi i insulinska rezistencija

Borkman et al. (1993)

- Scatterplot (dijagram rasturanja)
- Kako da opišemo odnos ove dve varijable?
- **Korelacija**




1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

Povezanost

1

-1

Može biti + ili -

Povezanost

0.4 0 -0.4

Može biti + ili -

Koleracija

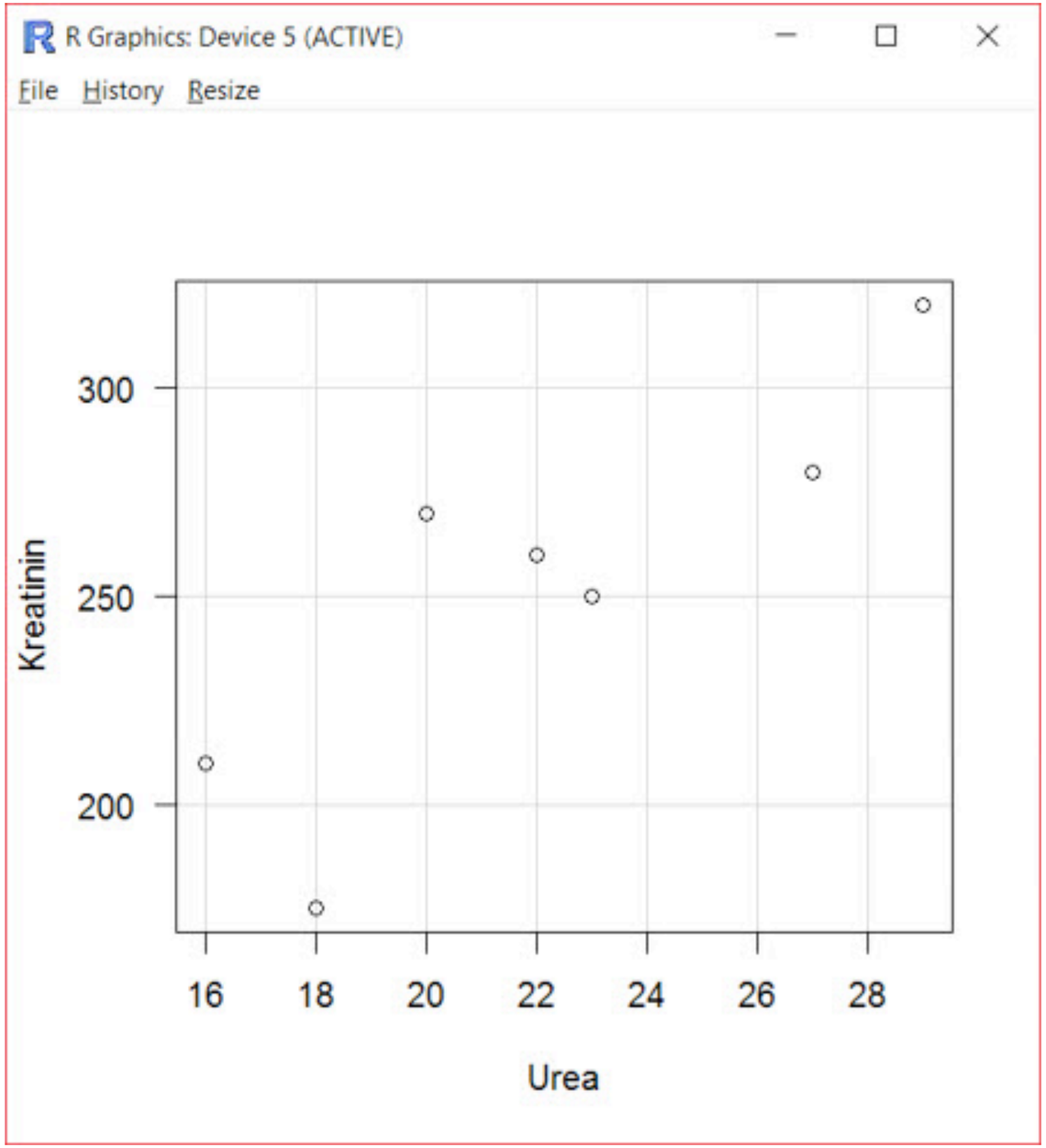
0 0 0

Varijable mogu biti povezane i nelinearno

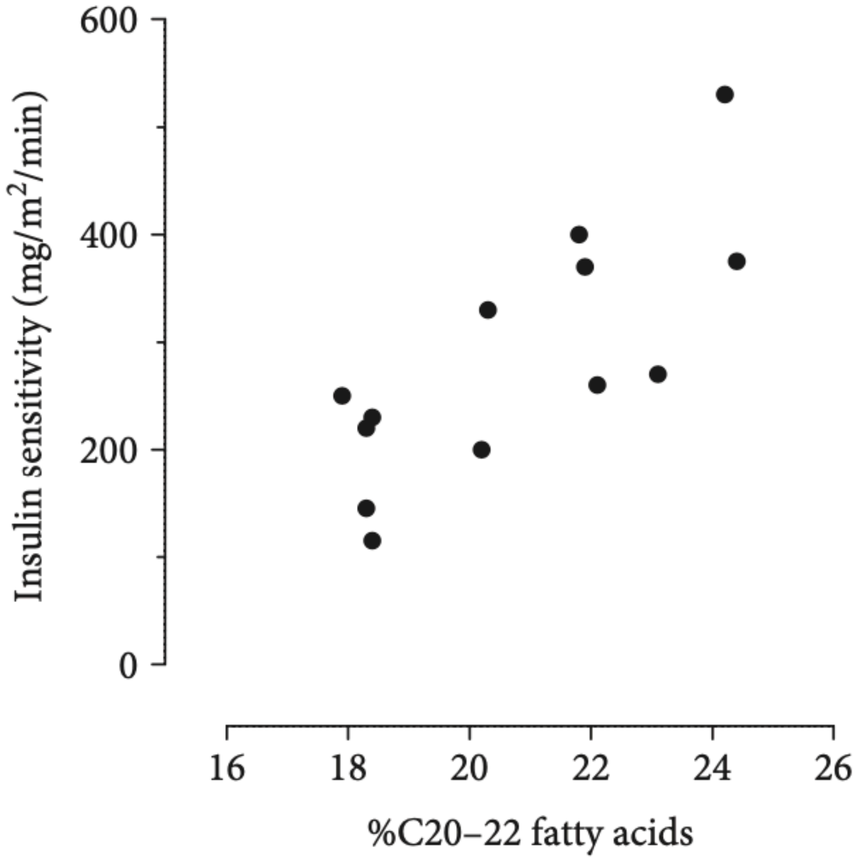
Povezanost

Rezime

- **Smer** (pozitivan **+**, negativan **-**)
- **Stepen** (od -1 do 1 = “jačina”)
- **Oblik** (Linearan, nelinearan)
- Scatterplot (dijagram rasturanja)



Napraviti dijagram rasturanja u kome će se vrednosti uree naneti na x osu, a vrednosti kreatinina na y osu.



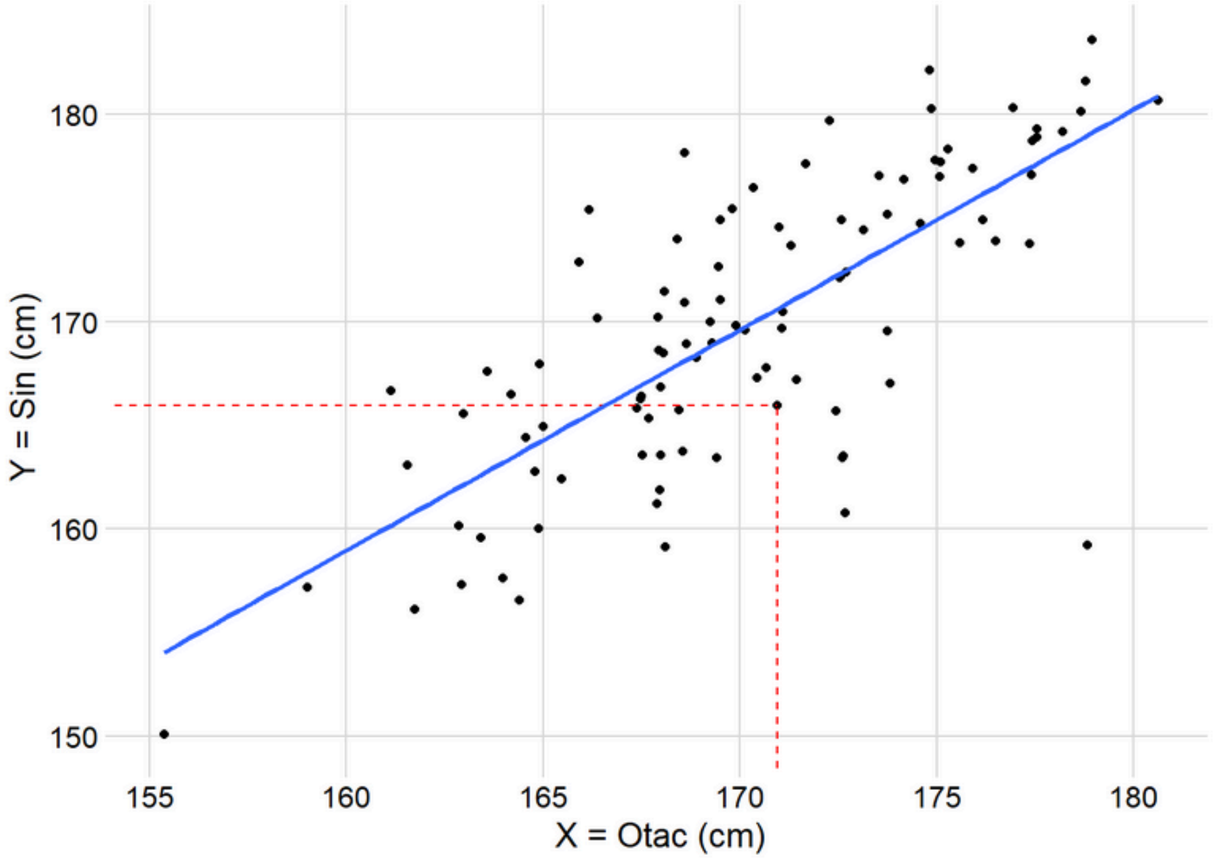
rb	Urea_mmol_L	Cr_umol_L
1	27	280
2	29	320
3	16	210
4	18	175
5	20	270
6	23	250
7	22	260

Korelacioni koeficijent

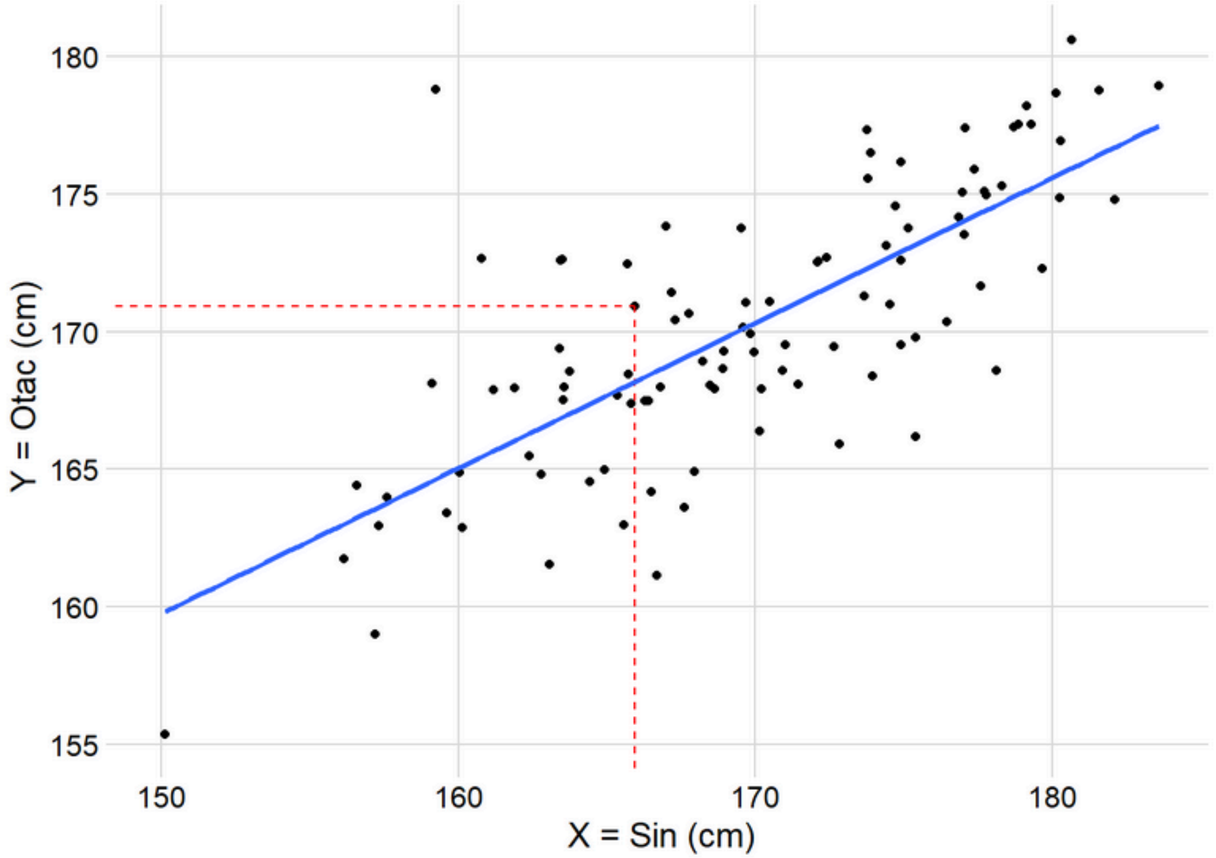
Koeficijent korelacije je statistika koja kvantifikuje jačinu povezanosti varijabli.

- Pearsonov koeficijent korelacije (r)
- Spearmanov koeficijent korelacije ranga (ρ)

Simetričnost korelacija



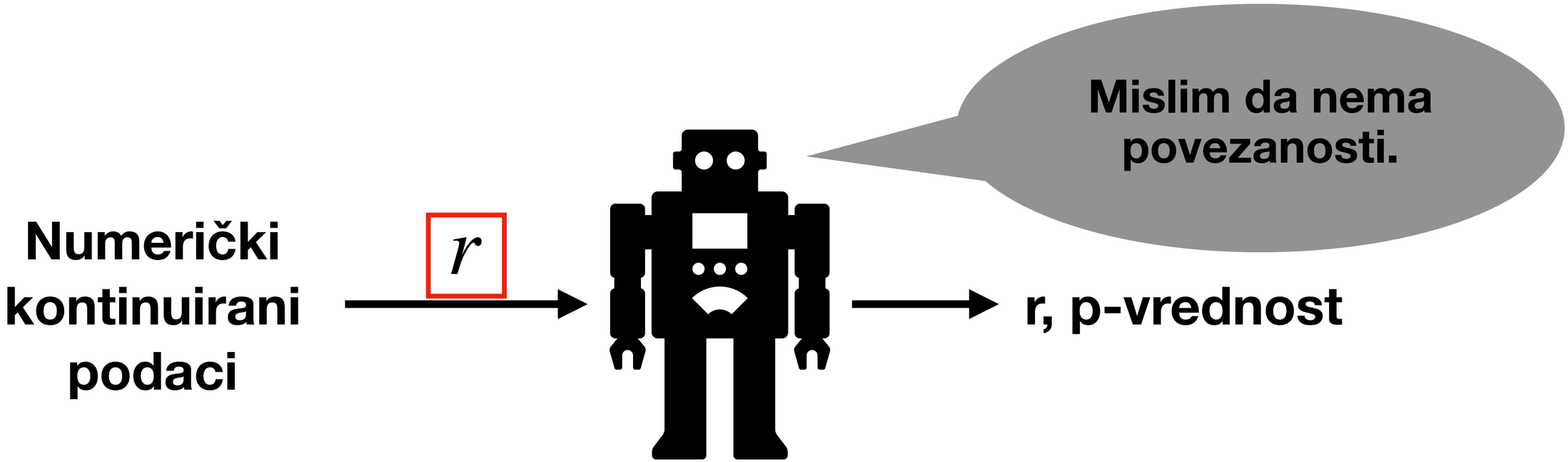
(a) Očevi predviđaju sinove



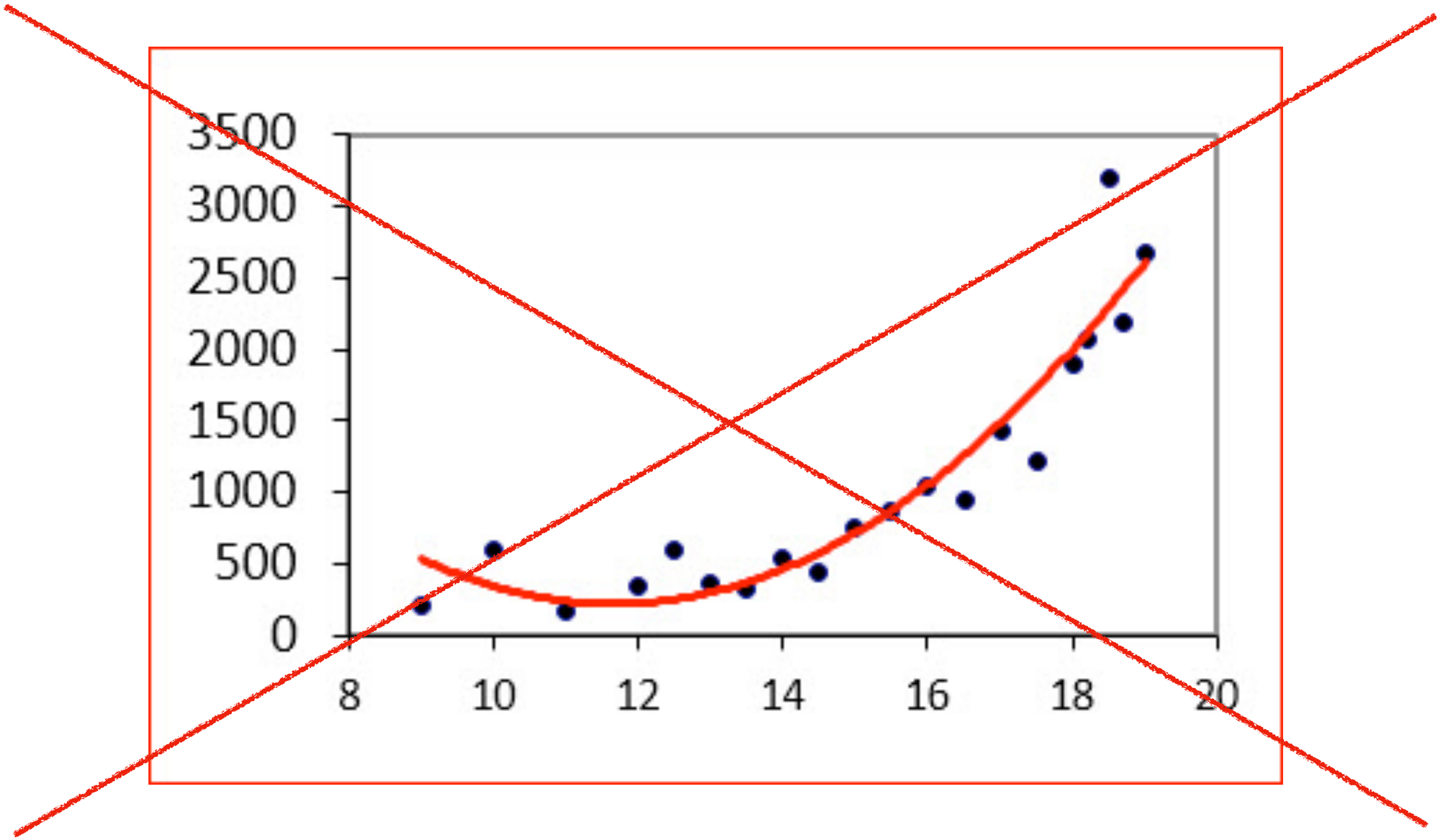
(b) Sinovi predviđaju očeve

$$V_d = \frac{V_{otac} + V_{majka} \pm 13}{2}\ cm$$

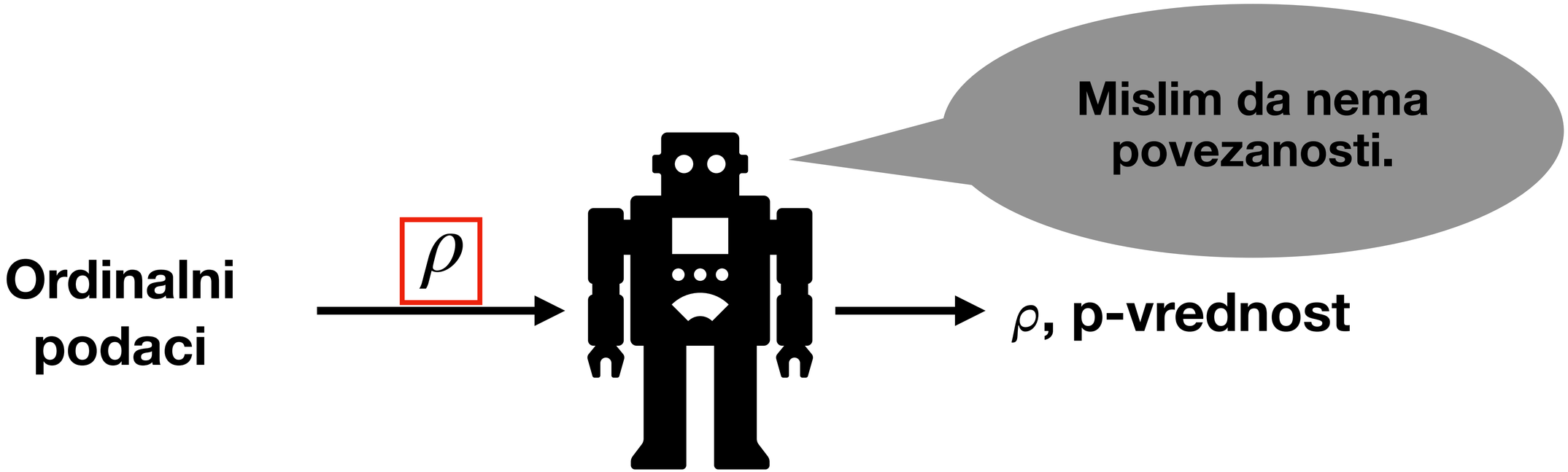
Pearsonov koeficijent korelacije



1. **Nezavisne opservacije**
2. **Normalna raspodela u populaciji**
3. **Svaka varijabla mora da ima svog para**
4. **Linearan odnos varijabli (scatterplot)**


3500
3000
2500
2000
1500
1000
500
0
8
10
12
14
16
18
20

Spearmanov koeficijent korelacije rangova



1. **Varijabla nema normalnu raspodelu**
2. **Podaci su ordinalni ili rangovi**
3. **Svaka varijabla mora da ima svog para**
4. **Nelinearan odnos varijabli (scatterplot)**
5. **Monotoničnost***

Interpretacija koeficijenta korelacije

Koeficijent	Jačina povezanosti (Interpretacija jednako važi i za vrednosti sa negativnim predznakom)
0.70 – 1.00	Jaka povezanost
0.30 – 0.69	Osrednja povezanost
< 0.30	Slaba povezanost
oko 0.00	Nema povezanosti

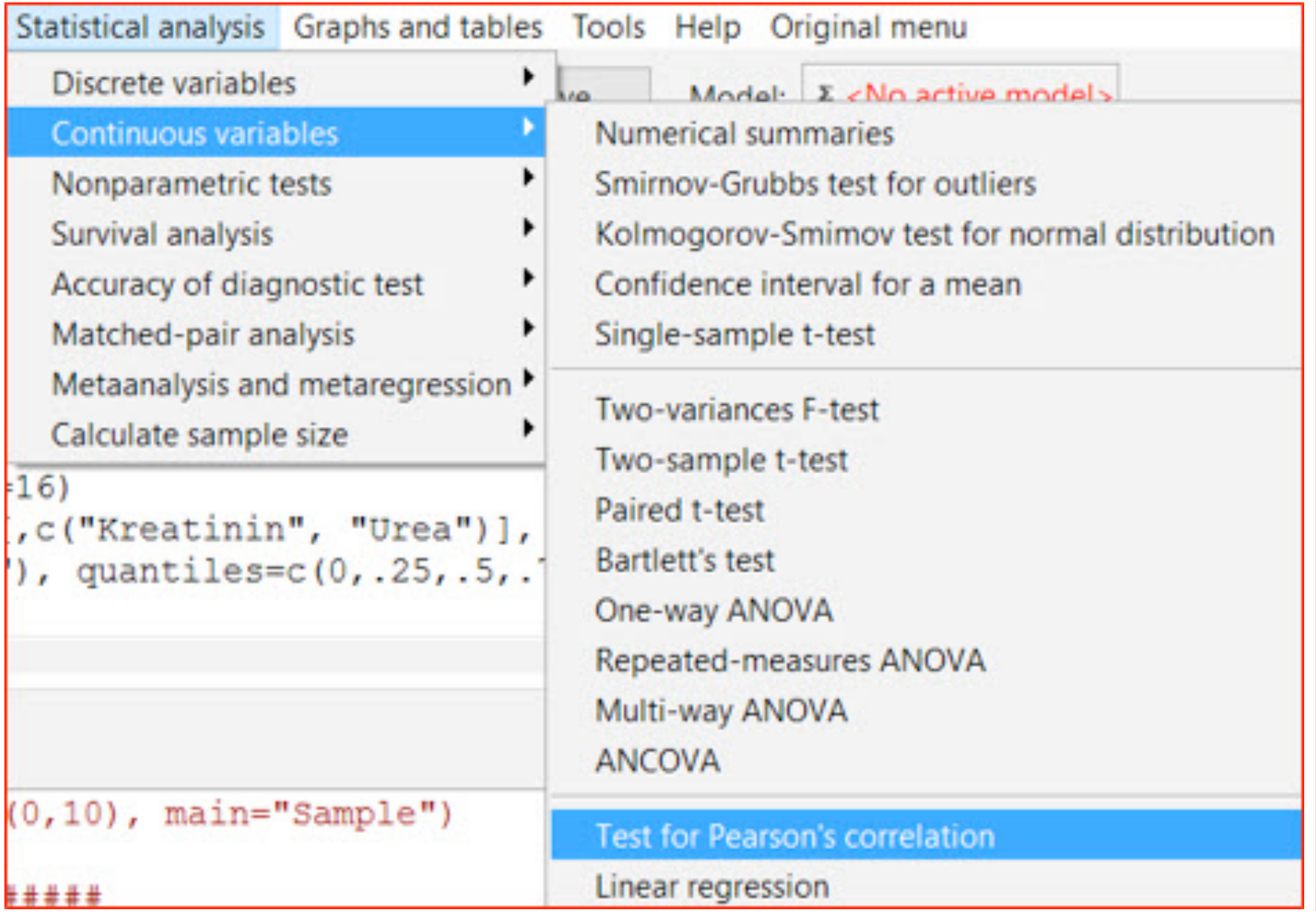
Ako je $p \leq 0.05$? Ako je $p > 0.05$?

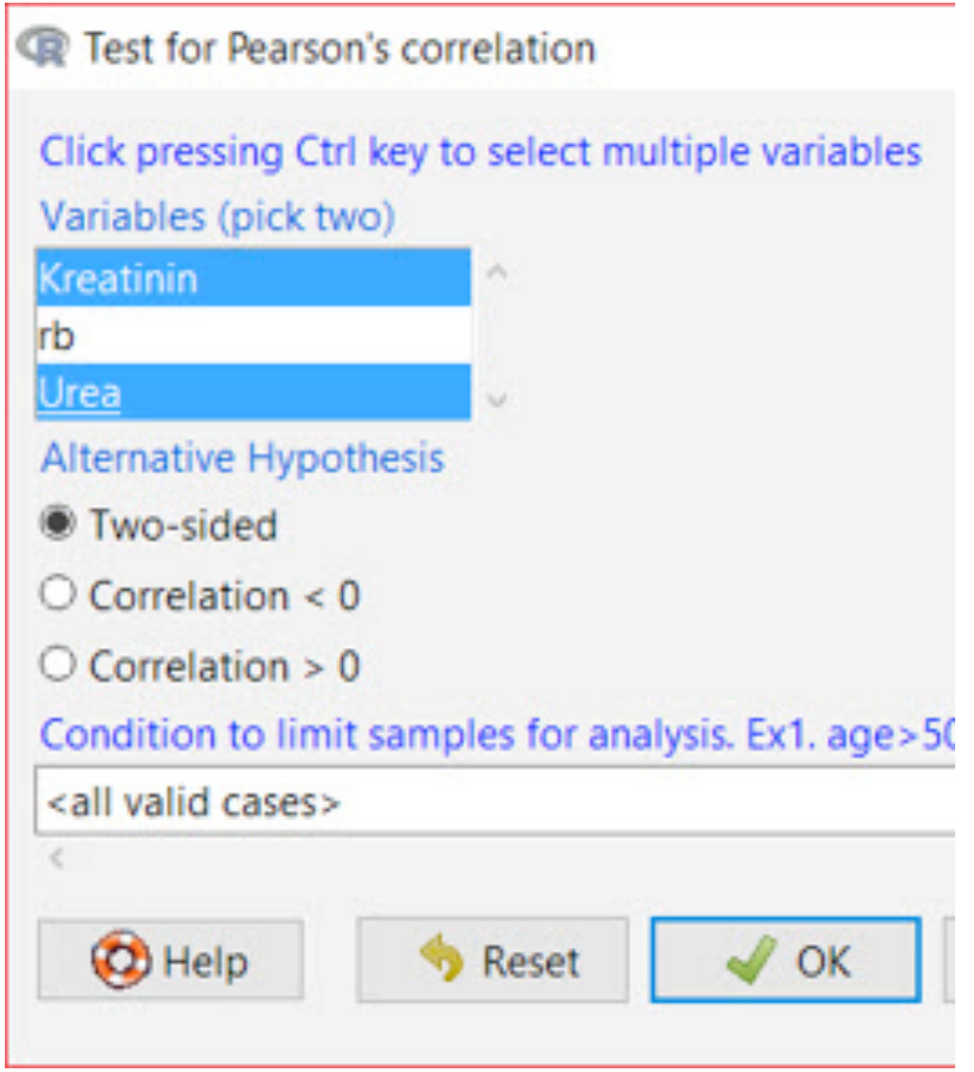
Odrediti koeficijent korelacije između uree i kreatinina.

Odrediti koeficijent korelacije između uree i kreatinina.

1. **Proveriti normalnost (nummerical summaries)**
2. **Pearson / Spearman ?**

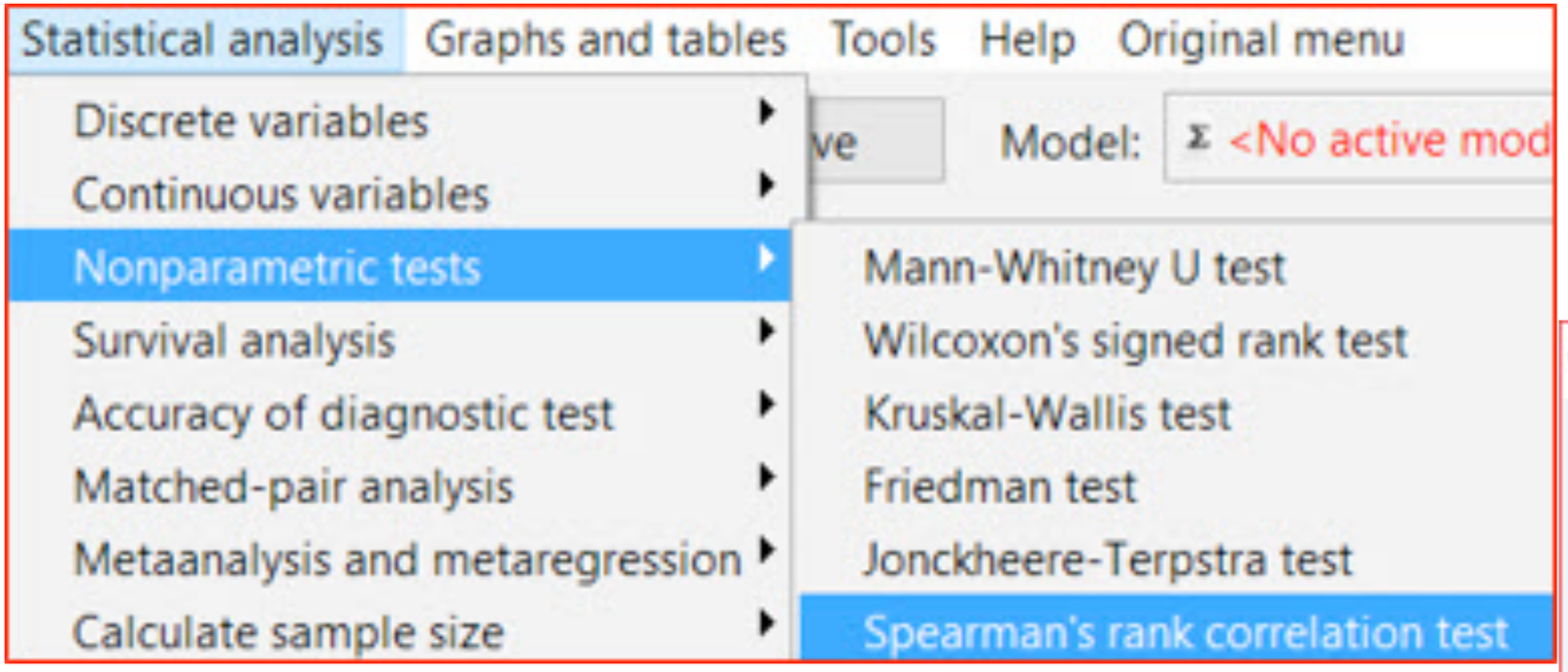
```
> numSummary(Dataset[,c("Kreatinin", "Urea")], statistics=c("mean", "sd",
+   "quantiles", "cv"), quantiles=c(0,.25,.5,.75,1))
                mean         sd        cv  0% 25% 50% 75% 100% n
Kreatinin 252.14286 47.421615 0.1880744 175 230 260 275  320 7
Urea       22.14286  4.670067 0.2109062  16  19  22  25   29 7
```

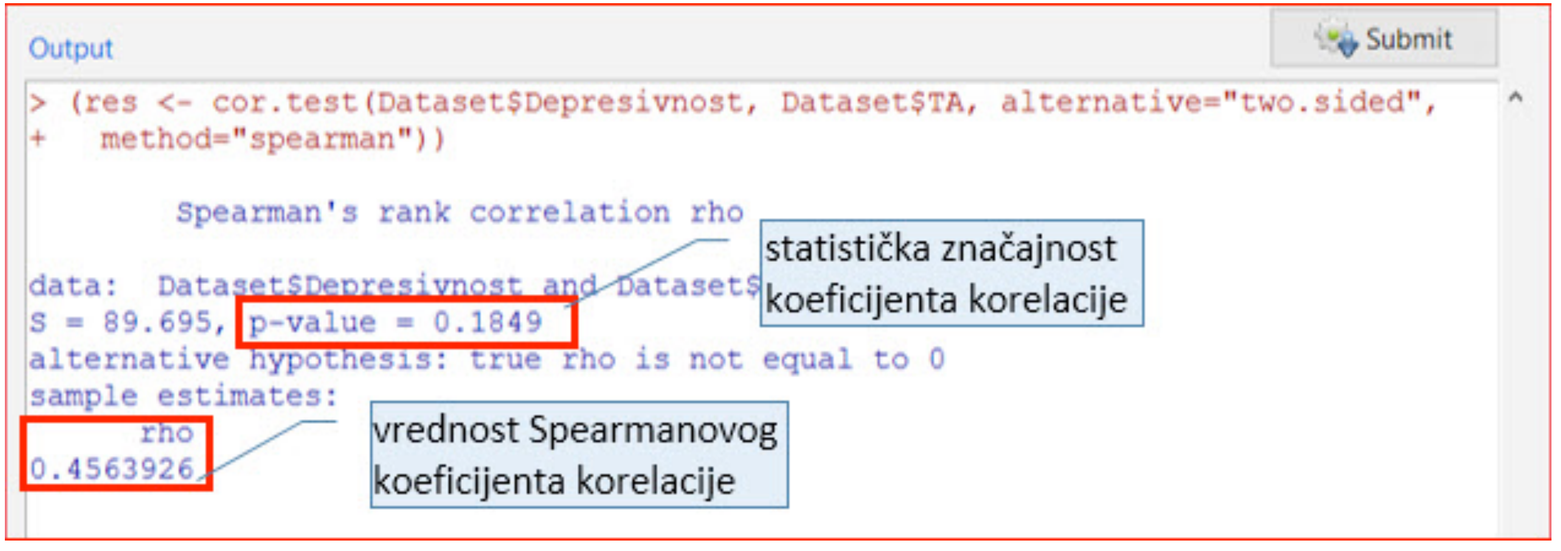

Statistical analysis
Graphs and tables
Tools
Help
Original menu
Discrete variables
Continuous variables
Nonparametric tests
Survival analysis
Accuracy of diagnostic test
Matched-pair analysis
Metaanalysis and metaregression
Calculate sample size
Model:
<No active model>
Numerical summaries
Smirnov-Grubbs test for outliers
Kolmogorov-Smimov test for normal distribution
Confidence interval for a mean
Single-sample t-test
Two-variances F-test
Two-sample t-test
Paired t-test
Bartlett's test
One-way ANOVA
Repeated-measures ANOVA
Multi-way ANOVA
ANCOVA
Test for Pearson's correlation
Linear regression
=16)
,c("Kreatinin", "Urea")],
'), quantiles=c(0,.25,.5,.
(0,10), main="Sample")
#####

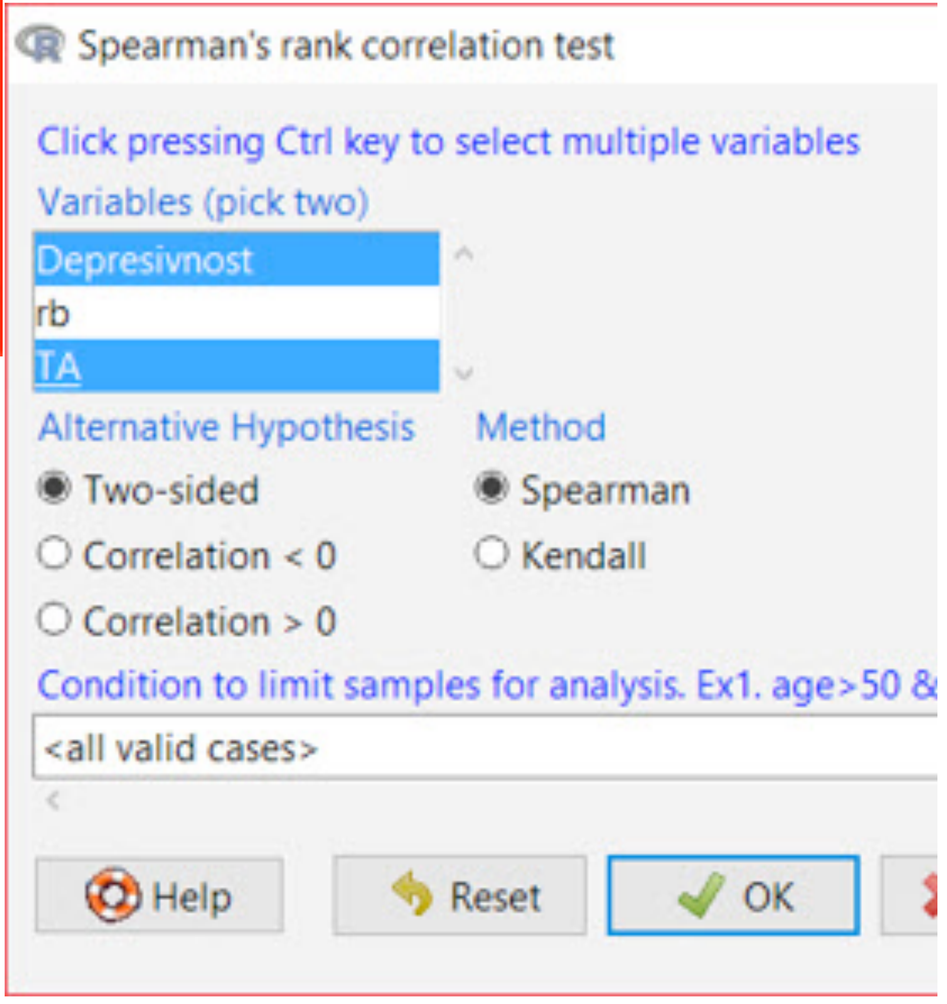

Test for Pearson's correlation
Click pressing Ctrl key to select multiple variables
Variables (pick two)
Kreatinin
rb
Urea
Alternative Hypothesis
Two-sided
Correlation < 0
Correlation > 0
Condition to limit samples for analysis. Ex1. age>50
<all valid cases>
Help
Reset
OK

Da li postoji povezanost depresije i sistolne arterijske tenzije? Testirati za nivo značajnosti 0.05.

rb	skor	ta
1	23	139
2	19	109
3	26	113
4	23	128
5	19	124
6	17	105
7	23	116
8	26	135
9	20	120
10	19	124


Statistical analysis
Graphs and tables
Tools
Help
Original menu
Discrete variables
Continuous variables
Nonparametric tests
Survival analysis
Accuracy of diagnostic test
Matched-pair analysis
Metaanalysis and metaregression
Calculate sample size
Model:
<No active mod
Mann-Whitney U test
Wilcoxon's signed rank test
Kruskal-Wallis test
Friedman test
Jonckheere-Terpstra test
Spearman's rank correlation test


Output
Submit
> (res <- cor.test(Dataset$Depresivnost, Dataset$TA, alternative="two.sided",
+ method="spearman"))
Spearman's rank correlation rho
data: Dataset$Depresivnost and Dataset$
S = 89.695, p-value = 0.1849
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.4563926
statistička značajnost koeficijenta korelacije
vrednost Spearmanovog koeficijenta korelacije


Spearman's rank correlation test
Click pressing Ctrl key to select multiple variables
Variables (pick two)
Depresivnost
rb
TA
Alternative Hypothesis
Two-sided
Correlation < 0
Correlation > 0
Method
Spearman
Kendall
Condition to limit samples for analysis. Ex1. age>50 &
<all valid cases>
Help
Reset
OK

Preuzimanjem baza DZ odgovori na sledeća istraživačka pitanja:

1. Da li postoji povezanost između starosti i ukupnog holesterola?

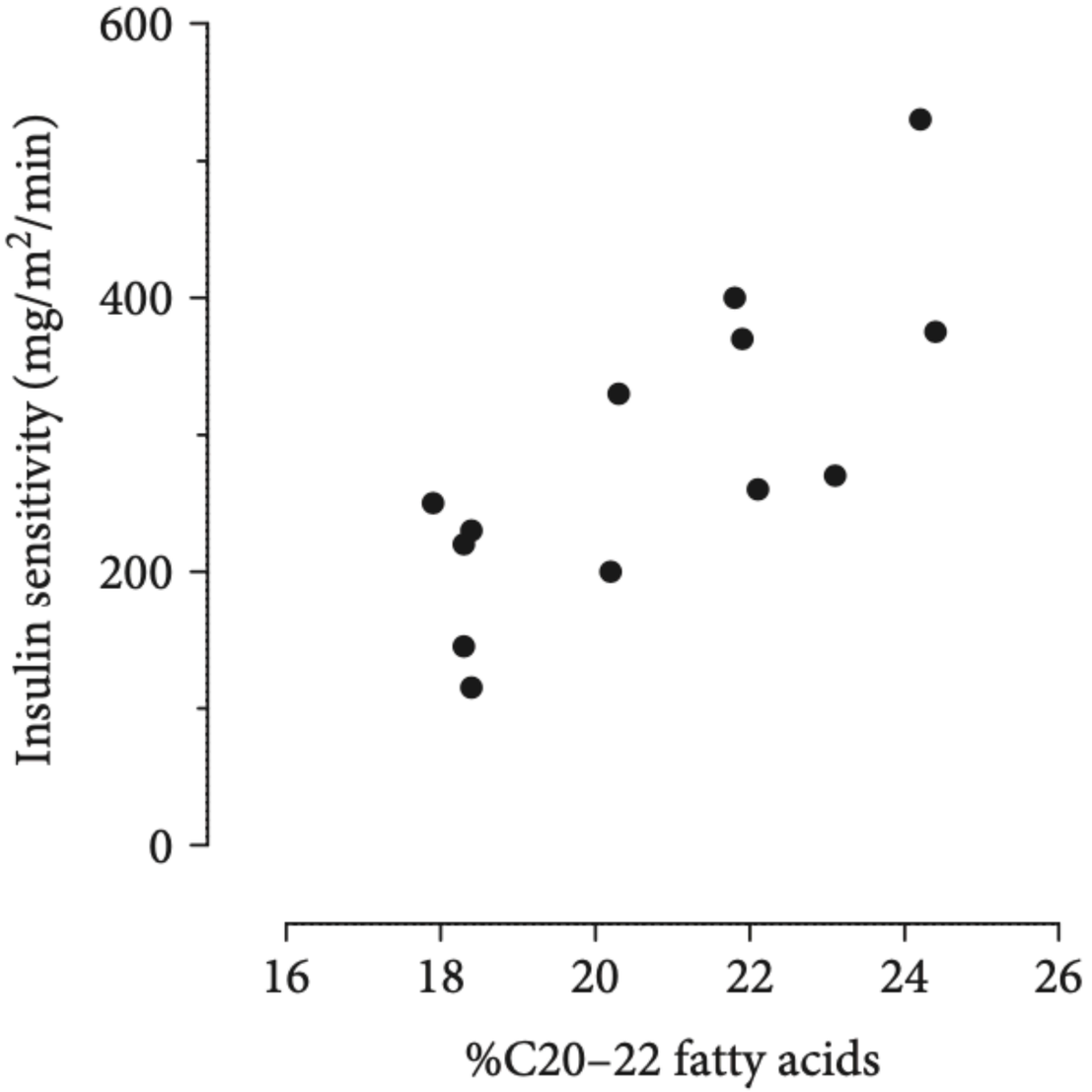
2. Da li postoji povezanost između triglicerida i stepena uhranjenosti?

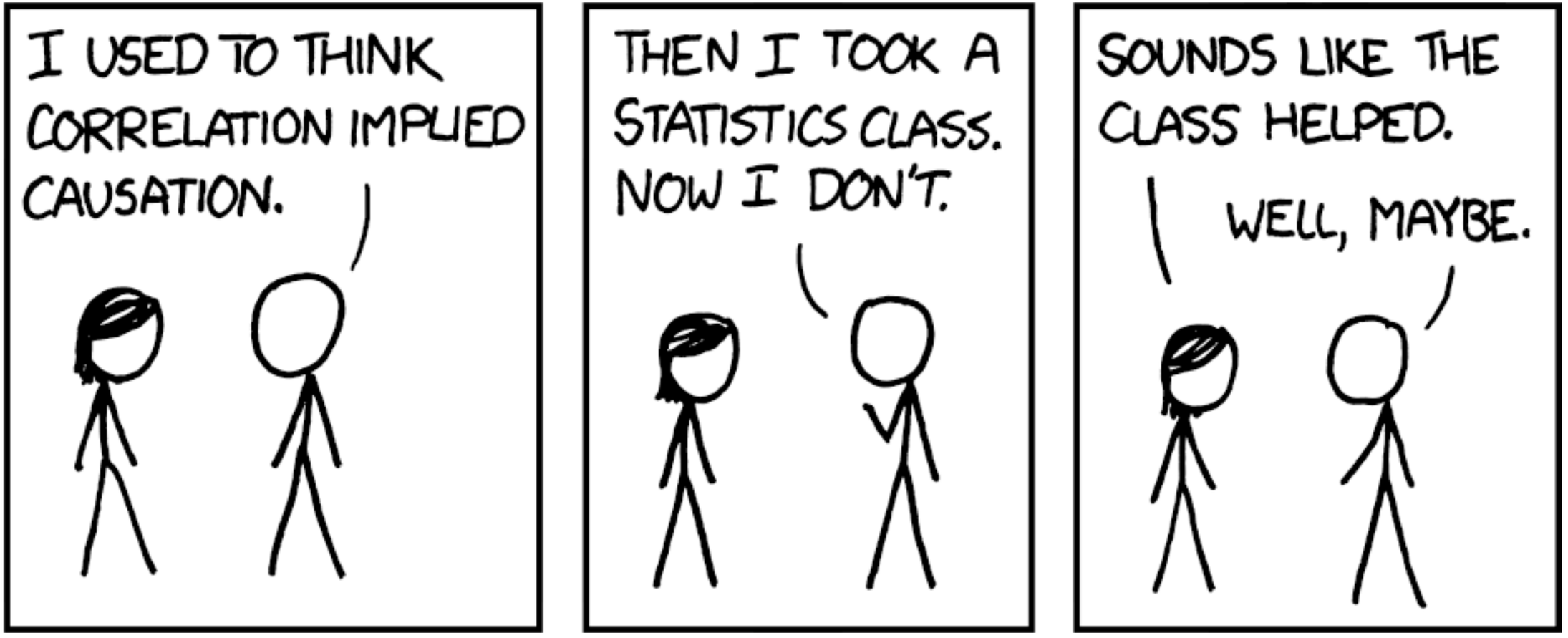
Primer: lipidi i insulinska rezistencija (2)

Borkman et al. (1993)

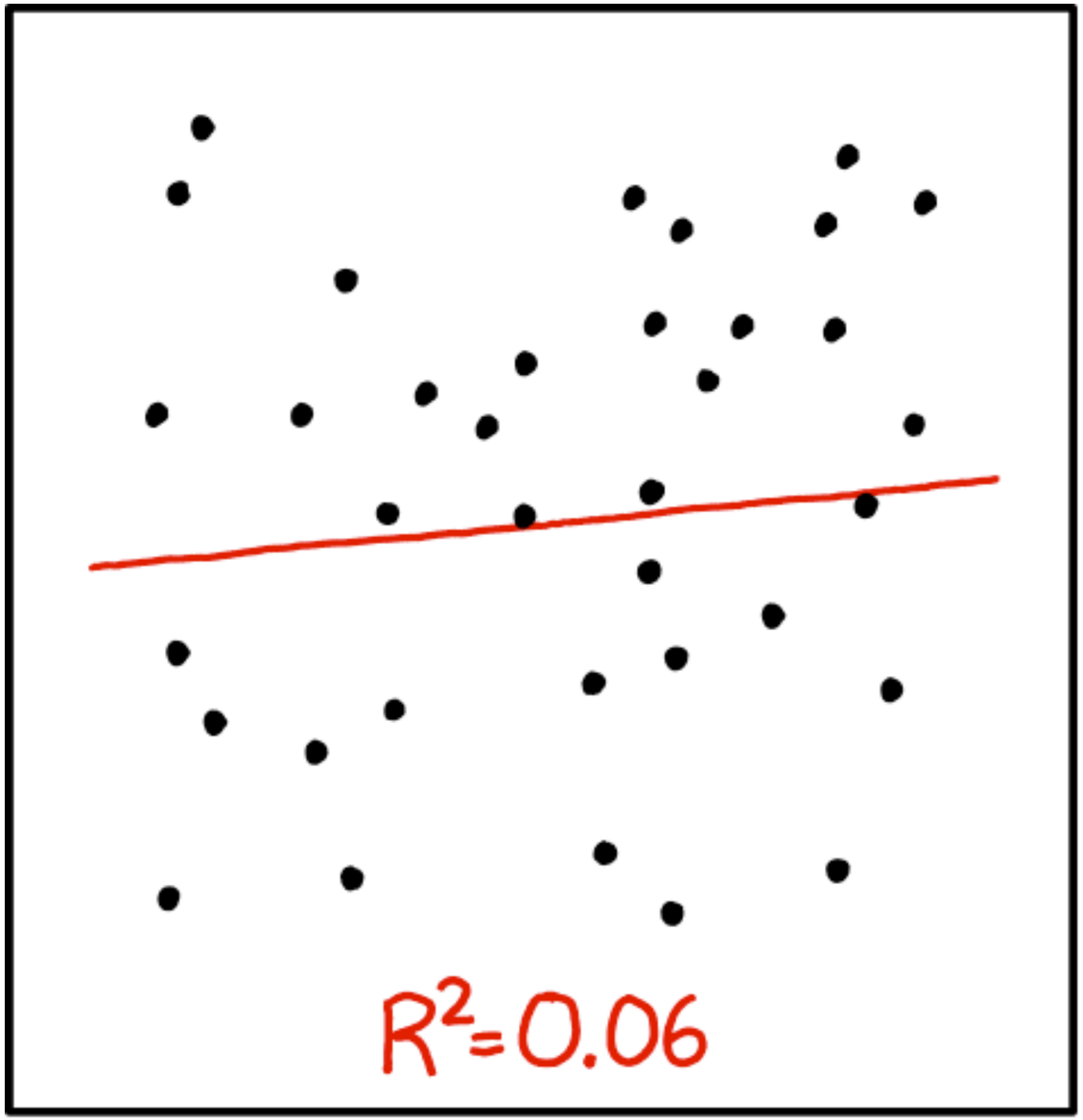
CORRELATION	
r	0.77
95% CI	0.3803 to 0.9275
r^2	0.5929
P (two-tailed)	0.0021
Number of XY pairs	13

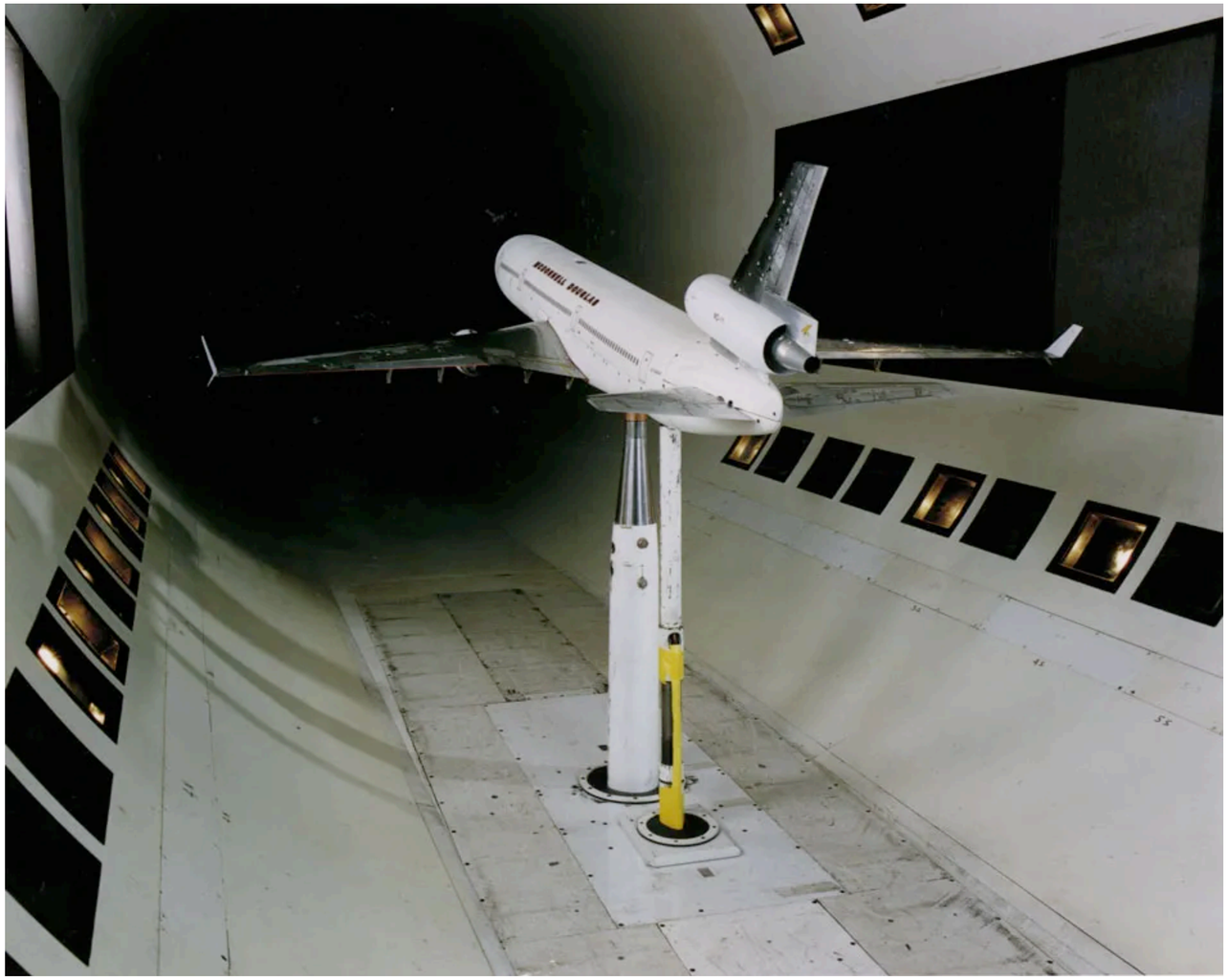
Opisali smo povezanost




I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Regresija



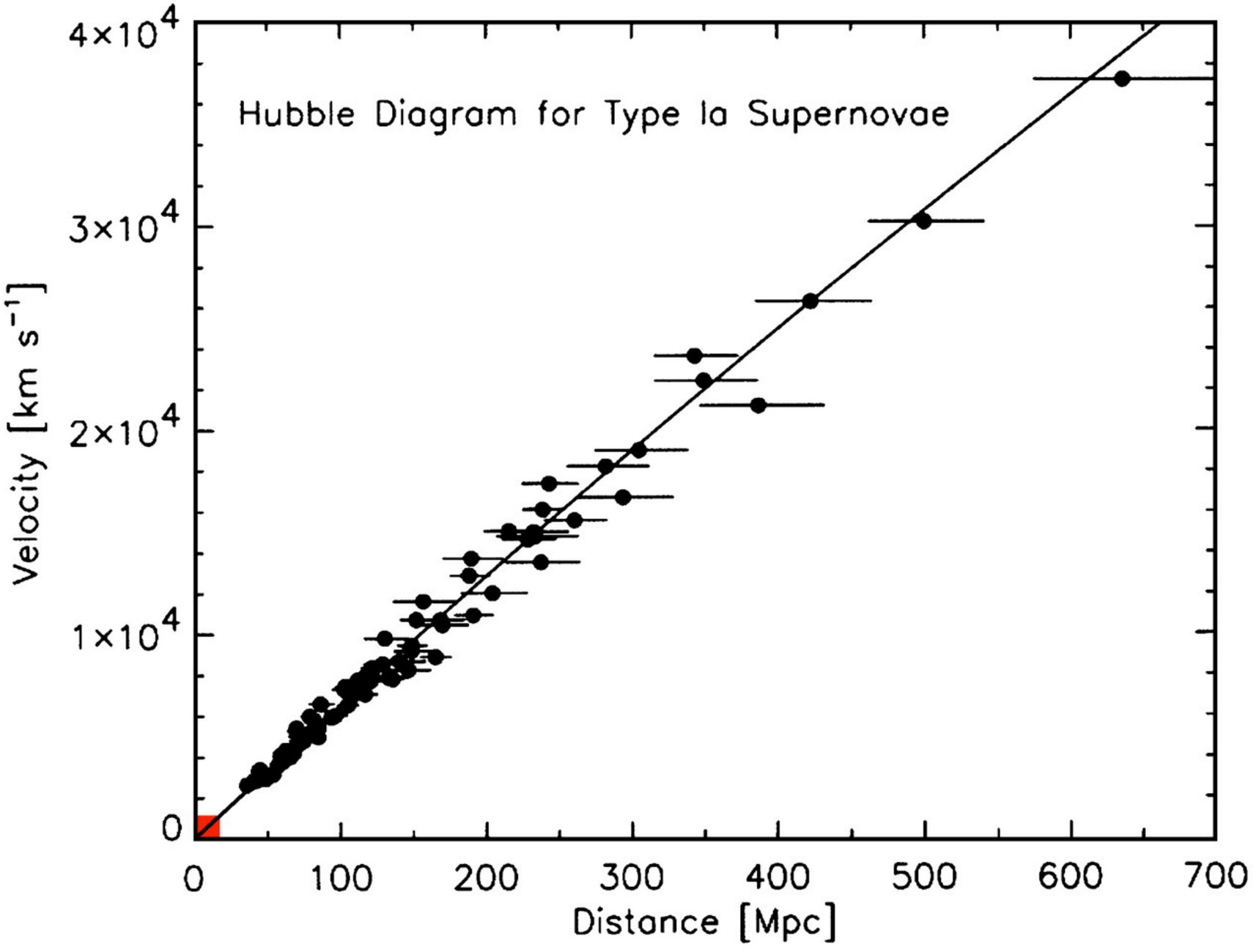


Matematički model

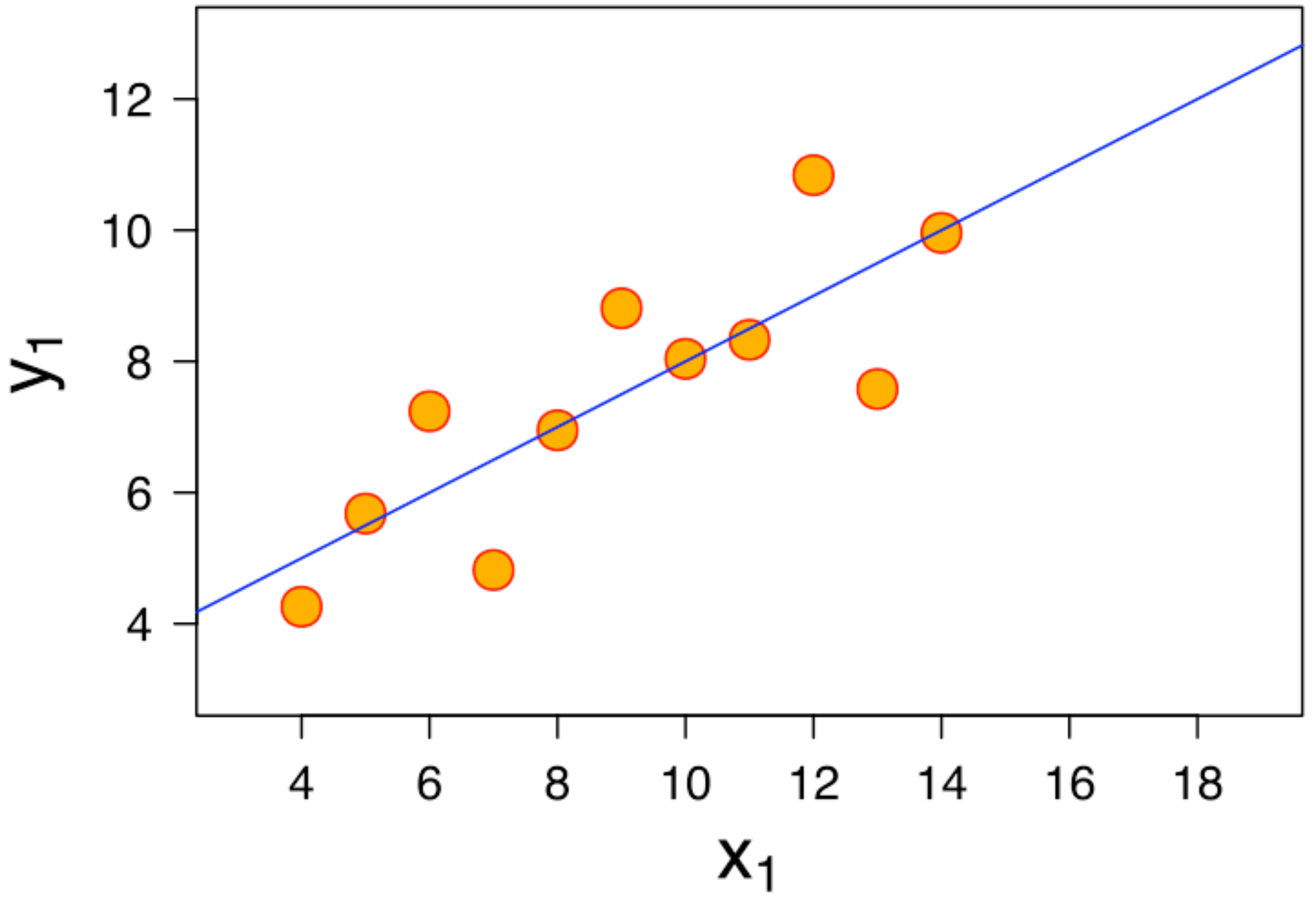
$$y = \alpha + \beta x$$

Stvarnost

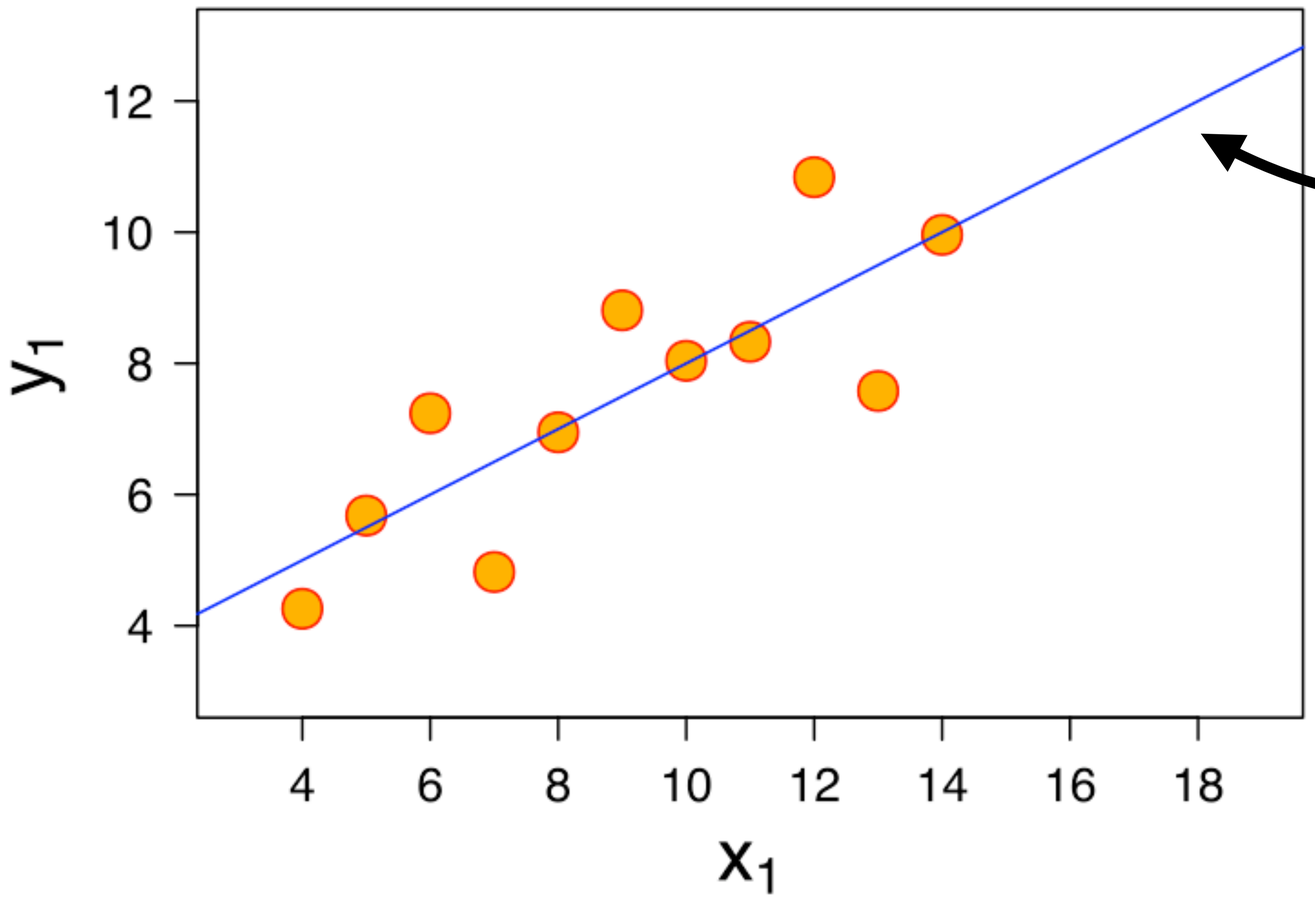
$$y = \text{komplikovano...}$$


Hubble Diagram for Type Ia Supernovae
Velocity [km s^{-1}]
4×10^4
3×10^4
2×10^4
1×10^4
0
0
100
200
300
400
500
600
700
Distance [Mpc]

Anatomija regresione jednačine



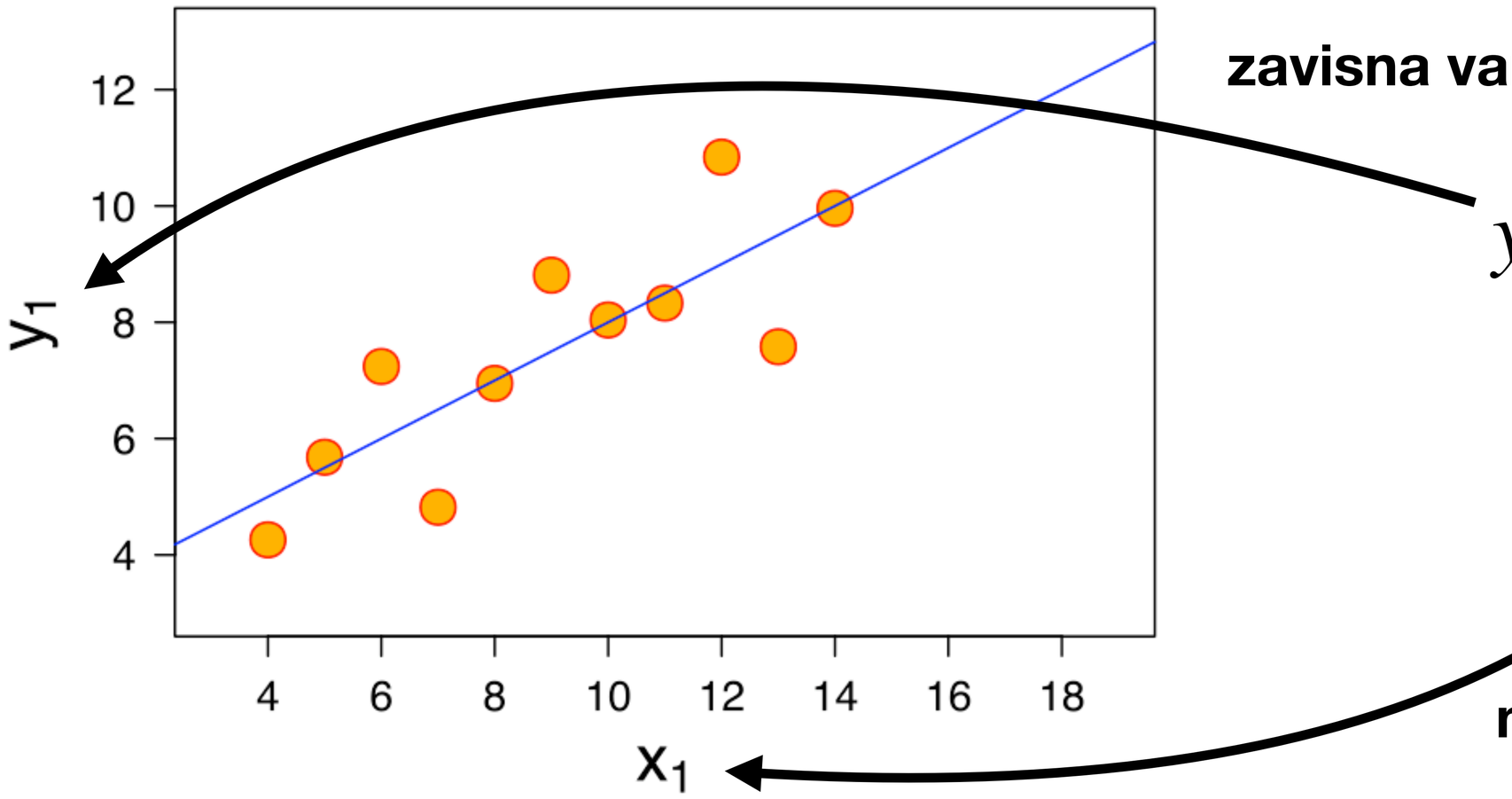
Anatomija regresione jednačine



Linearan odnos

$$y = \alpha + \beta x$$

Anatomija regresione jednačine

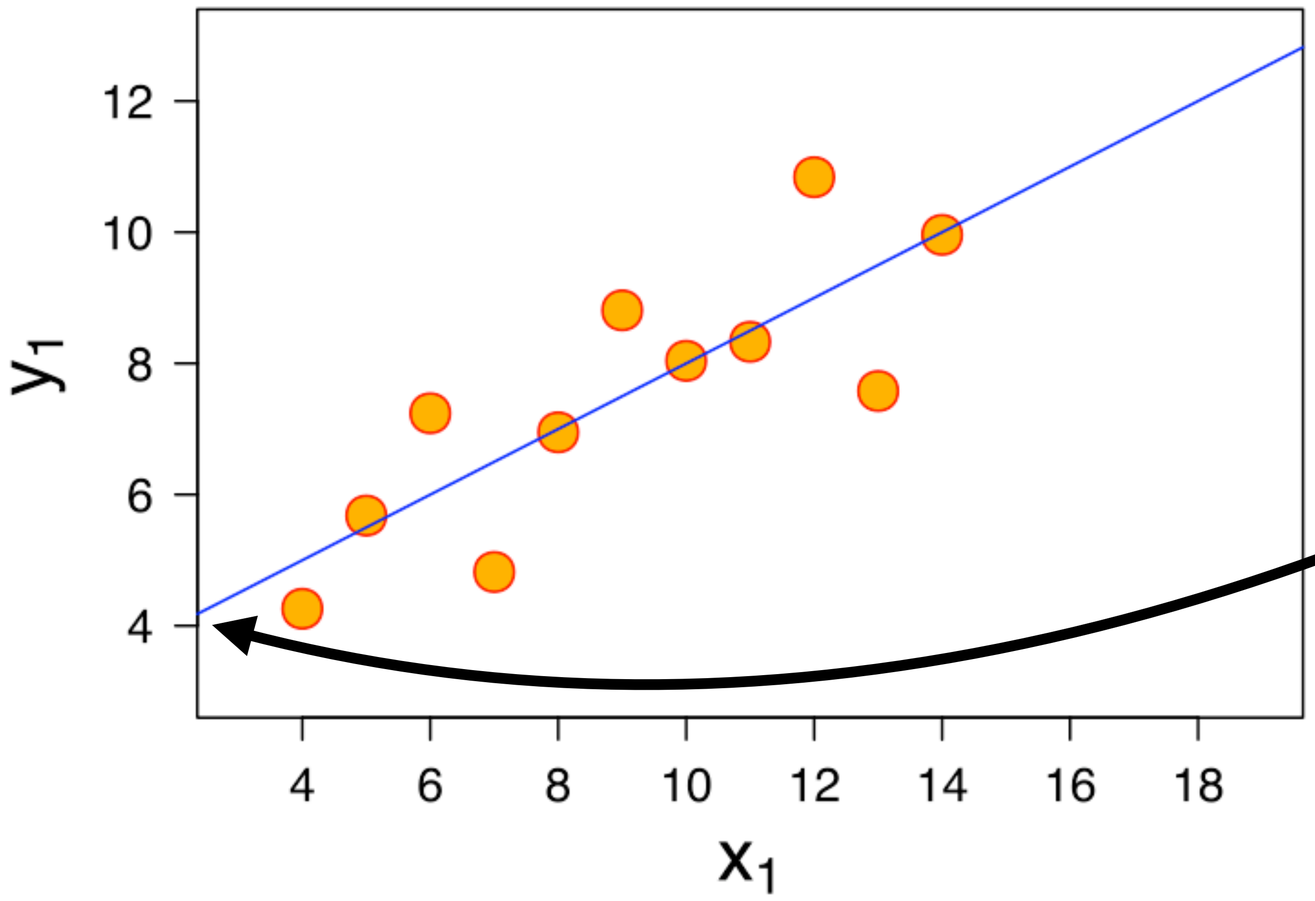


zavisna varijabla

$$y = \alpha + \beta x$$

nezavisna varijabla

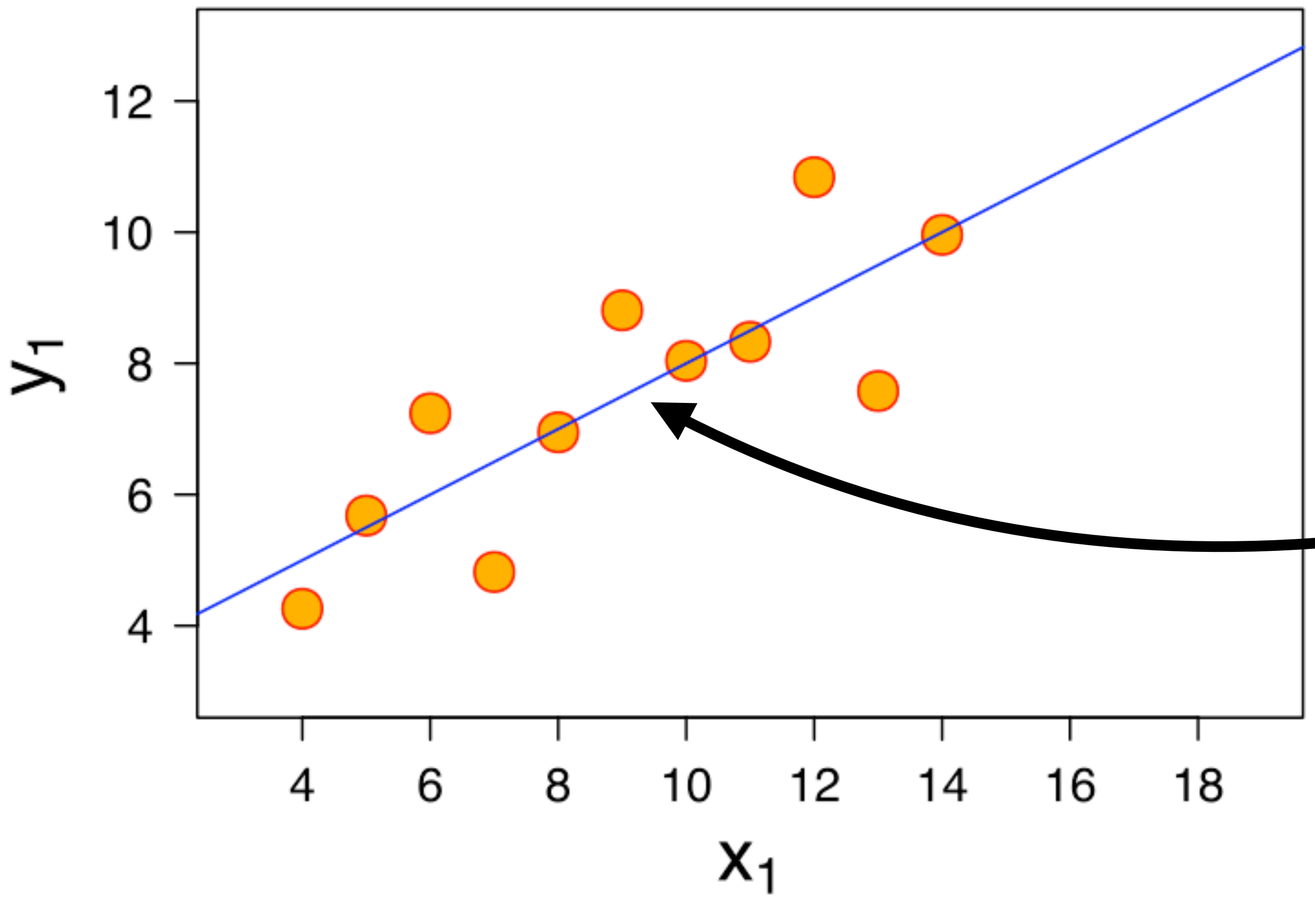
Anatomija regresione jednačine



regresiona konstanta a:
kad je x = 0

$$y = \alpha + \beta x$$

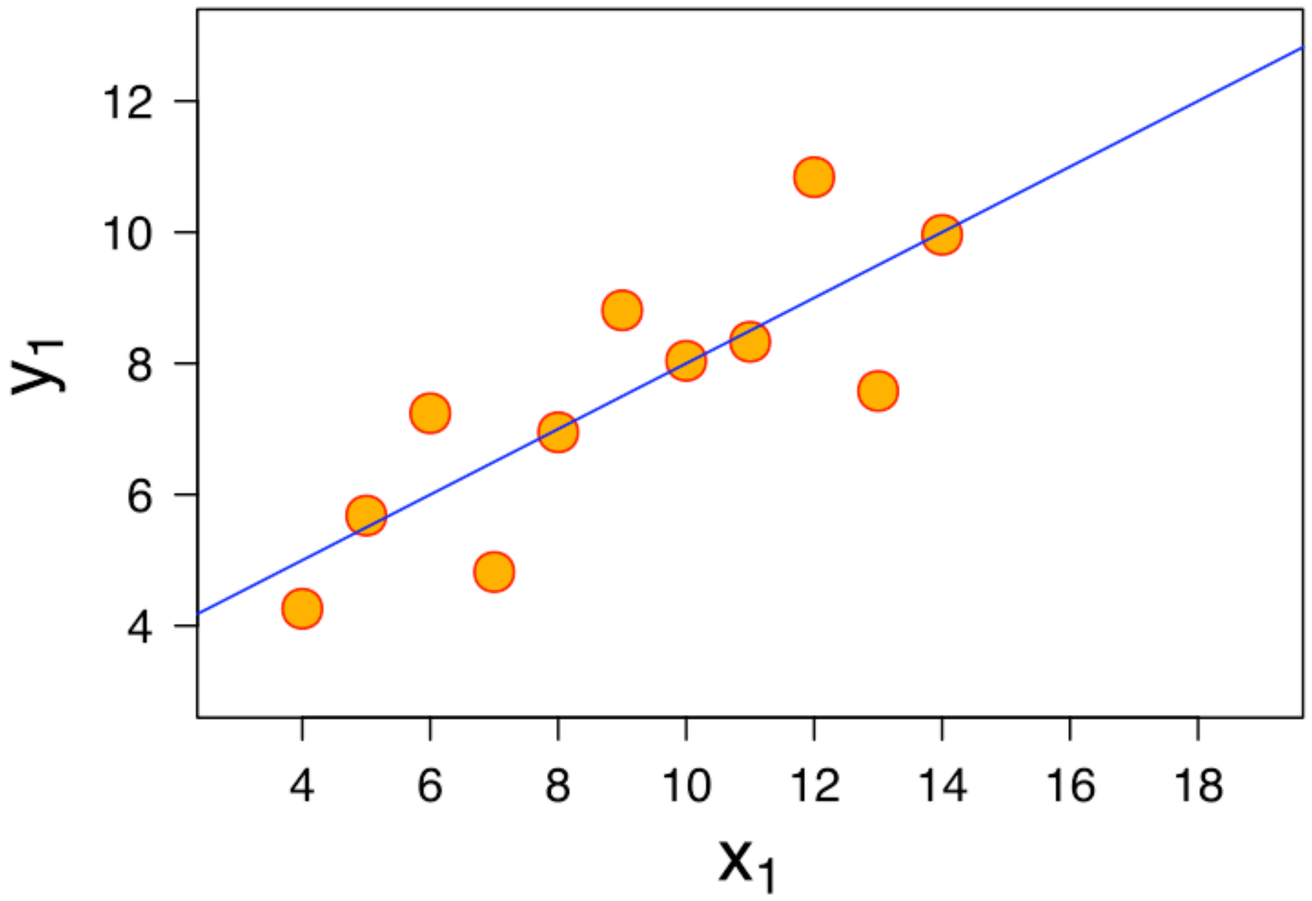
Anatomija regresione jednačine



$$y = \alpha + \beta x$$

regresioni koeficijent b: nagib / vrednost za koju se Y promeni ako povećamo X za 1.

Anatomija regresione jednačine



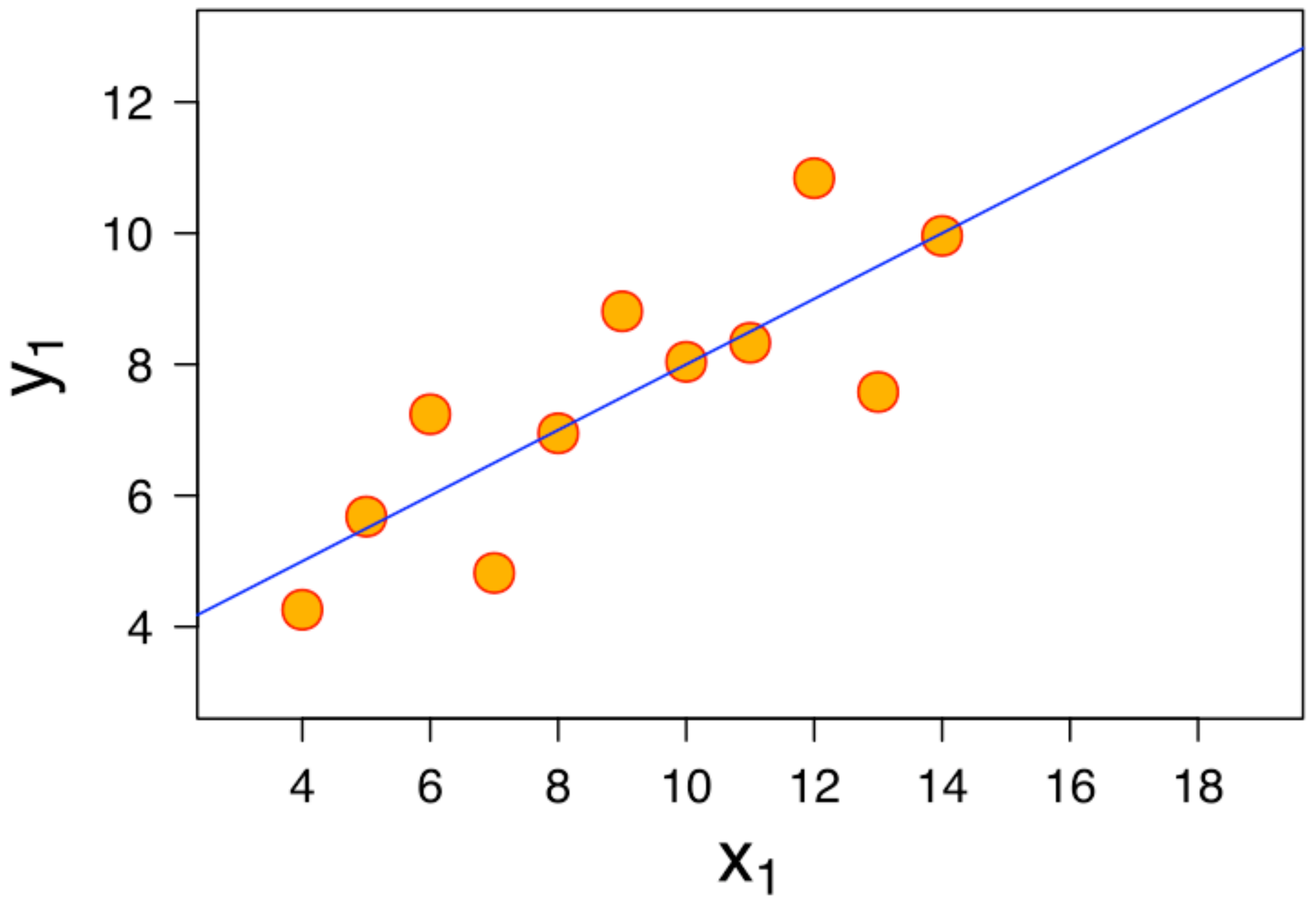
$$y = \alpha + \beta x$$

Koeficijent determinacije

$r^2 = cor(x, y)^2$, od 0 do 1

Udeo objašnjene varijacije

Anatomija regresione jednačine



$$y = \alpha + \beta x$$

$$y = -4.16 + 0.13x$$

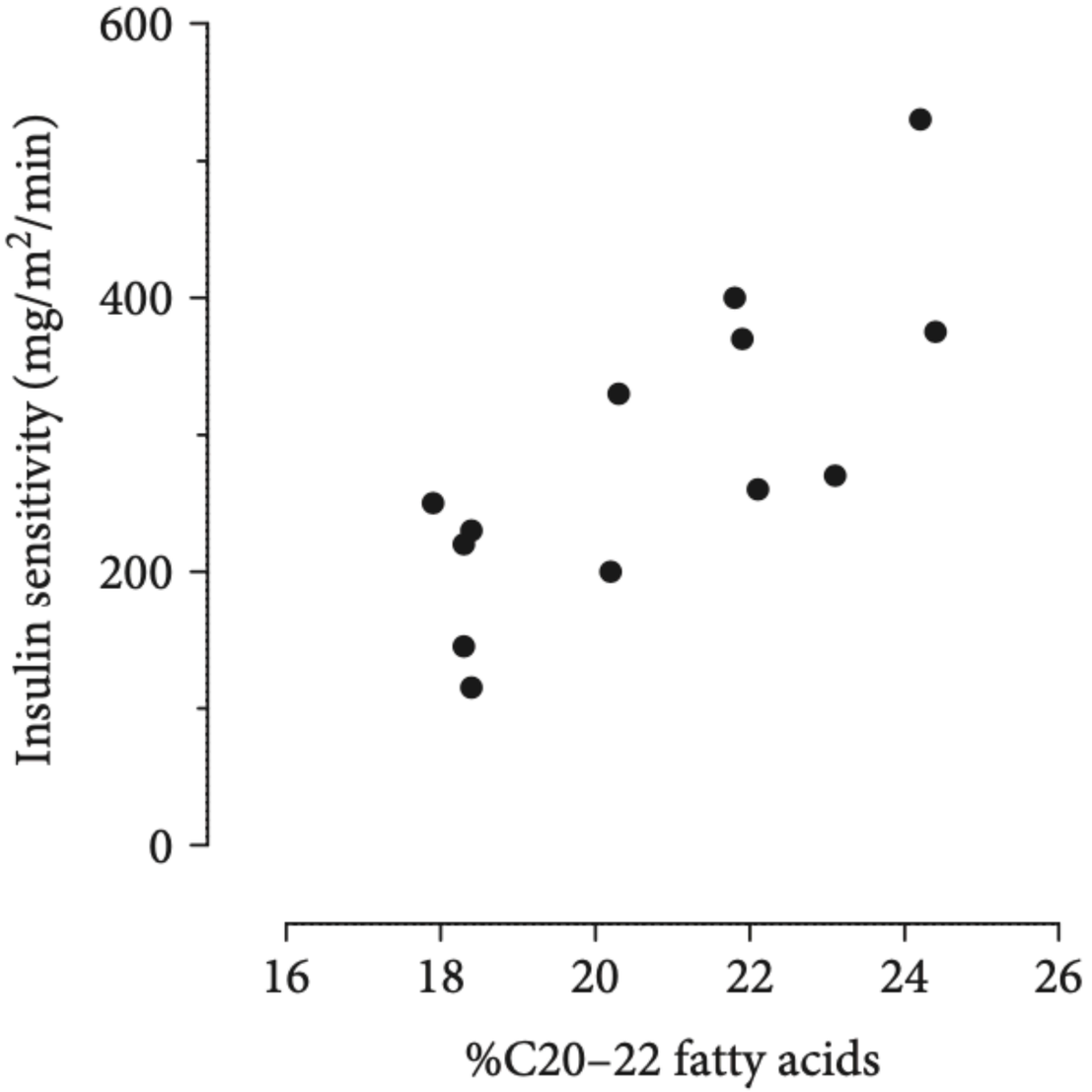
$$r^2 = 0.5 \times 100 = 50\,\%$$

Primer: lipidi i insulinska rezistencija (2)

Borkman et al. (1993)

CORRELATION	
r	0.77
95% CI	0.3803 to 0.9275
r^2	0.5929
P (two-tailed)	0.0021
Number of XY pairs	13

Opisali smo povezanost



Primer: inteligencija i koncentracije spermatozoida

Arden et al. (2008)

	r	r^2	P VALUE
Log (sperm count per ml)	0.15	0.023	0.0019
Log (sperm count per ejaculate)	0.19	0.036	< 0.0001
Fraction of sperm that are motile (%)	0.14	0.020	0.0038

Primer: inteligencija i koncentracije spermatozoida

Arden et al. (2008)

	r	r^2	P VALUE
Log (sperm count per ml)	0.15	0.023	0.0019
Log (sperm count per ejaculate)	0.19	0.036	< 0.0001
Fraction of sperm that are motile (%)	0.14	0.020	0.0038

Primer: inteligencija i koncentracije spermatozoida

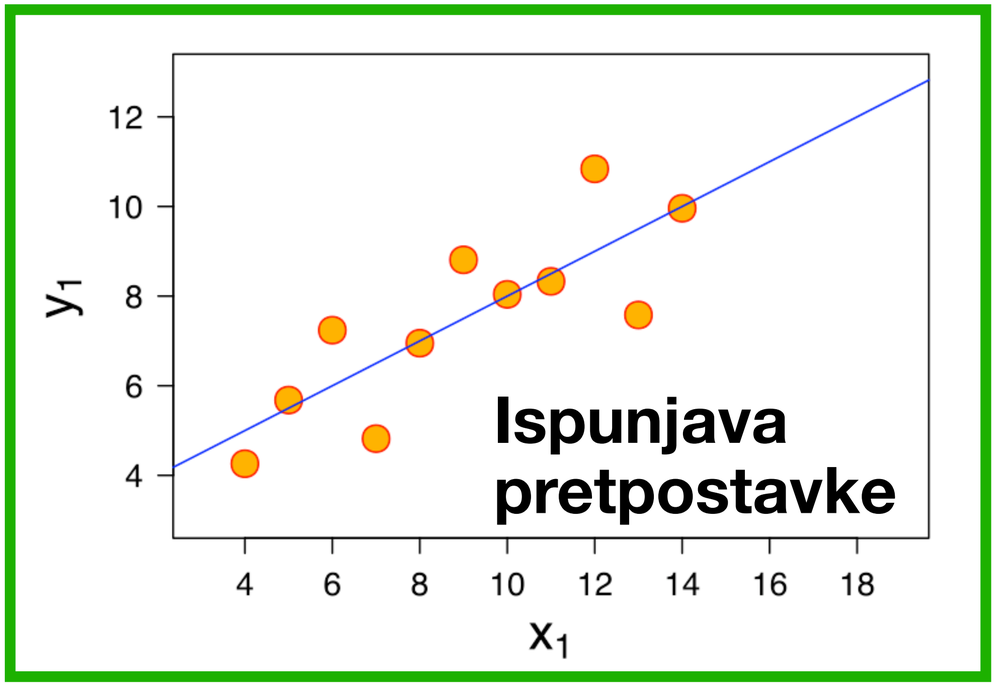
Arden et al. (2008)

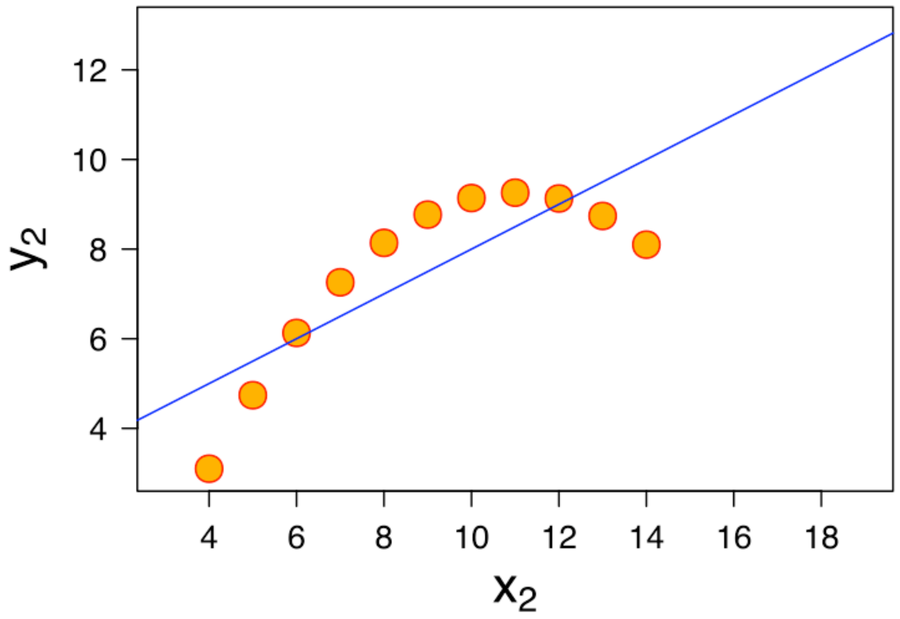
% objašnjene varijanse

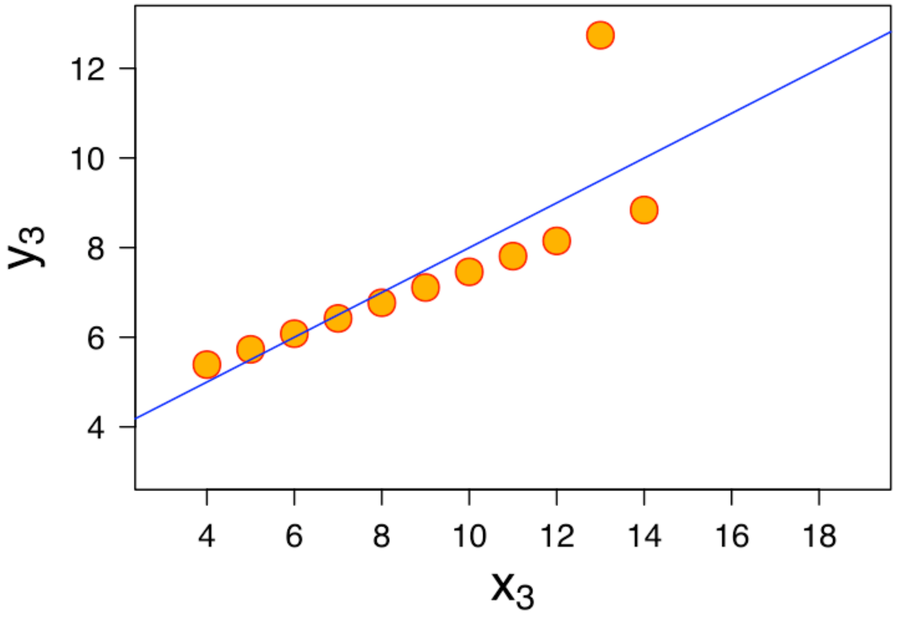
	r	r^2	P VALUE
Log (sperm count per ml)	0.15	0.023	0.0019
Log (sperm count per ejaculate)	0.19	0.036	< 0.0001
Fraction of sperm that are motile (%)	0.14	0.020	0.0038

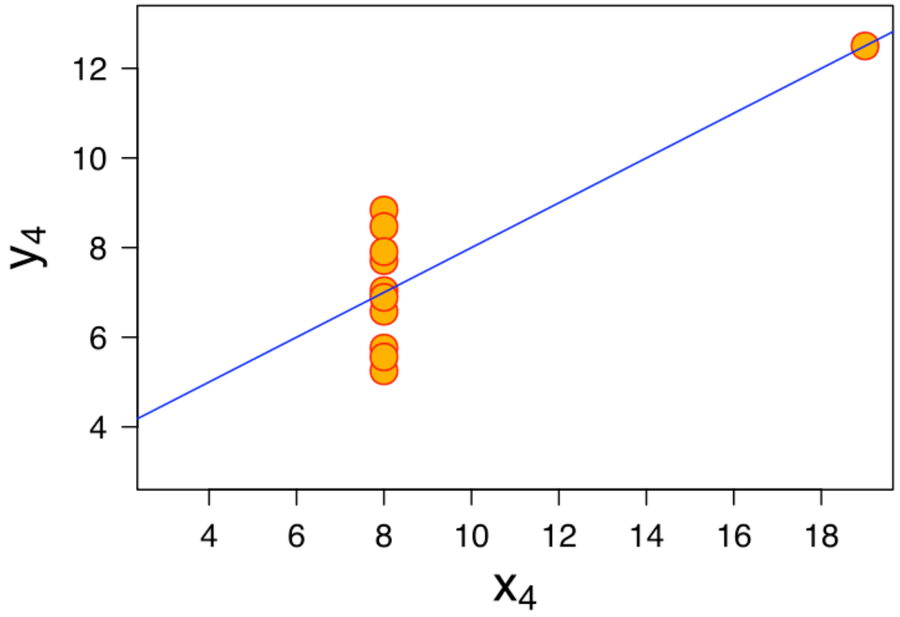
Pretpostavke linearna regresije

1. Normalnost
2. Linearnost
3. Nezavisnost (merenja nisu ponovljena)
4. Jednakost varijansi


Ispunjava
pretpostavke
y1
x1


y2
x2

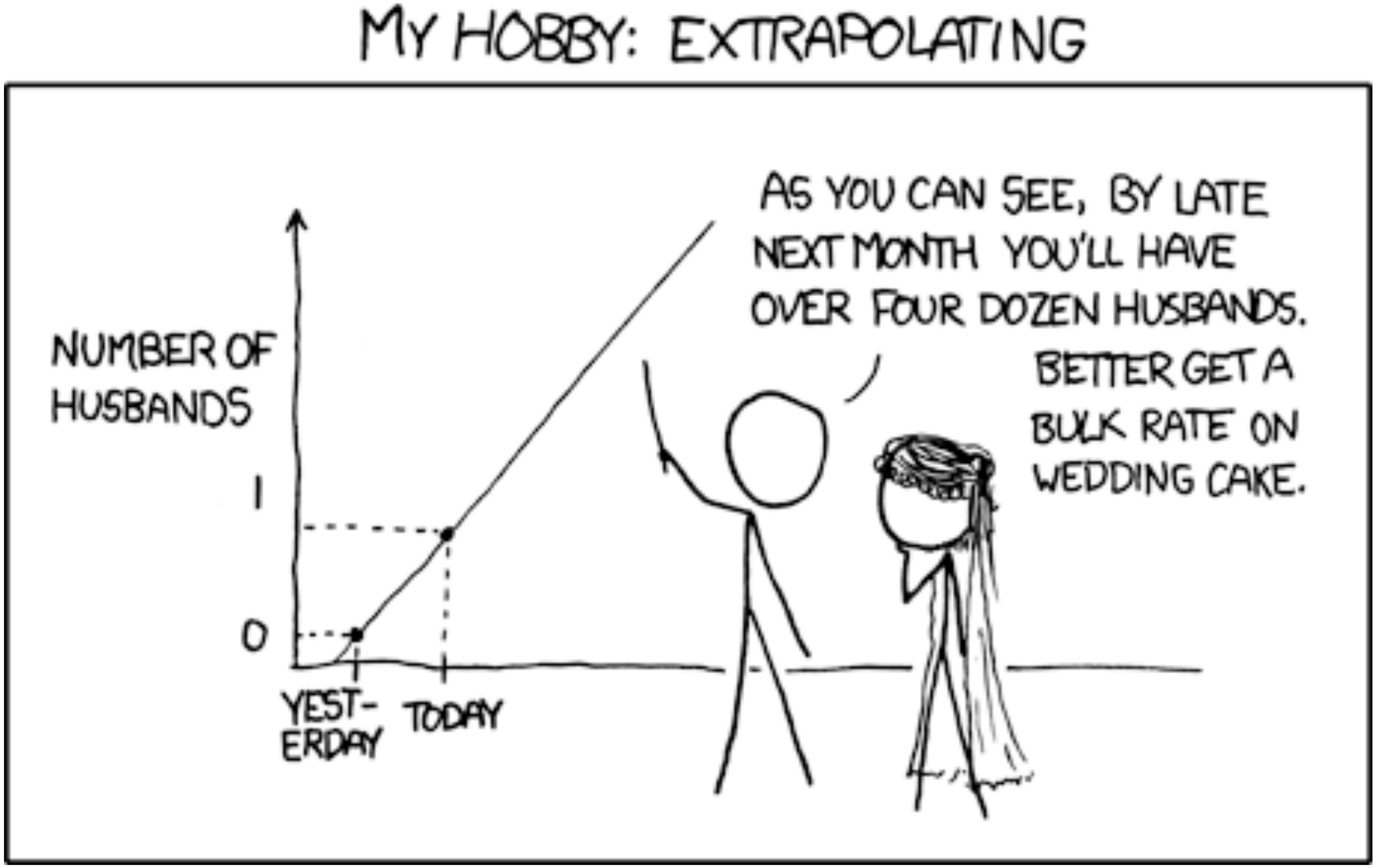

y3
x3

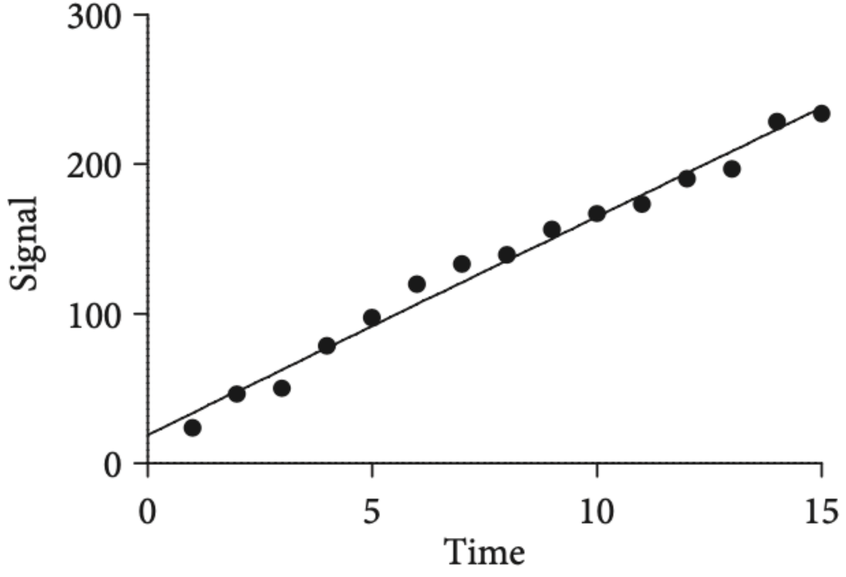

y4
x4

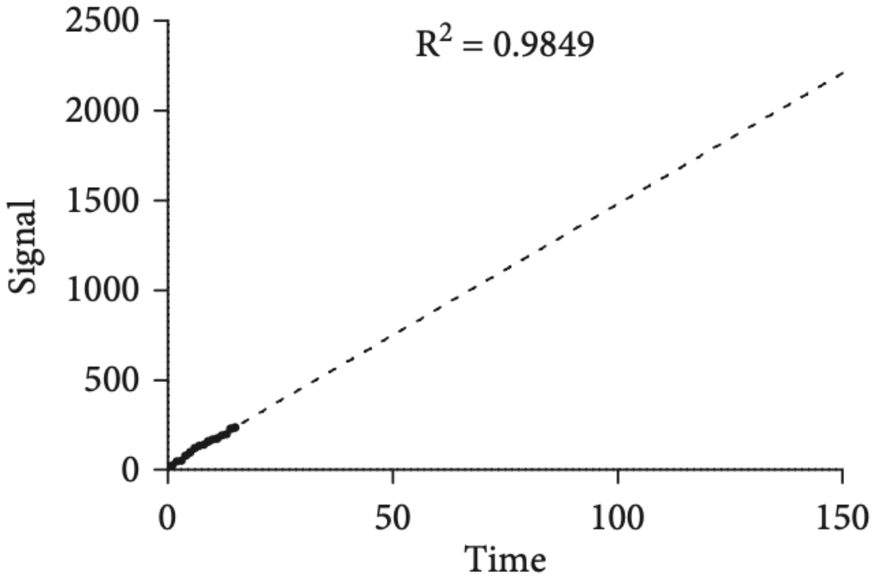
unutar

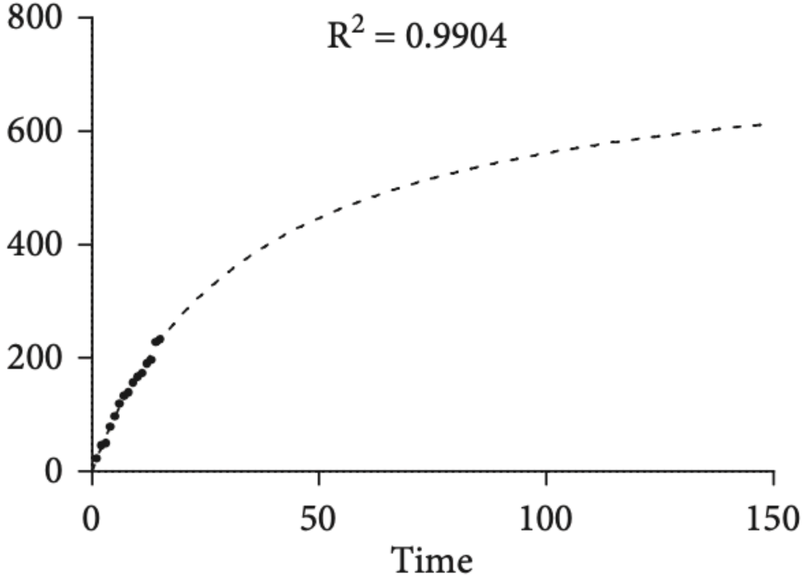
Ekstrapolacija i interpolacija

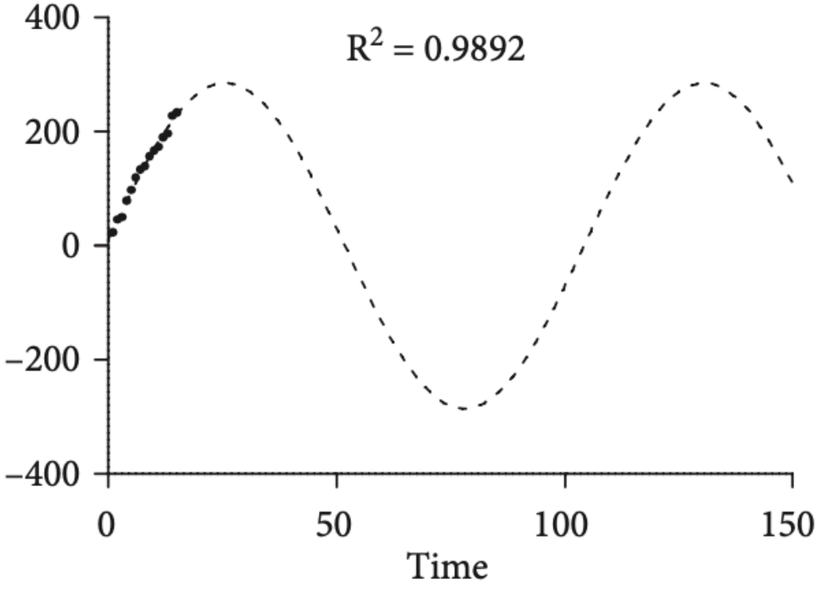
Izvan

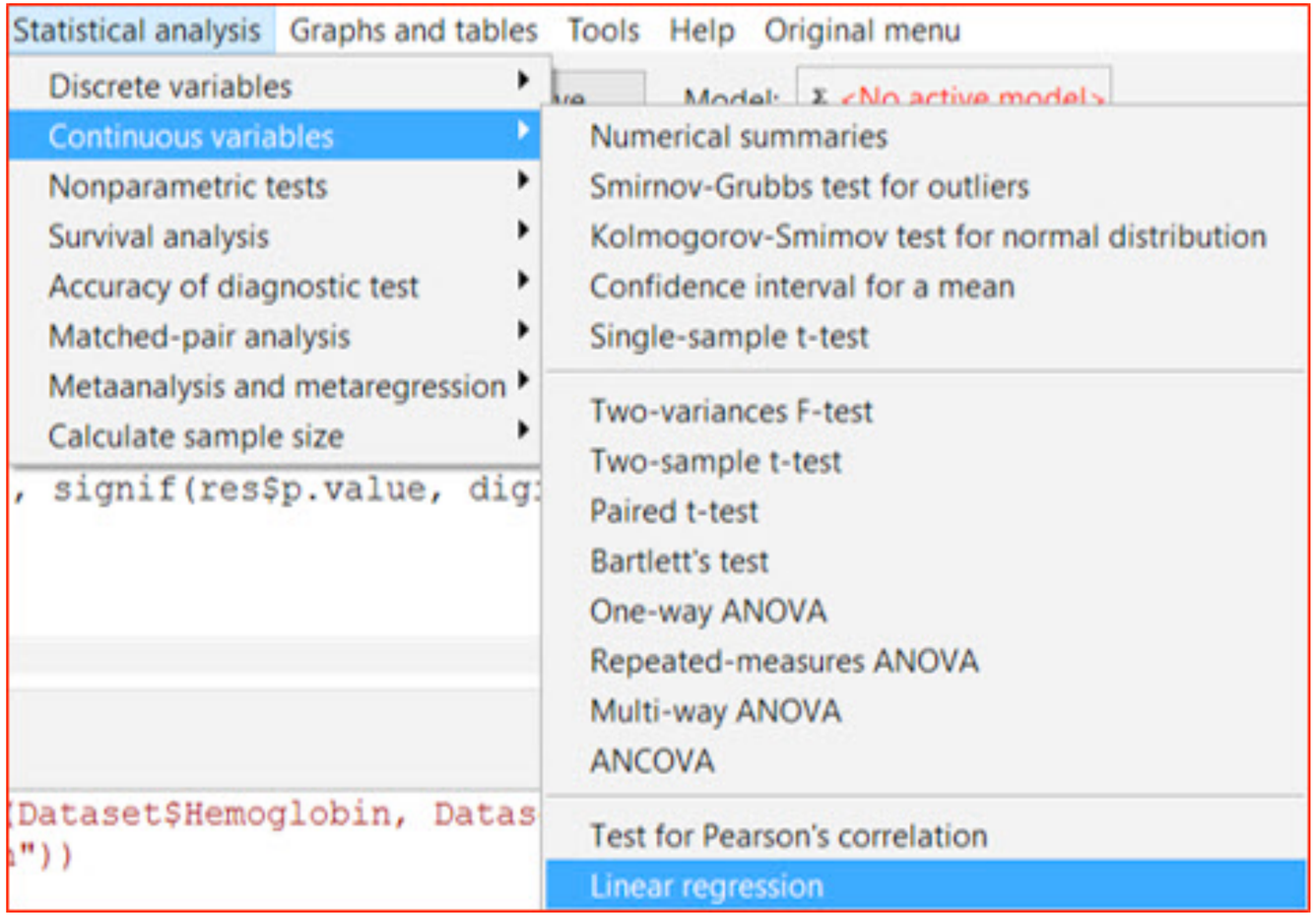


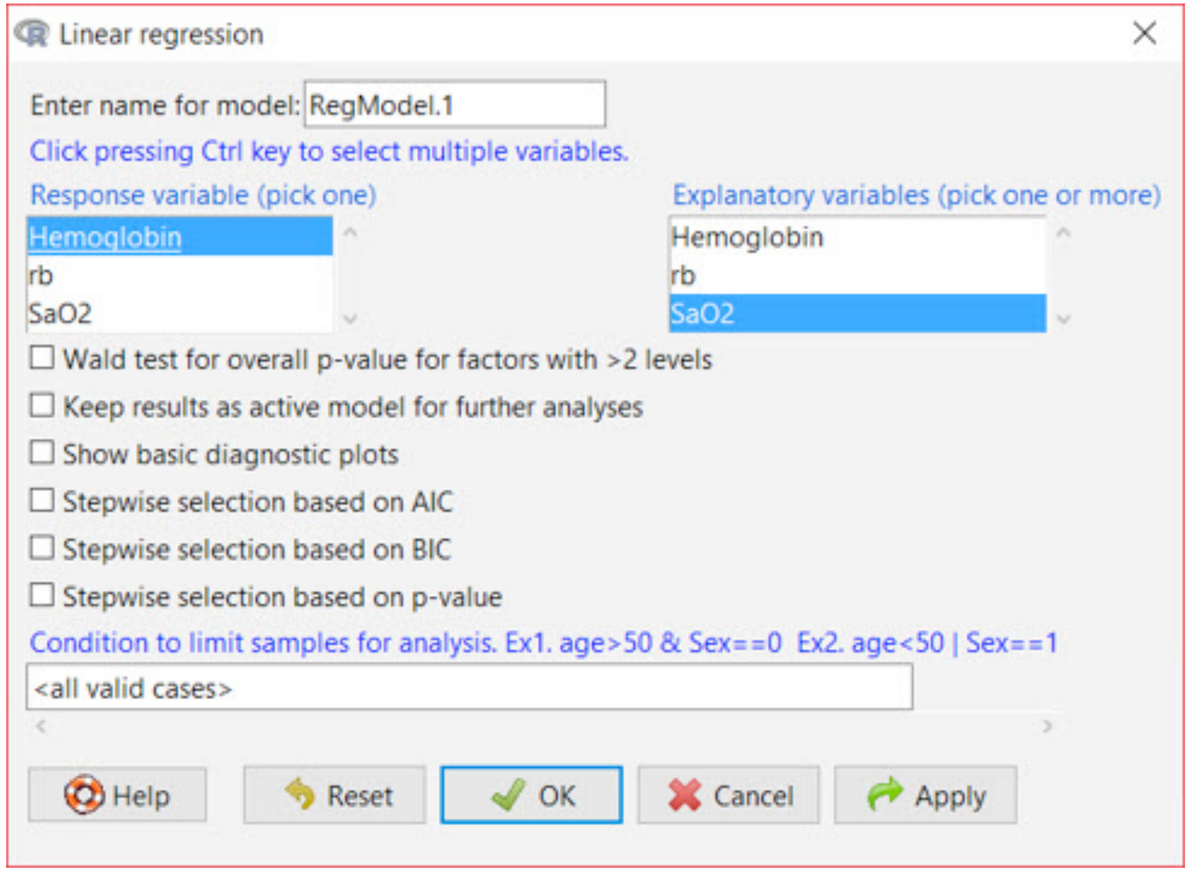

300
200
100
0
Signal
0
5
10
15
Time


2500
2000
1500
1000
500
0
Signal
R2 = 0.9849
0
50
100
150
Time


800
600
400
200
0
R2 = 0.9904
0
50
100
150
Time


400
200
0
−200
−400
R2 = 0.9892
0
50
100
150
Time






```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  339.390     56.536   6.003  0.00184 **
SaO2          -1.984      0.673  -2.948  0.03195 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

regresiona konstanta α

koeficijent nagiba β

$$y = 338.39 - 1.984x$$

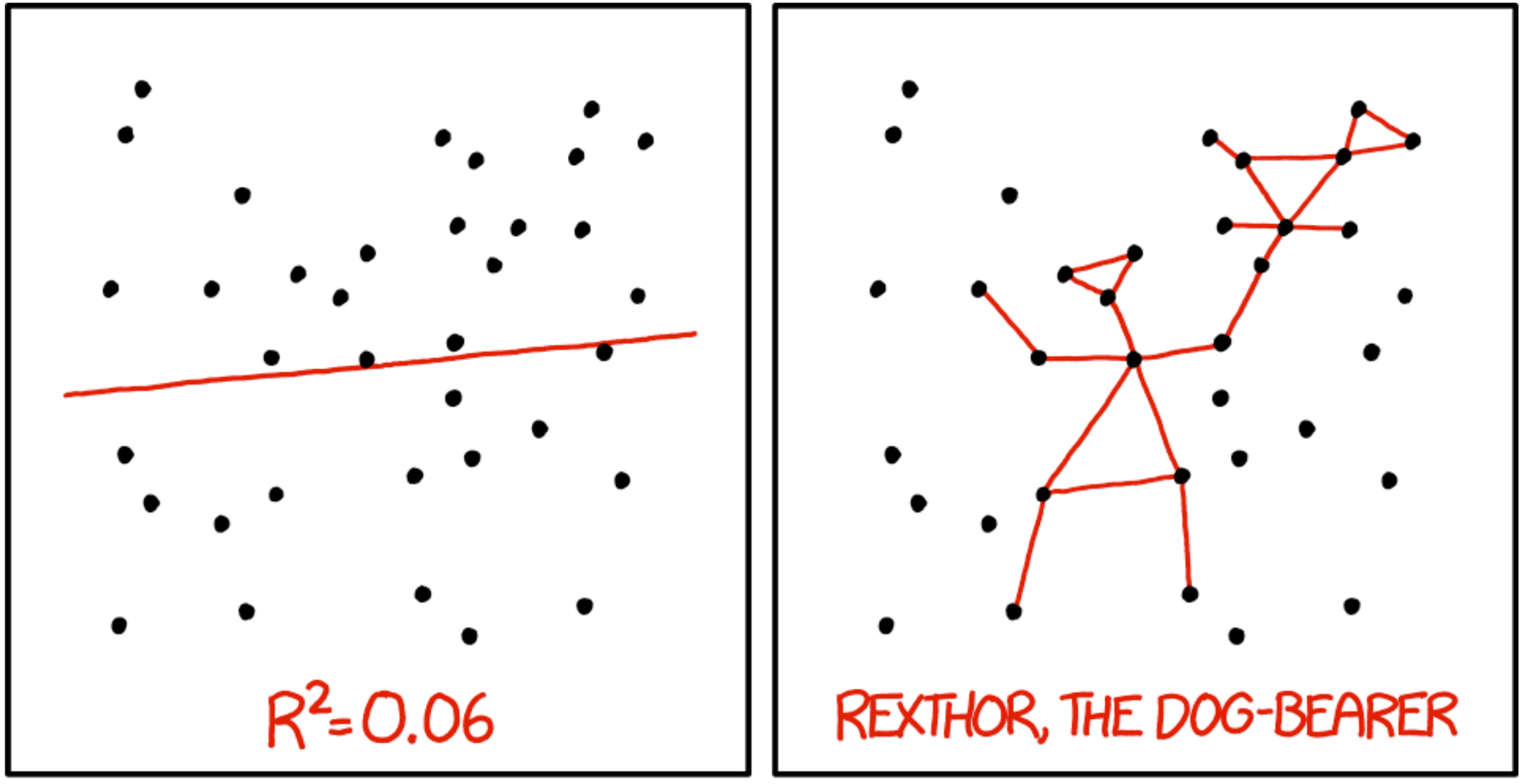
Koristeći dopunjenu bazu DZ odgovori na sledeća istraživačka pitanja:

1. Da li postoji povezanost između sistolne tenzije i ukupnog holesterola?

2. Za ispitanika sa ukupnim holesterolom od 6.2 mmol/l izračunaj vrednost sistolne tenzije (mmmHg).

Korelacija vs. Regresija

- Korelacija opisuje povezanost varijabli, ne određuje najbolju liniju
- Kod korelacija nije potrebno razmišljati o uzrocima (simetrija)
- Kod regresije $Y \leftarrow X \neq X \leftarrow Y$ (nema simetrije)



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.