Proučavanje povezanosti između kontinuiranih varijabli - korelacija i regresija

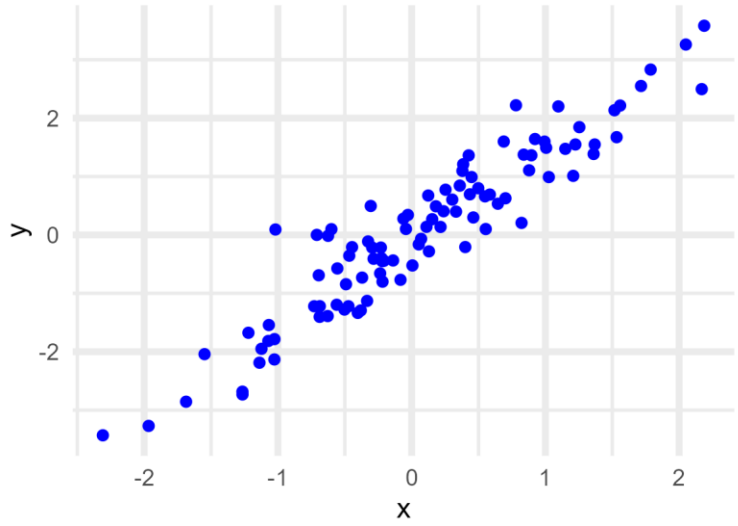
Izv.prof. Rosa Karlić

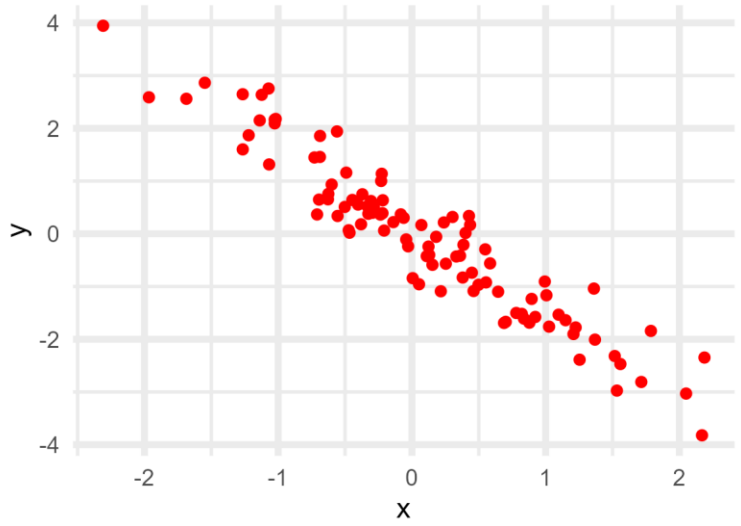
Predavanje 9, MZIRuB 2024/2025

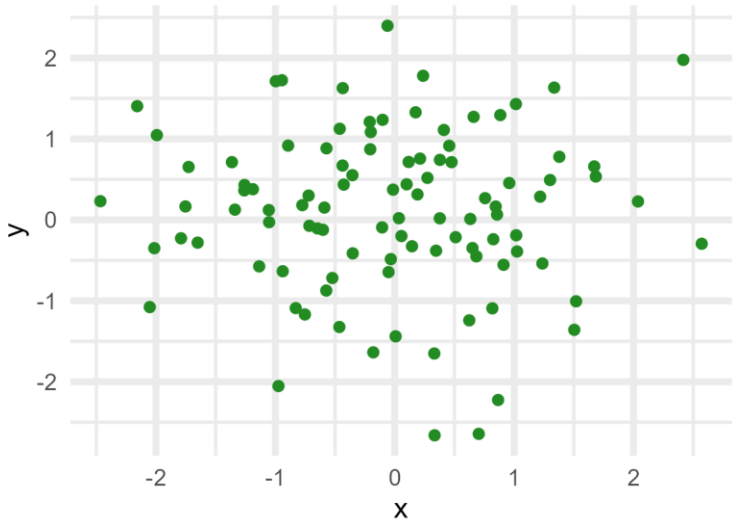
15.01.2024.

Dva moguća cilja

- Opisati odnose između dvije ili više kontinuiranih varijabli
- Koristiti navedene odnose za predviđanje vrijednosti varijabli

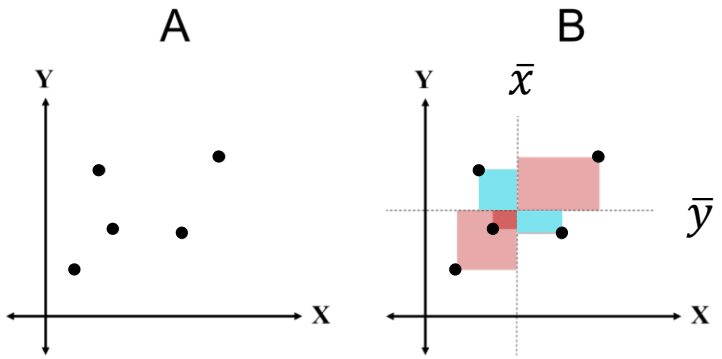




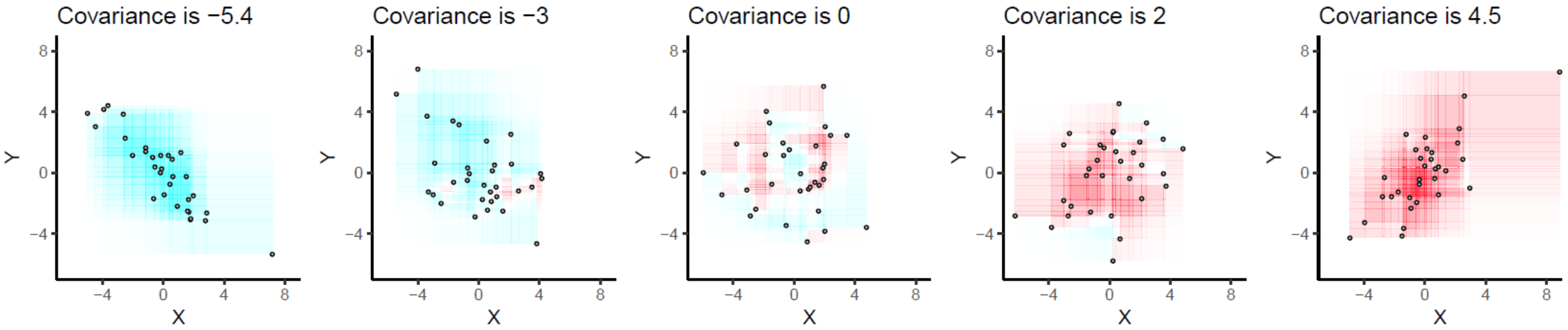


Kovarijanca

- Kovarijanca – koliko se jedna varijabla mijenja kad se druga varijabla mijenja σ_{XY}
- Kovarijanca uzorka: $s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$



Kovarijanca



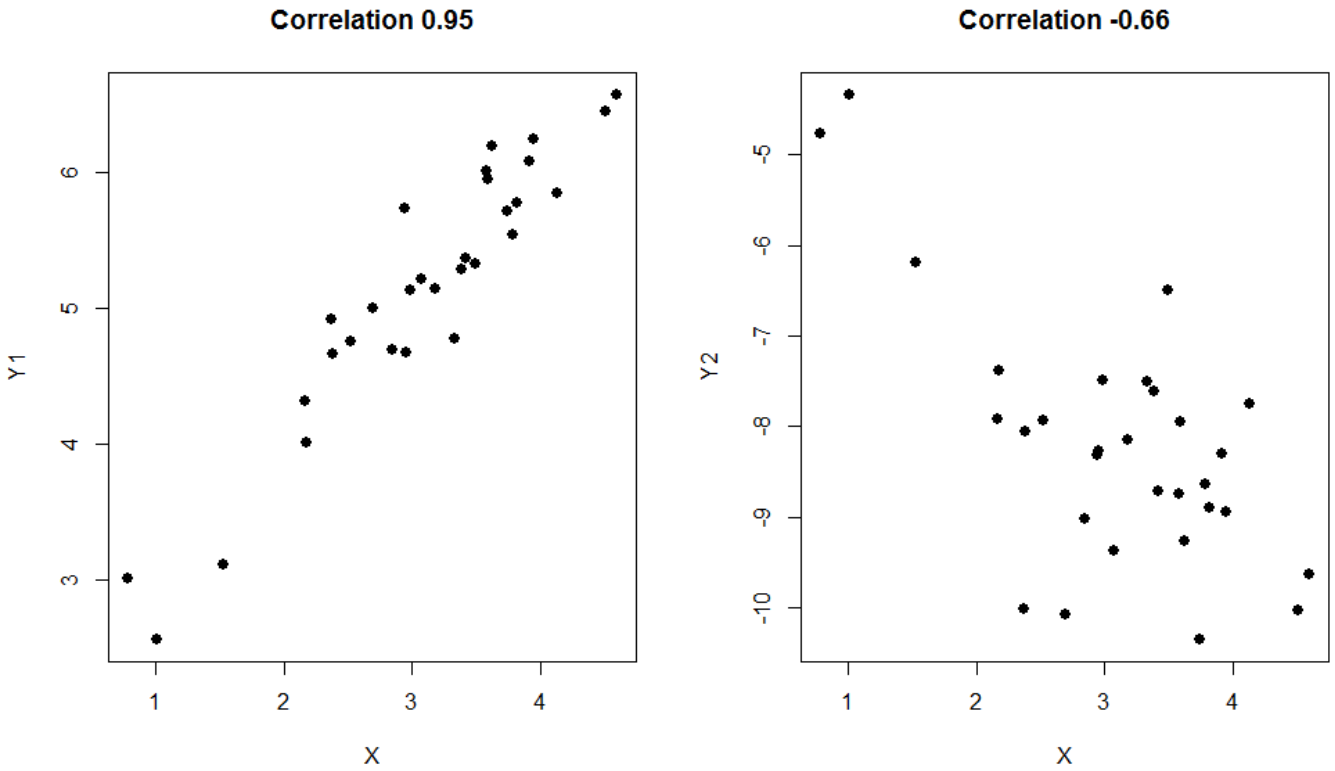
Izvor: https://stats.stackexchange.com/questions/18058/how-would-you-explain-covariance-to-someone-who-understands-only-the-mean

- Proporcionalna skali na kojoj su mjereni X i Y
- Osjetljiva na *outliere* (netipične vrijednosti)

Korelacija

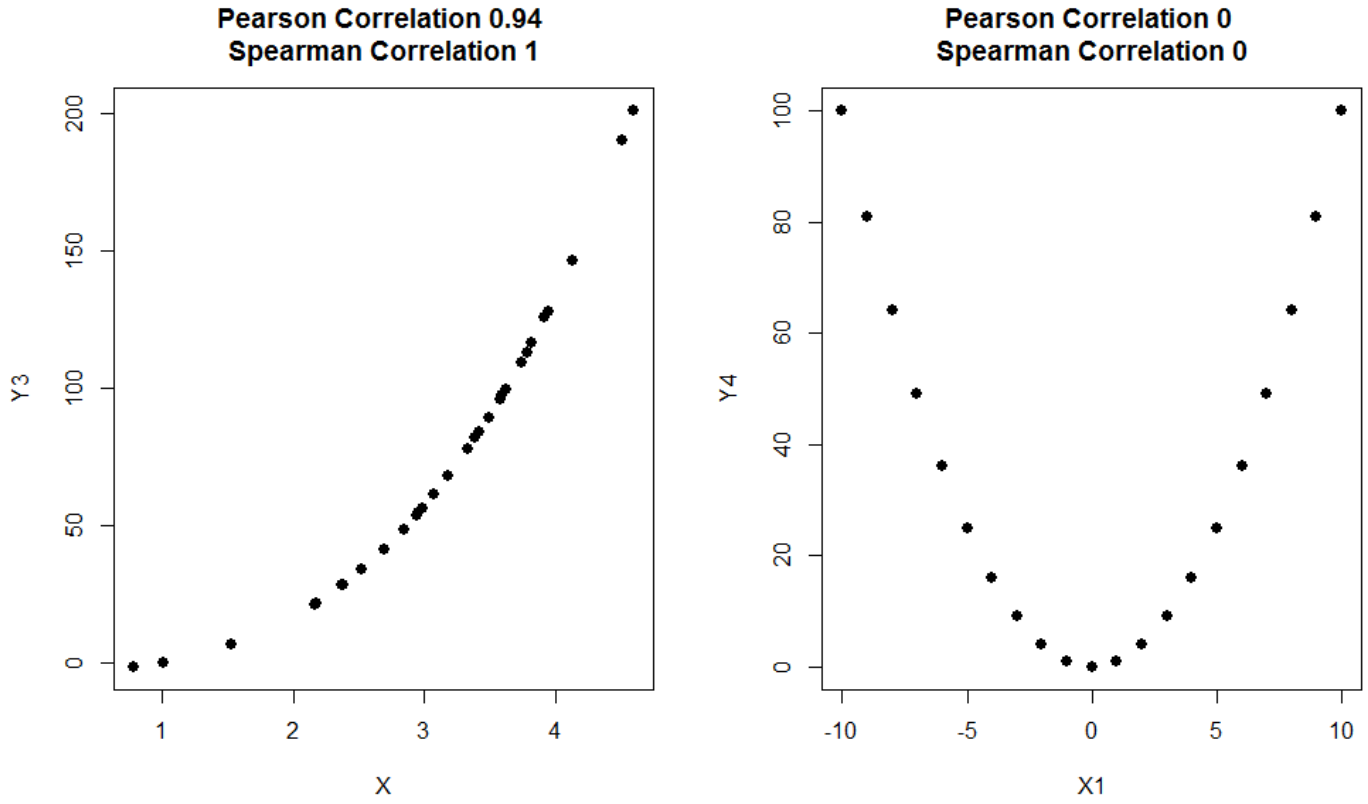
- Korelacija – kovarijanca normalizirana standardnom devijacijom
- Koeficijent korelacije – mjeri snagu odnosa između dviju varijabli (raspon od -1 do 1)
- Pearsonov koeficijent korelacije (r)– linearni odnosi
- Nulta hipoteza: nema linearnog odnosa između dvije varijable

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

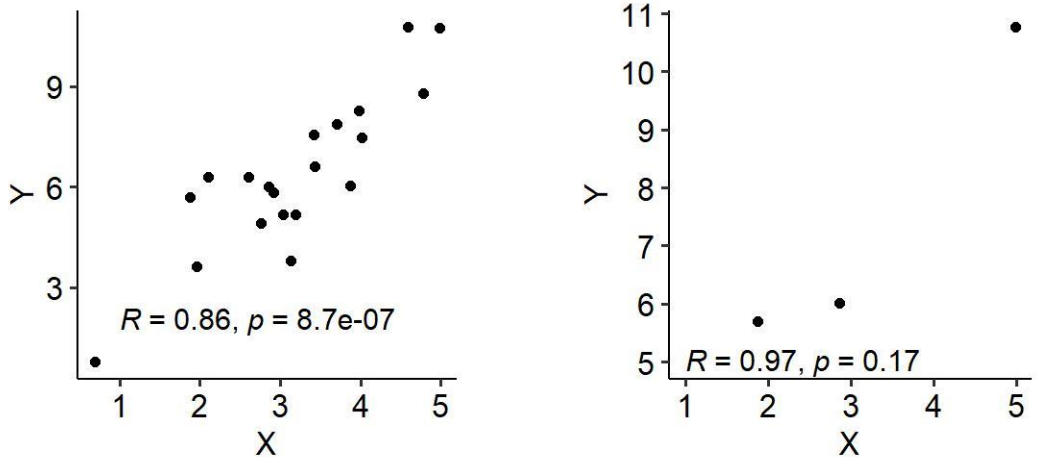


Spearmanov koeficijent korelacije (rho, ρ)

- Pearsonov koeficijent korelacije nakon što su vrijednosti varijabli pretvorene u rangove
- Nulta hipoteza: nema monotonog odnosa između dvije varijable



Značajna povezanost među varijablama?

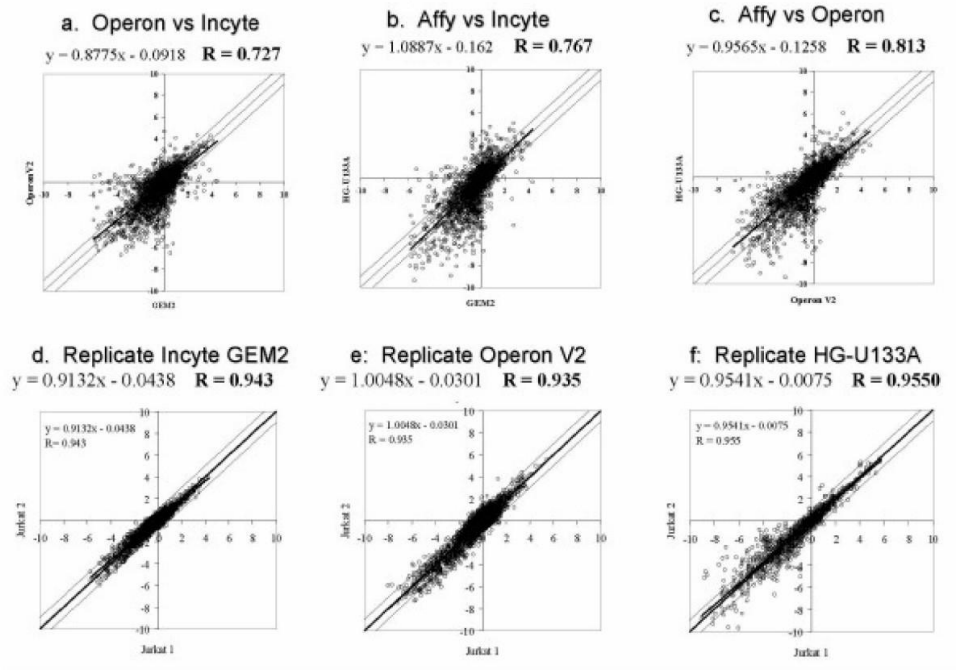


ABSOLUTE VALUE OF R	INTERPRETATION
< 0.19	Slight; almost no relationship
0.20–0.39	Low correlation; definite but small relationship
0.40–0.69	Moderate correlation; substantial relationship
0.70–0.89	High correlation; strong relationship
0.90–1.00	Very high correlation; very dependable relationship
≥ 0.30	Practically significant relationship

Izvor: https://doi.org/10.4102/sajhrm.v7i1.175

- Koeficijent determinacije = r^2 jačina povezanosti (R^2) – proporcija varijabilnosti u jednoj varijabli koja je objašnjena drugom varijablom

Korelacija u biološkim eksperimentima



- **a-f**. Scatter plot analysis to determine correlation coefficients between and within platforms using Jurkat RNA as an example. Correlations for all cell lines are given in Table 4. (a) Operon versus Incyte (b) Affymetrix versus Incyte (c) Affymetrix versus Operon (d) GEM2 versus GEM2 replicate correlation (e) Operon versus Operon (f) HG-U133A versus HG-U133A

Petersen, D., Chandramouli, G., Geoghegan, J. *et al.* Three microarray platforms: an analysis of their concordance in profiling gene expression. *BMC Genomics* **6**, 63 (2005). https://doi.org/10.1186/1471-2164-6-63

Korelacija u biološkim eksperimentima

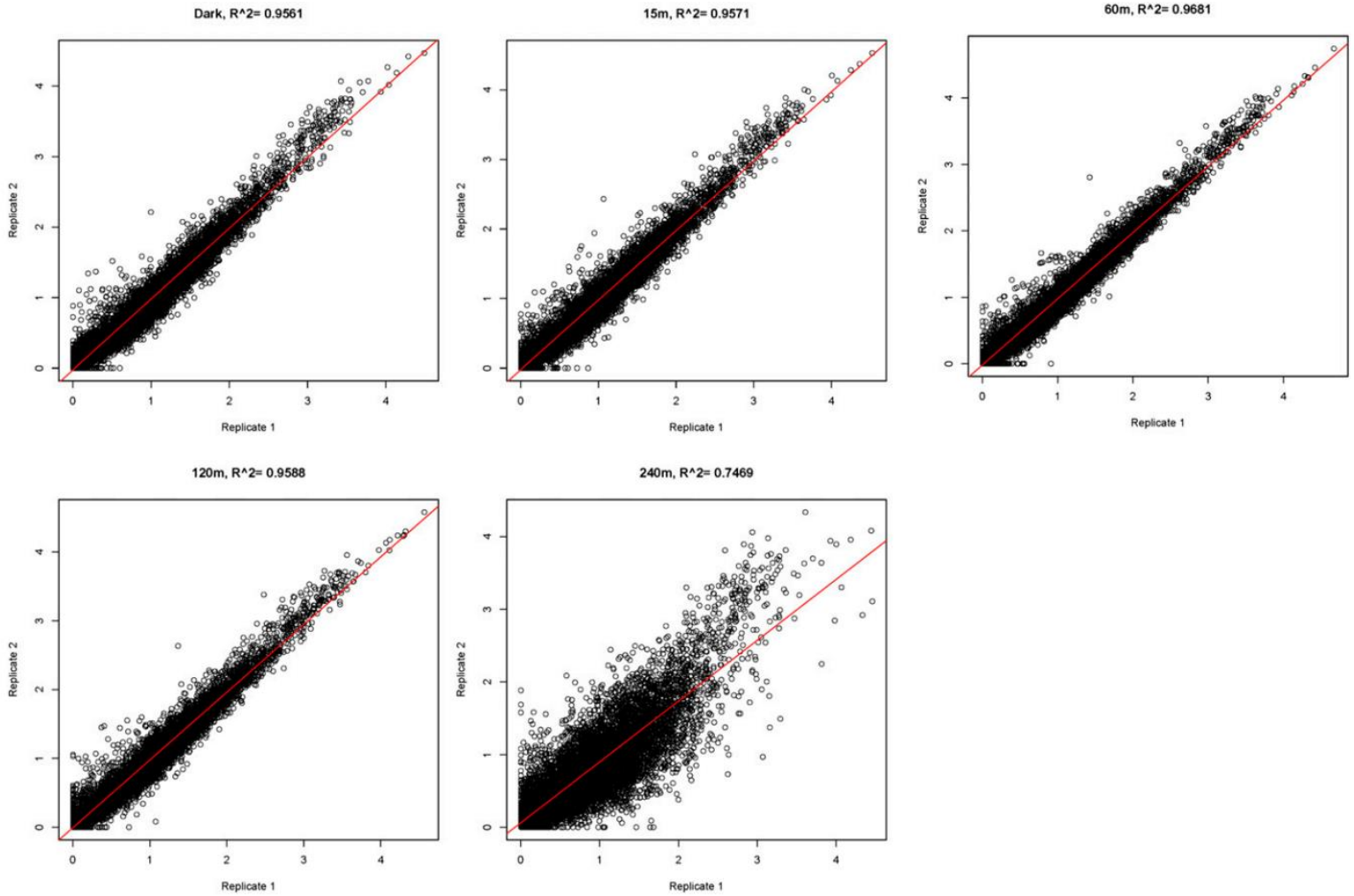


Figure 1 Comparison of RNA-seq replicate experiments. The FPKM for biological replicate 1 is plotted against biological replicate 2 for each gene, demonstrating strong correlation between replicate experiments at each time point. The correlation coefficient, R, is shown for each time point.

Pažnja kod izračuna korelacija – uzorkovanje iz više različitih populacija

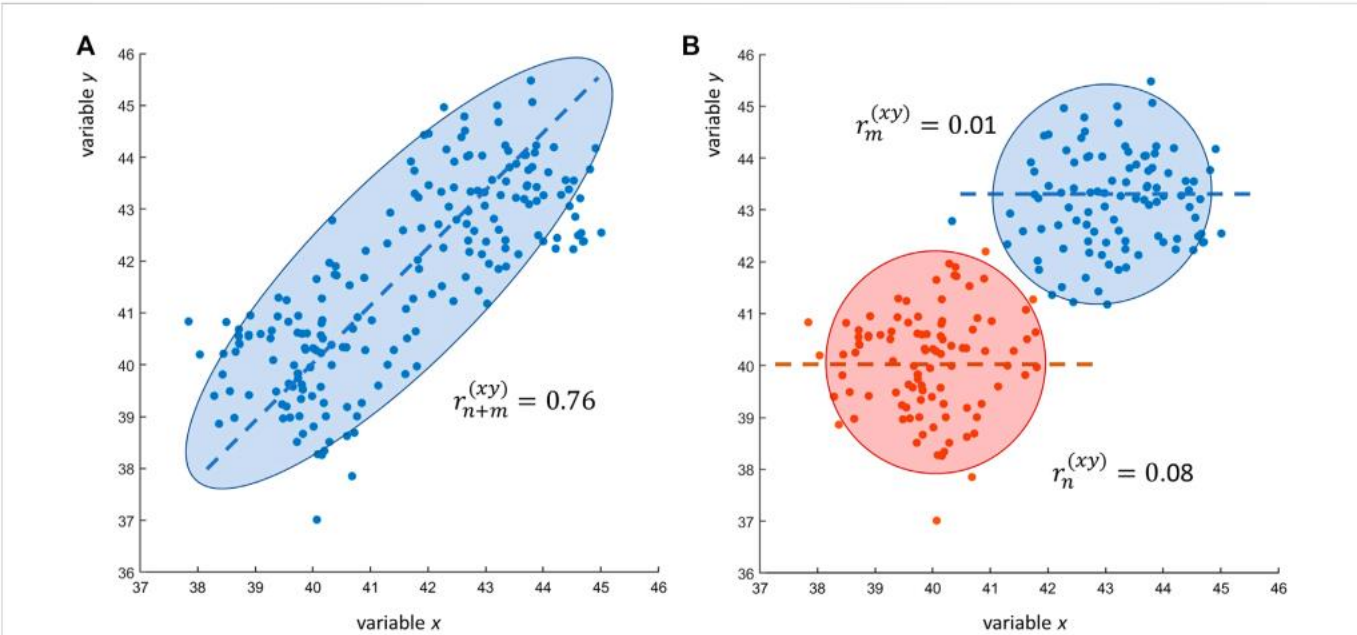


FIGURE 2
Research scenarios RS1 and RS2. **(A)** Scatter plot of $n + m = 200$ observations of two variables x and y: the Pearson's sample correlation between x and y is $r_{n+m}^{(xy)} = 0.76$. **(B)** The same scatter plot as in Panel A but with data points color coded to highlight the actual data structure: when taken separately the $n = m = 100$ observations of x and y (Condition A: blue; condition B: red) are uncorrelated: $r_n^{(xy)} = 0.01$ (data set A) and $r_n^{(xy)} = 0.08$ (data set B). The observed high correlation $r_{n+m}^{(xy)} = 0.76$ is an artifact due to the merging of two data set containing variables coming from two different populations: in this case x and y come from two independent normal distributions with population means $\mu_A = (43, 43)$ and $\mu_B = (40, 40)$ and unit variance $\nu^2 = 1$.

https://www.frontiersin.org/articles/10.3389/fsysb.2023.1042156/full

Pažnja kod izračuna korelacija – uzorkovanje iz više različitih populacija

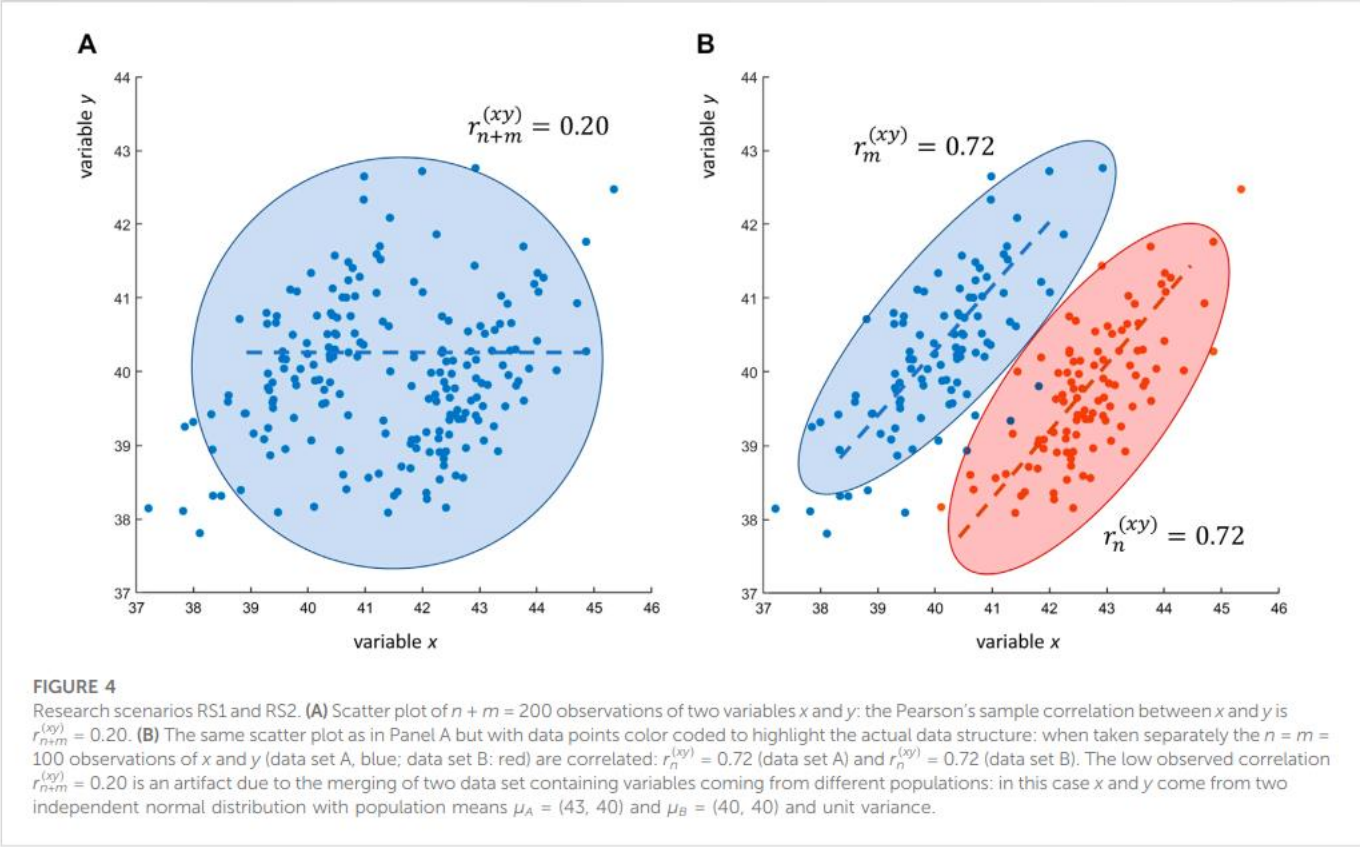


FIGURE 4
Research scenarios RS1 and RS2. **(A)** Scatter plot of $n + m = 200$ observations of two variables x and y: the Pearson's sample correlation between x and y is $r_{n+m}^{(xy)} = 0.20$. **(B)** The same scatter plot as in Panel A but with data points color coded to highlight the actual data structure: when taken separately the $n = m = 100$ observations of x and y (data set A, blue; data set B: red) are correlated: $r_n^{(xy)} = 0.72$ (data set A) and $r_n^{(xy)} = 0.72$ (data set B). The low observed correlation $r_{n+m}^{(xy)} = 0.20$ is an artifact due to the merging of two data set containing variables coming from different populations: in this case x and y come from two independent normal distribution with population means $\mu_A = (43, 40)$ and $\mu_B = (40, 40)$ and unit variance.

https://www.frontiersin.org/articles/10.3389/fsysb.2023.1042156/full

Pažnja kod izračuna korelacija – utjecaj netipičnih točaka (outliers)

Pearsonov koeficijent korelacije

Spearmanov koeficijent korelacije

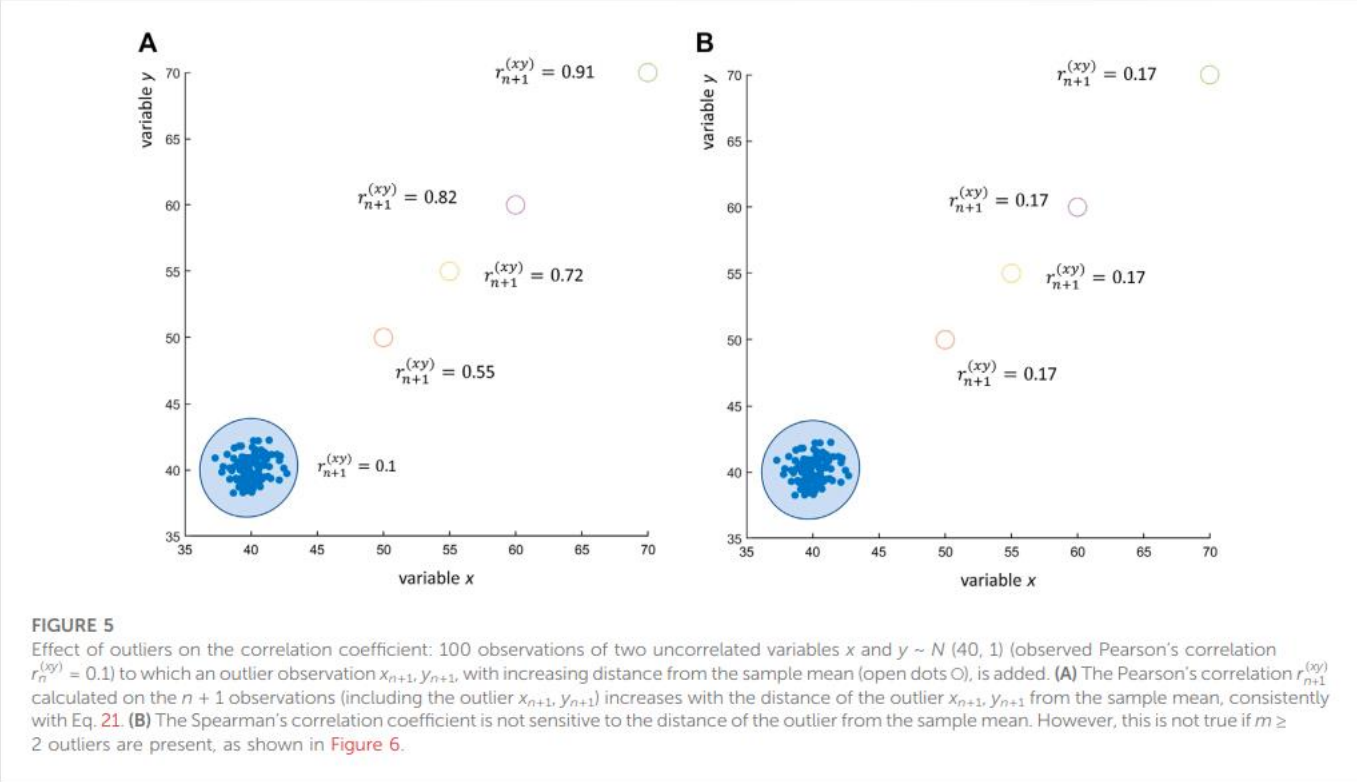
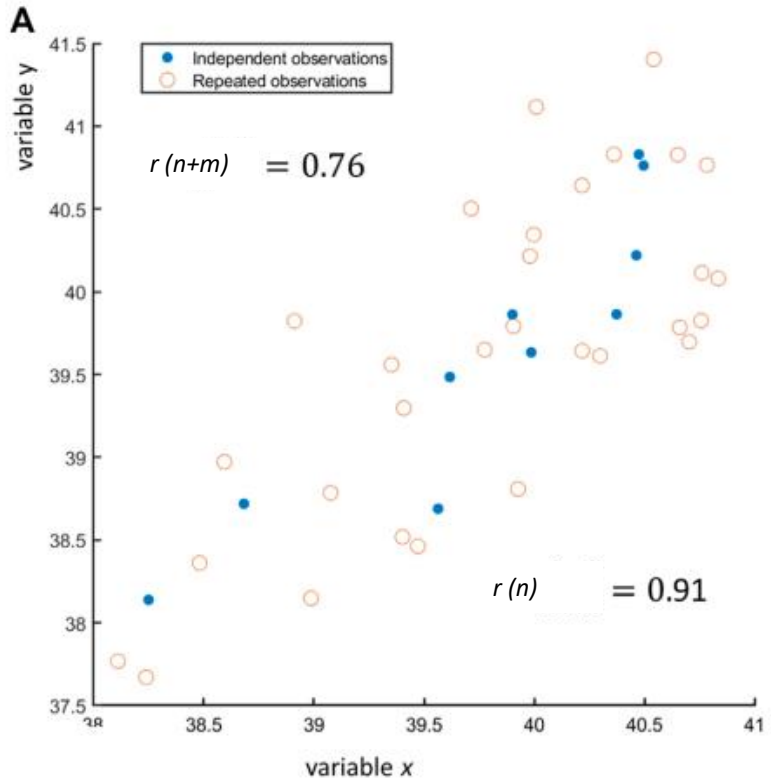


FIGURE 5
Effect of outliers on the correlation coefficient: 100 observations of two uncorrelated variables x and $y \sim N$ (40, 1) (observed Pearson's correlation $r_n^{(xy)} = 0.1$) to which an outlier observation x_{n+1}, y_{n+1}, with increasing distance from the sample mean (open dots ○), is added. **(A)** The Pearson's correlation $r_{n+1}^{(xy)}$ calculated on the $n + 1$ observations (including the outlier x_{n+1}, y_{n+1}) increases with the distance of the outlier x_{n+1}, y_{n+1} from the sample mean, consistently with Eq. 21. **(B)** The Spearman's correlation coefficient is not sensitive to the distance of the outlier from the sample mean. However, this is not true if $m \geq 2$ outliers are present, as shown in Figure 6.

https://www.frontiersin.org/articles/10.3389/fsysb.2023.1042156/full

Pažnja kod izračuna korelacija – uzorci nisu međusobno neovisni



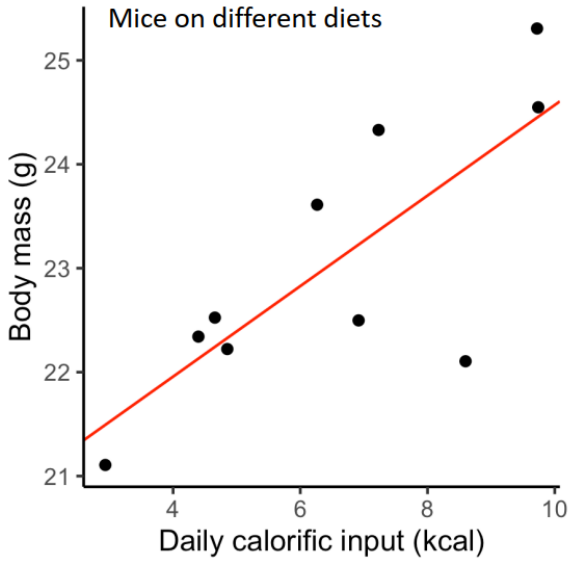
https://www.frontiersin.org/articles/10.3389/fsysb.2023.1042156/full

Linearna regresija

- Jednostavan kvantitativni model
- Zavisnu varijablu (*output, response*) pokušavamo modelirati kao linearnu kombinaciju jedne ili više nezavisnih varijabli (*input, predictor*)
- Cilj je pronaći linearni model koji najbolje opisuje naše podatke:

$$y_i = \beta_0 + \beta_i x_i + \epsilon_i$$

	Daily calorific input (kcal)	Body mass (g)
1	8.6	22.1
2	2.9	21.1
3	4.4	22.3
4	4.9	22.2
5	9.7	25.3
6	9.7	24.5
7	6.3	23.6
8	4.7	22.5
9	7.2	24.3
10	6.9	22.5



Adapted from https://dag.compbio.dundee.ac.uk/workshops/statistics_lectures/13_Linear_models.pdf

Linearna regresija

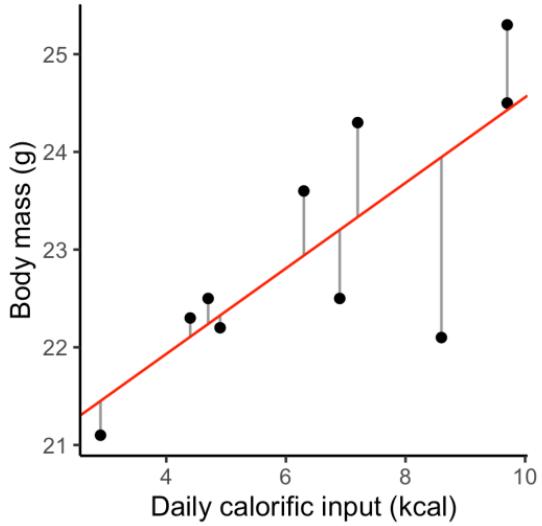
```
> f <- lm(mass ~ kcal, data=ms)
> summary(f)

Call:
lm(formula = mass ~ kcal, data = ms)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8462 -0.2947  0.1323  0.5608  0.9667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.1813     0.8750  23.065 1.33e-08 ***
kcal          0.4378     0.1269   3.449  0.00871 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8862 on 8 degrees of freedom
Multiple R-squared:  0.5979,        Adjusted R-squared:  0.5476
F-statistic:  11.9 on 1 and 8 DF,  p-value: 0.008709
```



- β_0 - **odsječak na osi Y, prosječna vrijednost Y ako su svi X jednaki 0**, u našem primjeru predstavlja predviđenu tjelesnu masu (u gramima) kada je unos kalorija nula.)
- β_j **- prosječno povećanje Y kad se X_j poveća za 1 jedinicu i svi ostali X su konstantni,** u našem primjeru prosječna promjena tjelesne mase (u gramima) povezana s povećanjem unosa kalorija za jednu jedinicu (1 kcal).

Adapted from https://dag.compbio.dundee.ac.uk/workshops/statistics_lectures/13_Linear_models.pdf

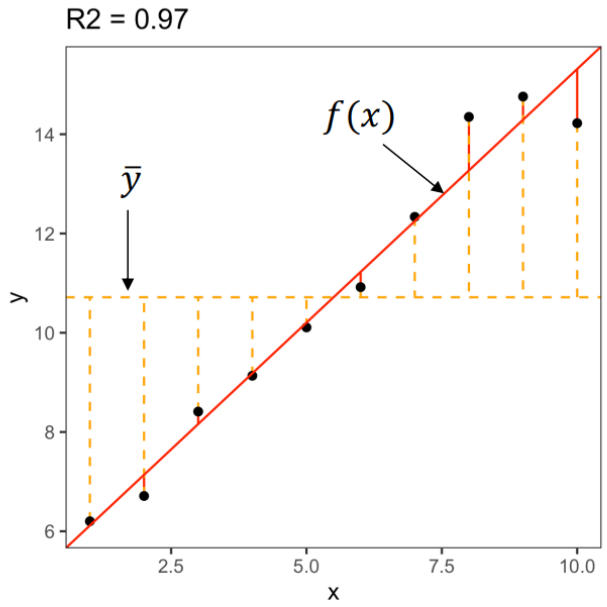
Linearna regresija

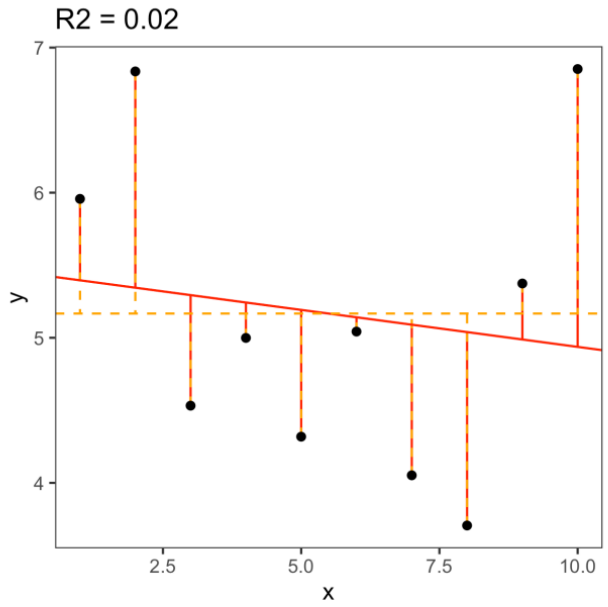
$$R^2 = 1 - \frac{\text{variance explained by the model}}{\text{total variance}}$$

$$R^2 = 1 - \frac{\sum(y_i - f(x_i))^2}{\sum(y_i - \bar{y})^2}$$

It is a measure of fit quality.
The higher R^2, the better fit.

Adjusted R^2 takes into account number of model parameters.

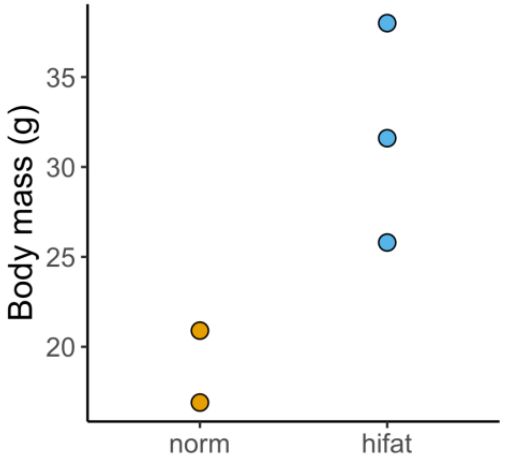




Adapted from https://dag.compbio.dundee.ac.uk/workshops/statistics_lectures/13_Linear_models.pdf

Kategorički prediktori

	Body mass (g)	Diet
1	16.8	norm
2	20.9	norm
3	25.8	hifat
4	38.0	hifat
5	31.6	hifat



- Kategorički prediktori moraju se kodirati numeričkim vrijednostima kako bi se mogli koristiti u linearnoj regresiji (*dummy coding*)

- Jedna kategorija je referentna (*baseline*) i kodira se sa 0, a druga kategorija sa 1
- Naš primjer: „normalna" dijeta se kodira sa 0, a „hifat" sa 1

Rezultat analize:

```
> f1 <- lm(mass ~ diet, data = mdat)
> summary(f1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.900      3.708   5.098   0.0146 *
diethifat     12.900      4.787   2.695   0.0741 .
```

H_0: effect size is equal zero

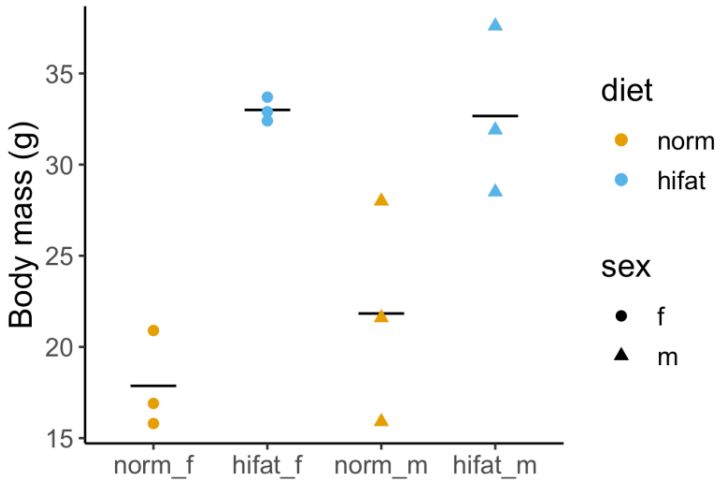
dietfat = difference between normal and high-fat diet

- <u>Interpretacija koeficijenata</u>:
- β_0 Srednja vrijednost tjelesne mase za skupinu na "Normalnoj" prehrani.
- β_1 Srednja razlika u tjelesnoj masi između skupine s „hifat" prehranom i skupine s "normalnom" prehranom. (Razlika u usporedbi s referentnom kategorijom).
- Npr. Srednja vrijednost razlike u tjelesnoj masi miševa na „hifat" prehrani u usporedbi s miševima na „normalnoj" prehrani je 12.9 g (ali koeficijent nije statistički značajan - ne bismo ga trebali tumačiti!)

Adapted from https://dag.compbio.dundee.ac.uk/workshops/statistics_lectures/13_Linear_models.pdf

Više od jednog prediktora

	Body mass (g)	Diet	Sex
1	16.9	norm	f
2	20.9	norm	f
3	15.8	norm	f
4	28.0	norm	m
5	21.6	norm	m
6	15.9	norm	m
7	32.4	hifat	f
8	33.7	hifat	f
9	32.9	hifat	f
10	28.5	hifat	m
11	37.6	hifat	m
12	31.9	hifat	m



- U linearnoj regresiji možemo koristiti i više od jednog (numeričkog ili kategoričkog) prediktora – višestruka regresija
- Može doći do promjene vrijednosti i/ili p-vrijednosti koeficijenta u usporedbi s jednostavnom regresijom (samo 1 prediktor)

Rezultat analize:

```
> f <- lm(mass ~ diet + sex, data = mds)
> summary(f)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.942      2.005   9.448 5.73e-06 ***
diethifat     12.983      2.315   5.608 0.000331 ***
sexm           1.817      2.315   0.785 0.452780
```

Sex not significant

Adapted from https://dag.compbio.dundee.ac.uk/workshops/statistics_lectures/13_Linear_models.pdf

Linearna regresija s interakcijama

- Moguće je kao prediktore uključiti i interakcije prediktorskih varijabli (npr. Ako očekujemo da prehrana ima različit utjecaj ovisno o spolu miševa)

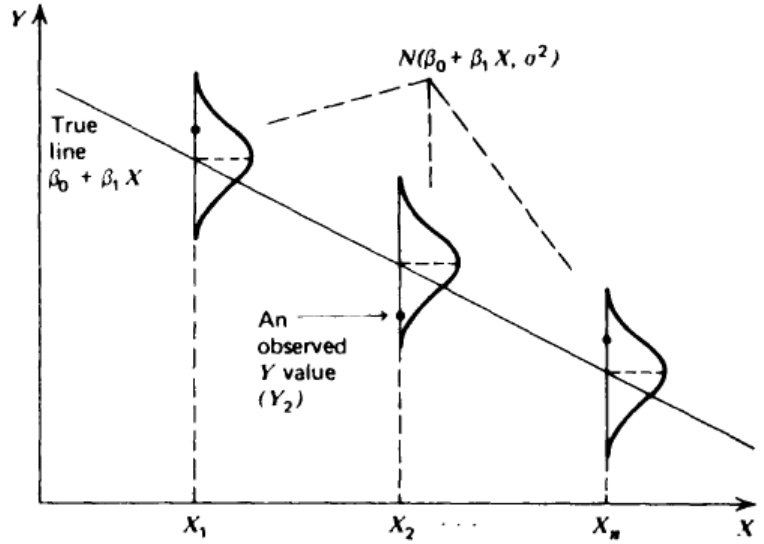
```
> f <- lm(mass ~ diet + sex + diet:sex, data = mds)
> summary(f)$coefficients
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      17.867      2.335   7.652 6.01e-05 ***
diethifat        15.133      3.302   4.583  0.00179 **
sexm              3.967      3.302   1.201  0.26400
diethifat:sexm   -4.300      4.670  -0.921  0.38407
```

Interaction not significant, we are overfitting

Adapted from https://dag.compbio.dundee.ac.uk/workshops/statistics_lectures/13_Linear_models.pdf

Pregled reziduala

- Reziduali sadrže informacije o tome zašto model možda ne odgovara podacima
- Reziduali – uočene pogreške ako je model točan
- Pretpostavke o pogreškama:
  - Pogreške su neovisne
  - Slijede normalnu distribuciju sa srednjom vrijednošću 0 i konstantnom varijancom σ2
  - Nakon ispitivanja reziduala možemo zaključiti vrijede li ove pretpostavke ili ne



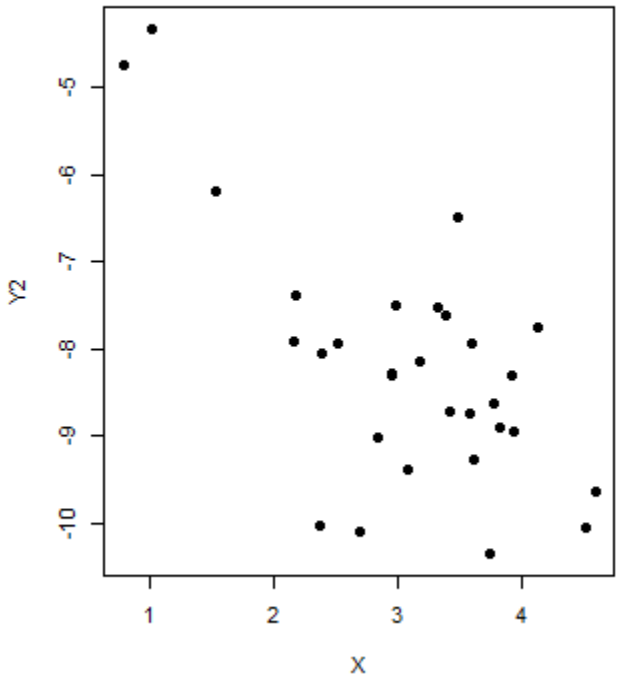
$$\varepsilon_i \sim N(0, \sigma^2)$$

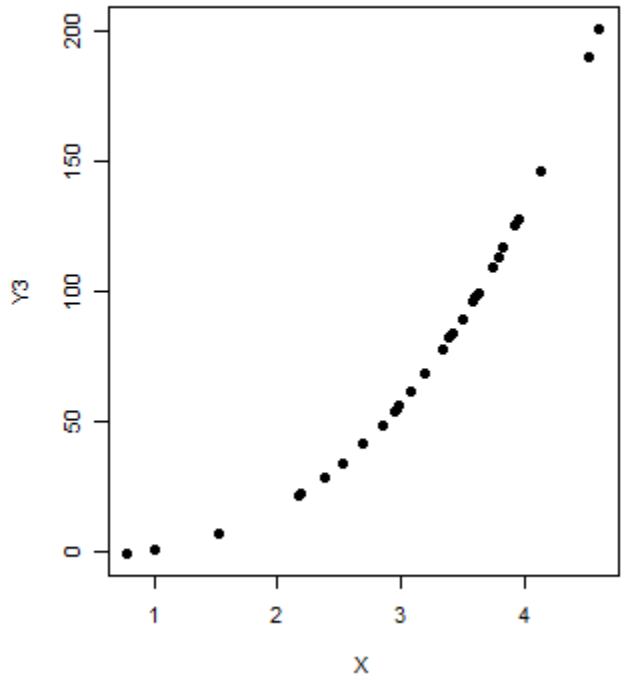
Testiranje normalnosti reziduala

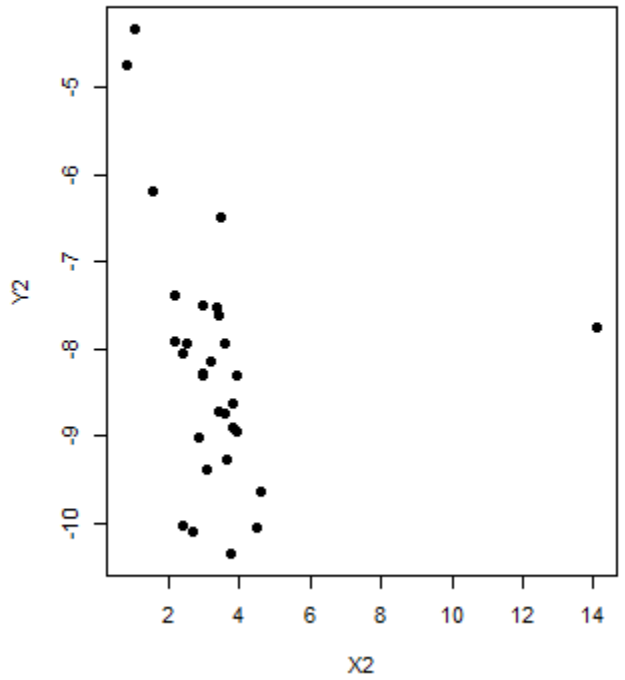
- Grafičke metode:
  - Histogram
  - Kvantil-kvantil prikaz s obzirom na standardnu normalnu distribuciju

- Testovi za normalnost
  - Anderson-Darling test, Shapiro-Wilk test, Lilliefors test (adaptacija Kolmogorov-Smirnoff testa), D'Agostino-Pearson test…

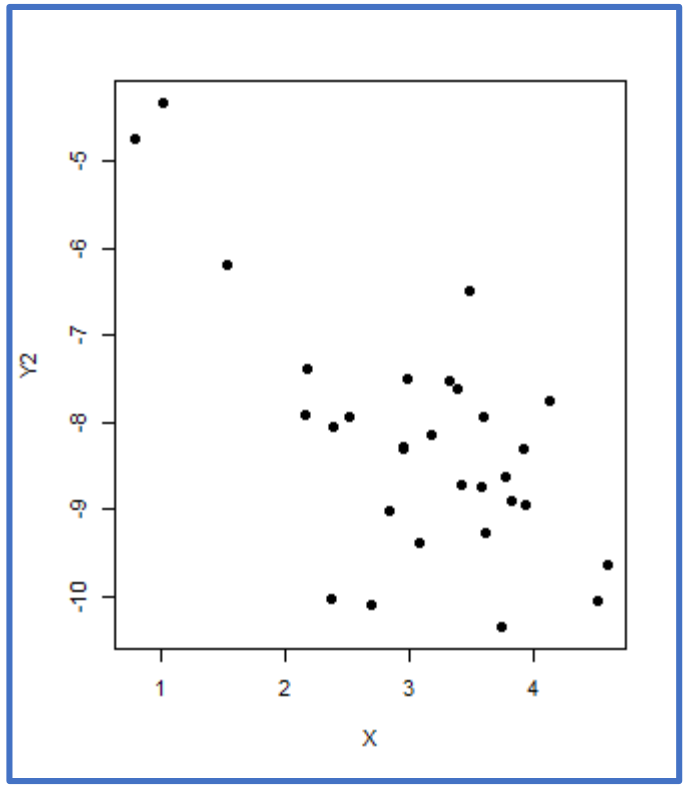
Je li linearan model prikladan za sljedeće analize?

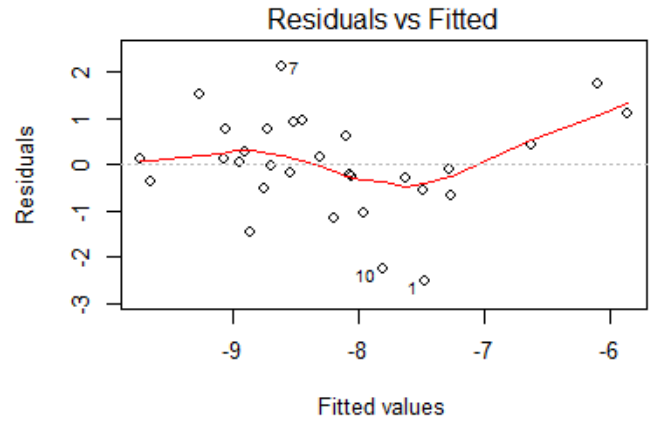


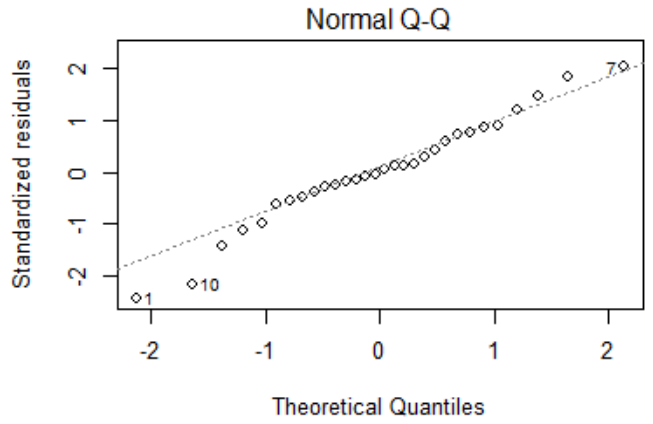


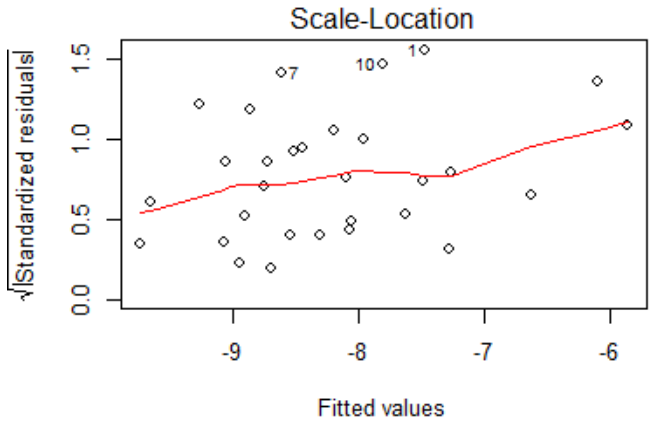


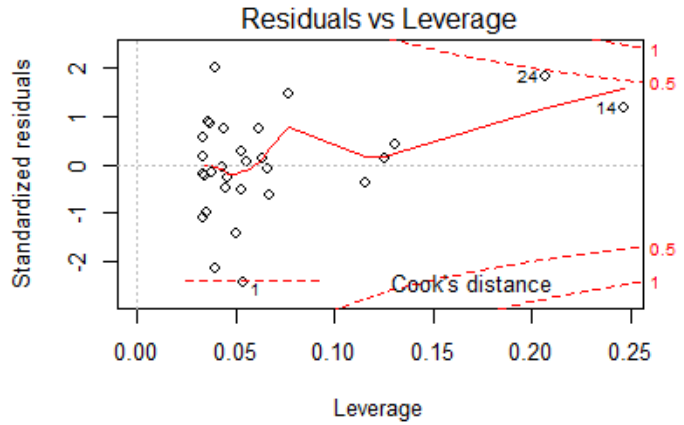
Grafički prikaz reziduala



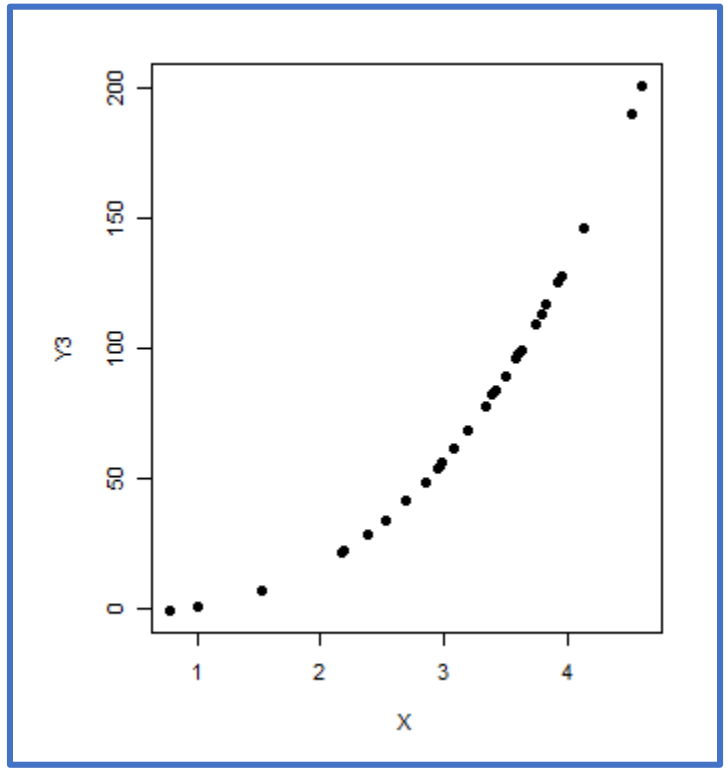


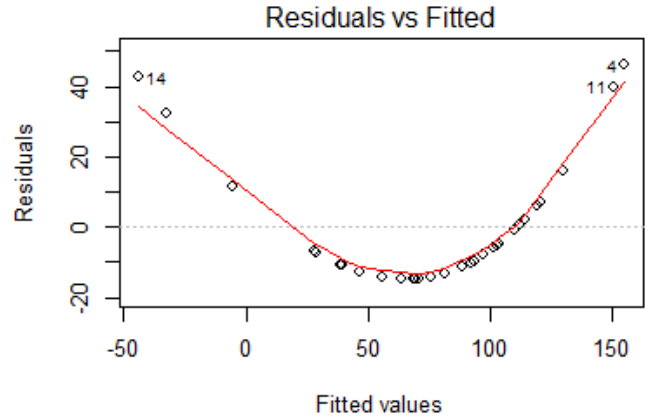


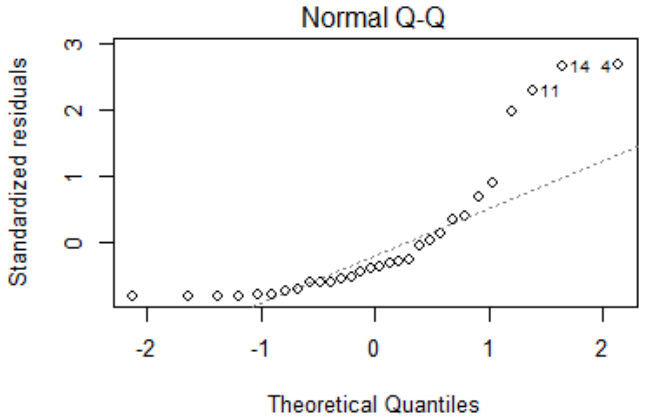


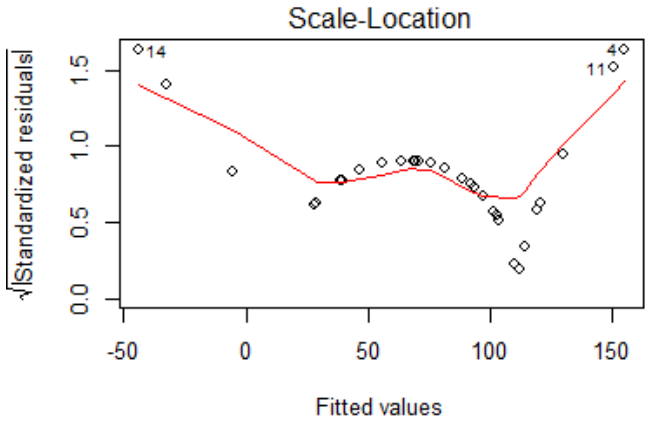


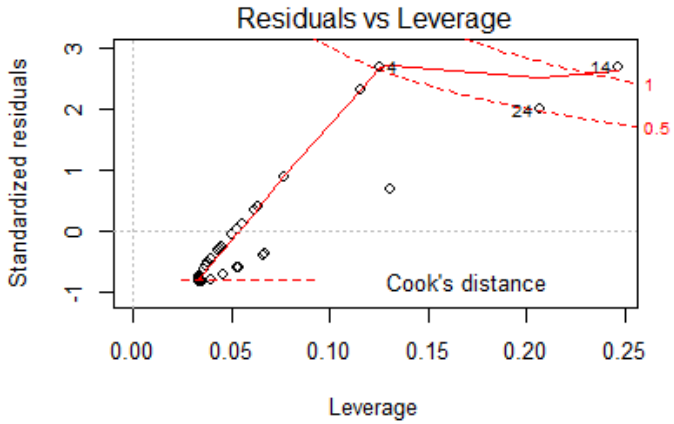
Ne-linearni odnosi



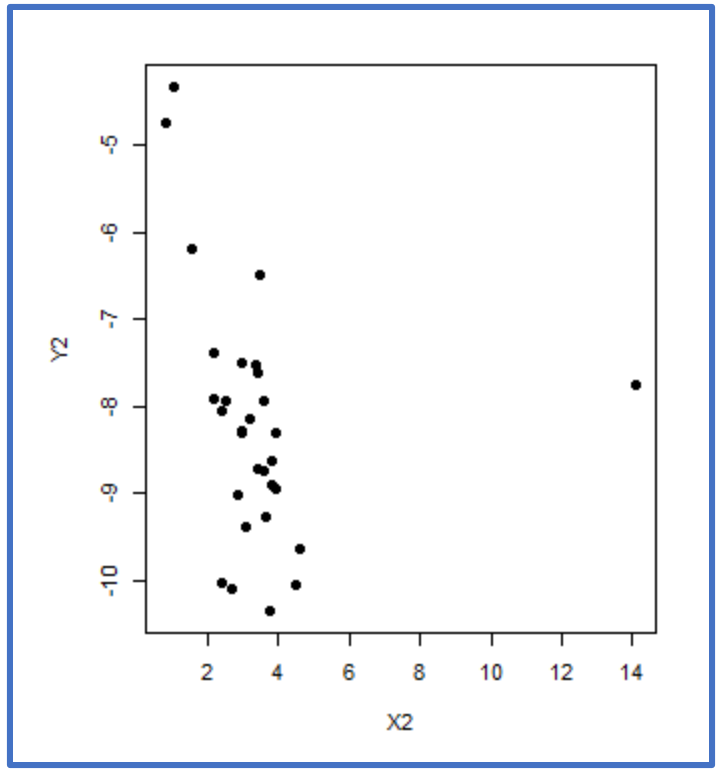


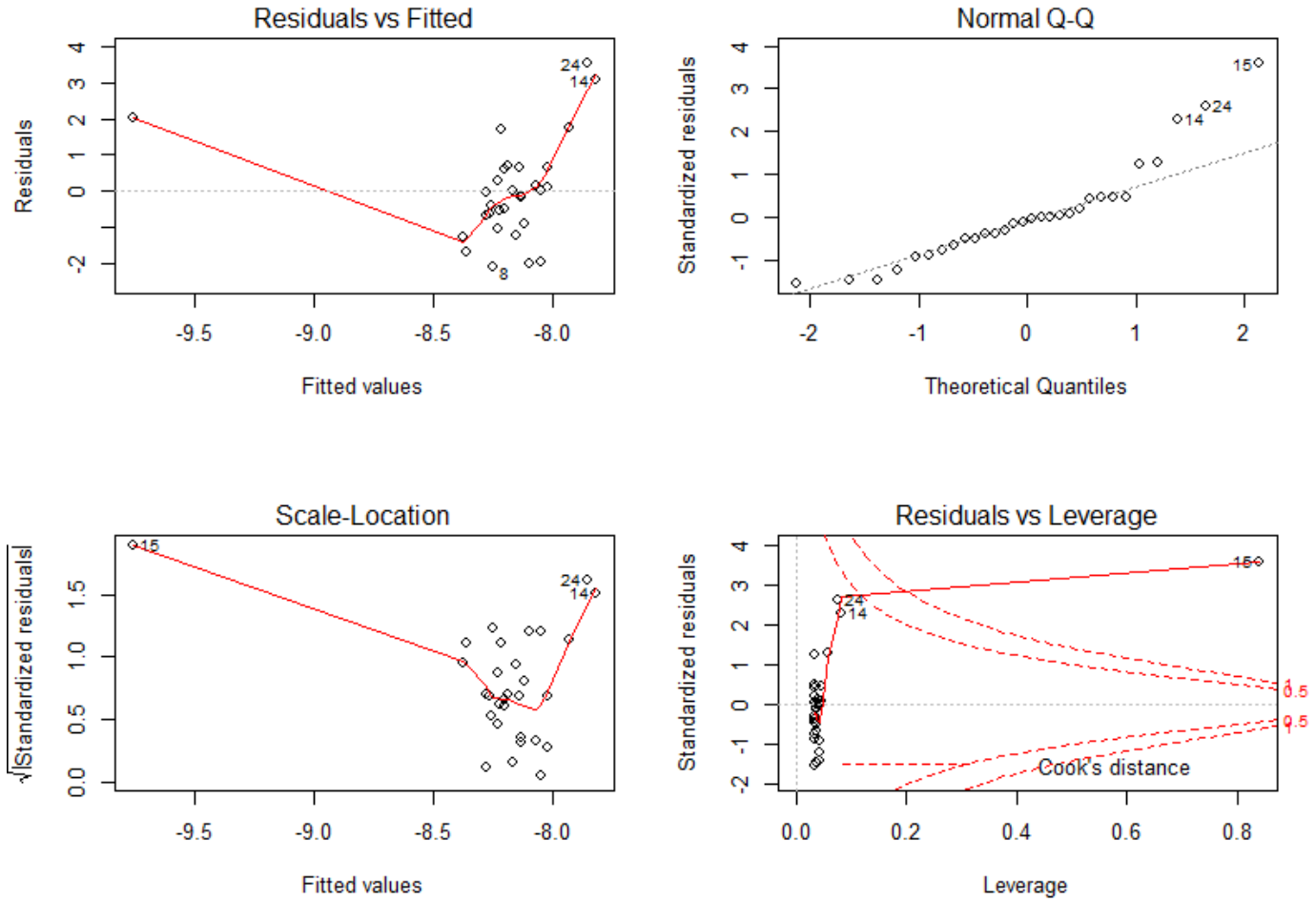






High leverage points (točke visokog utjecaja)

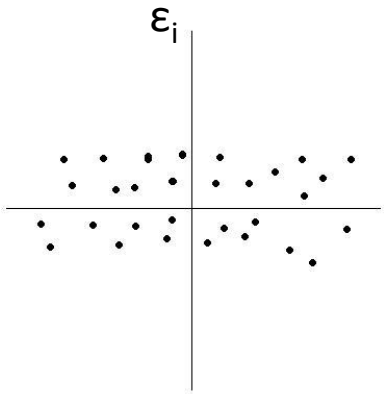




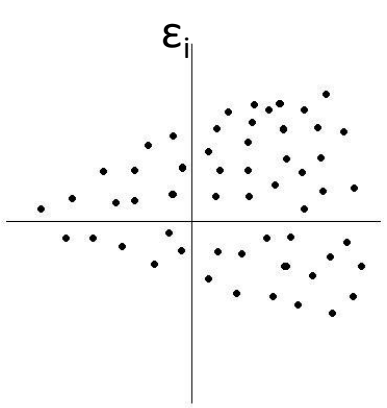
Opservacije čiji je utjecaj (leverage) prevelik (izvan iscrtanih crvenih linija) treba dodatno pregledati i potencijalno isključiti iz analize

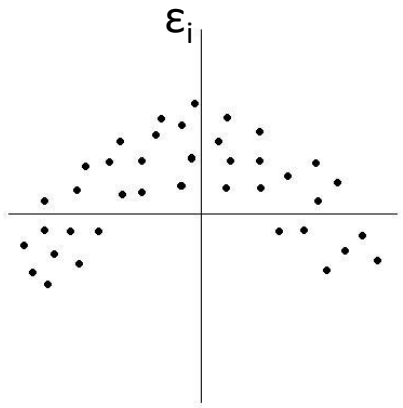
Provjeriti vremenske učinke, nestalnu varijancu, potrebu za transformacijom i zakrivljenost

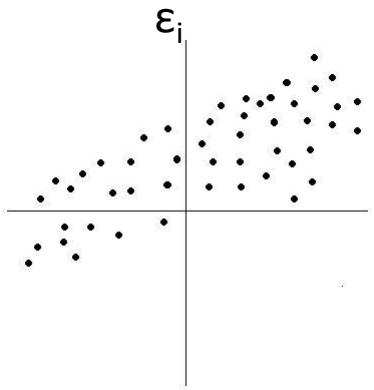
- Zadovoljavajući dijagram reziduala trebao bi pokazivati slučajni uzorak



- Nezadovoljavajući prikazi reziduala:



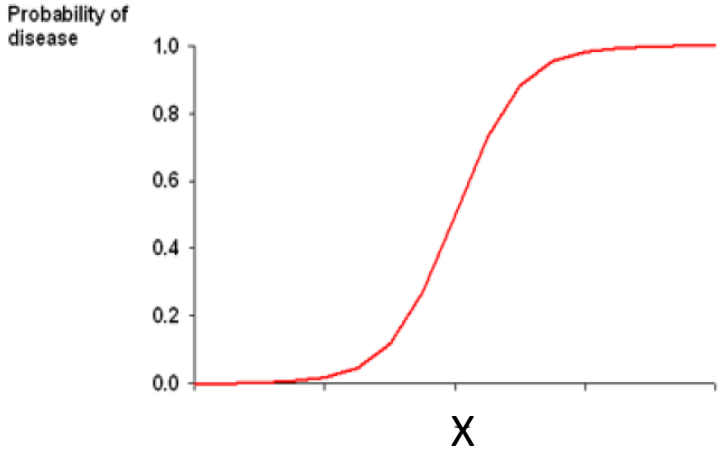




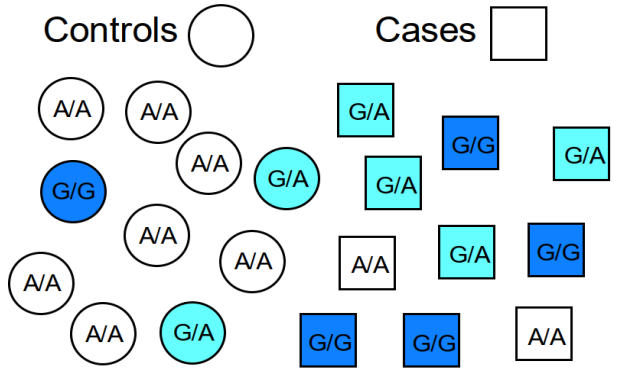
Logistička regresija

$$p(X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$

Predviđanje kategoričkih zavisnih varijabli
Npr. Želimo na temelju genotipa (G/G, G/A, A/A) boluje li ispitanik od neke bolesti (slučaj, *case*) ili ne (kontrola, *control*)
Predviđamo **p(X) = Pr(Y = 1 | X)** – vjerojatnost da ispitanik boluje od proučavane bolesti (vjerojatnost da pripada u kategoriju slučaj/*case*) s obzirom na genotip



Studija slučaja s kontrolom (Case/control study)



G alel je povezan s bolešću

Zavisna varijabla: Y (1 ako je slučaj, 0 ako je kontrola)
Prediktor: X (0, 1 ili 2 pojavljivanja alela G)

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta o + \beta_1 X + \varepsilon$$

log odds ratio

β_1 = razlika u logaritmu omjera izgleda (log odds ratio) za slučajeve u usporedbi s kontrolama

$e^{(\beta)}$ = razlika u izgledima = omjer izgleda (Odds Ratio, OR)

Efekt alela je OR:

OR > 1 povećan rizik da bude slučaj (odnosno da boluje od bolesti)

OR < 1 smanjen rizik

Adapted from https://www.colorado.edu/ibg/isg-workshop/isgw-online-resources/b1-introduction-gwas-part-1

Literatura:

Korelacija:

- Saccenti E. What can go wrong when observations are not independently and identically distributed: a cautionary note on calculating correlations on combined data sets from different experiments or conditions. Front Syst Biology. Jan. 2023;3.
- Udovičić M, Baždarić K, Bilić-Zulle L, Petrovečki M. What we need to know when calculating the coefficient of correlation?. Biochem Med (Zagreb). 2007;17:10-15

Regresija:

- Dunn, P.K. (2019) Scientific Research Methods: An introduction to quantitative research in science and health. Available at: https://srm-course.netlify.com (Accessed: January 17, 2025).
  - Poglavlje 35, regression: https://srm-course.netlify.app/regression