

Linearni modeli sa više prediktora

Primjeri – usporedbe prosjeka

I u regresiju i u ANOVA-u možemo „dodavati” prediktore

Kvantitativni prediktori - Multipla regresija

Kategorijski prediktori - Faktorska ANOVA (interakcije!!)

Kategorijski i kvantitativni prediktori – ANCOVA (razlike u trendovima)

„Slaganje” efekata – crossed ili nested efekti

Stara terminologija je zamijenjena linearnim modelima

Fiksni efekti – razine faktora su unaprijed definirane, testira se razlika upravo između tih razina (sva „ciljana” istraživanja) – tretmani – zaključci samo o uključenim razinama

Razine fiksnih efekata možemo uspoređivati (zanimaju nas prosjeci)

Slučajni efekti – razine su izvan kontrole istraživača (godine, slučajne lokacije i sl.)

Zanimaju nas varijance – smanjuju pogrešku

Multipla linearna regresija

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \varepsilon$$

$\beta_0, \beta_1, \beta_2, \dots$ - koeficijenti (intercept + za svaki prediktor u modelu)

ε - pogreška (reziduali – razlike izmjerenih i procijenjenih y vrijednosti)

Regresijski parametri (procjenitelji) – parcijalni koeficijenti (partial slopes)

b_1 - promjena y za jediničnu promjenu x_1 ako se $x_2, x_3 \dots x_n$ drže konstantnima

b_2 - promjena y za jediničnu promjenu x_2 ako se $x_1, x_3 \dots x_n$ drže konstantnima

Osim djelovanja pojedinačnih prediktora – interakcija – zajedničko djelovanje 2 ili više prediktora (npr efekt zajedničkog djelovanja prva dva prediktora – $b_3 x_1 x_2 \dots$)

Interakcija uvodi „probleme” – efekti nižeg reda (pojedinačnih prediktora) često su korelirani sa efektima višeg reda (interakcijama)

Graf više nije pravac već svaki prediktor zauzima svoju dimenziju (2 prediktora – ravnina u prostoru)

Procjene koeficijenata (suma kvadrata) su analogne jednostavnoj regresiji, ali je broj stupnjeva slobode modela jednak broju parametara koji se procjenjuju -1 (tj. broju prediktora)

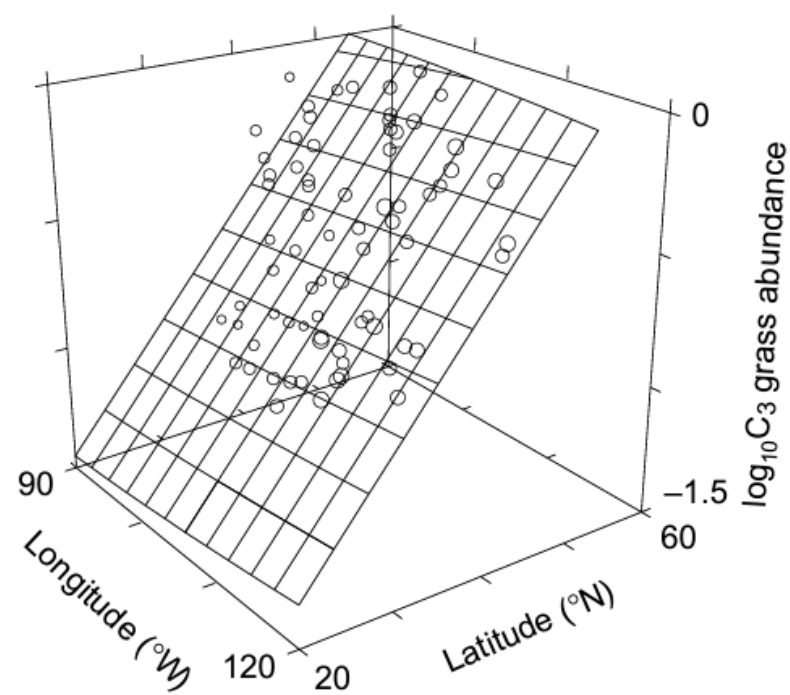


Figure 6.1 Scatterplot of the log-transformed relative abundance of C₃ plants against longitude and latitude for 73 sites from Paruelo & Lauenroth (1996) showing OLS fitted multiple regression linear response surface.

Source of variation	SS	df	MS
Regression	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	p	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{p}$
Residual	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - p - 1$	$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1}$
Total	$\sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$	

Testovi značajnosti

Testira se nulta hipoteza: $H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_j = 0$

(y se ne mijenja ako se mijenja x)

F-test (omjer varijanci)

$$F = MS_{Model} / MS_{Rezidual}$$

F se uspoređi sa F distribucijom (za $1; n-2$ stupnjeva slobode) – dobije se p – vjerojatnost prihvatanja H_0

Može se testirati i hipoteza za pojedine koeficijente (npr. $H_0: \beta_1 = 0$)

Princip: Ako je suma kvadrata modela sa b_1 dovoljno veća od sume kvadrata modela koji uključuje b_1 – možemo odbaciti H_0 ; ili t-test ($t = b_1 / s_{b1}$)

R funkcija `lm()`

Problem – kolinearnost prediktora

Prediktori su često korelirani (povezani) – problem sa inverzijom matrice X i povećanje standardnih pogrešaka procjena koeficijenata! – jako često u biologiji!

Korelacija – „povezanost” varijabli

Korelacijski koeficijent – r – ($-1 \leq r \leq 1$) – smjer (predznak) i jačina (apsolutna vrijednost) povezanosti

Scatter plot matrix – graf (scatter plot) svih parova varijabli

Tolerancija (tolerance) – niska tolerancija (manja od 0.1) – varijabla je korelirana s barem jednom drugom varijablom u modelu

Variance inflation factor (VIF) = $1 - \text{tolerancija}$

PCA – multivarijatna analiza

Primjer: skakanje insekata – veličina tijela i masa nogu

<https://mjkeough.github.io/examples/ercit.nb.html>

Bird abundance - svašta

<https://mjkeough.github.io/examples/loyn.nb.html>

ANOVA (višefaktorska) – prediktori su kategorijski

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

α – učinak faktora A

β – učinak faktora B

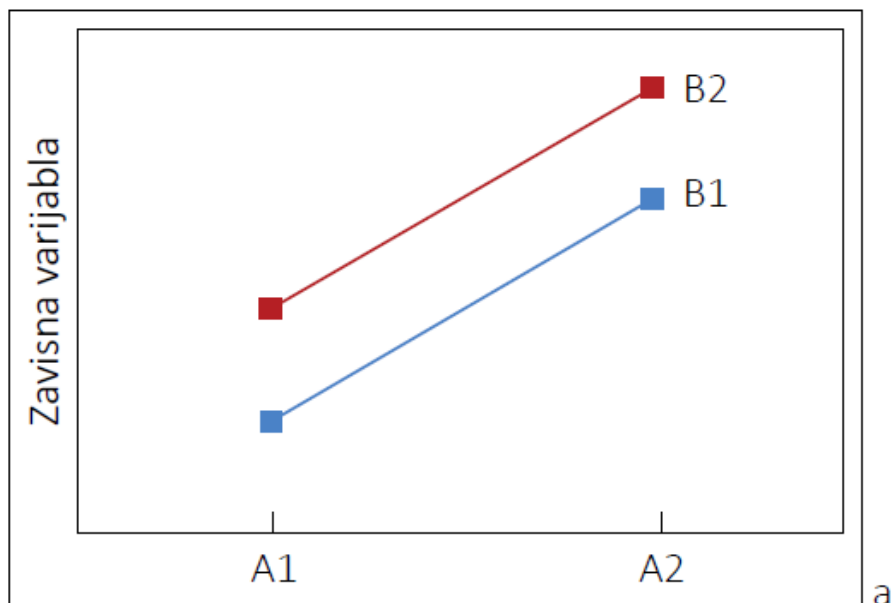
$\alpha\beta$ – učinak interakcije

Interakcija – zajedničko djelovanje faktora (neaditivno). Ako se ne dokaže – faktori A i B djeluju neovisno (aditivno)

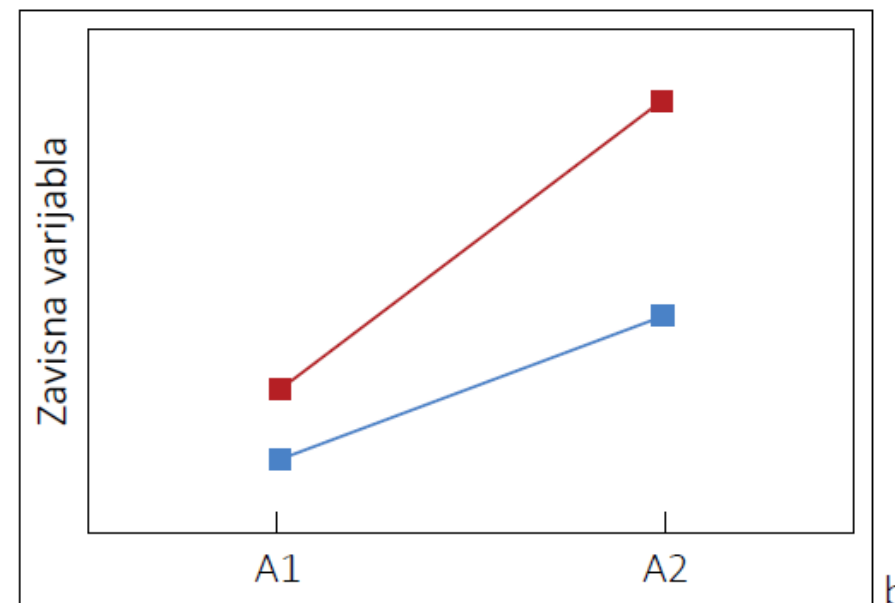
Dodatni faktori – dodatni izvori varijabilnosti („troše” stupnjeve slobode)

Elementi dizajna (repeticije *sensu stricto*, blokovi, nepotpuni blokovi...) se analiziraju kao dodatni faktori

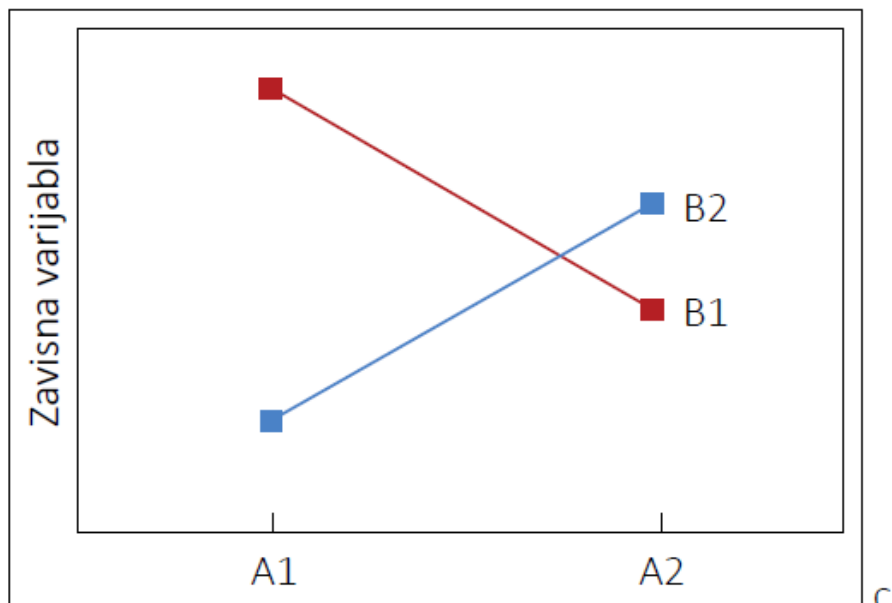
2 faktora (A i B)
Svaki sa 2 razine
(A1, A2 i B1, B2)



a



b



c

Slika 3.1 Shematski prikaz interakcije:
a) nema interakcije, b) interakcija bez promjene poretka,
c) interakcija s promjenom poretka.

Izvori varijabilnosti	df	SS	MS	F	EMS
Ukupno	$rab - 1$	SS_{UKUPNO}			
A	$a - 1$	SS_A	$\frac{SS_A}{a - 1}$	$\frac{MS_A}{MS_{ERR}}$	$\sigma^2 + rb\theta_A^2$
B	$b - 1$	SS_B	$\frac{SS_B}{b - 1}$	$\frac{MS_B}{MS_{ERR}}$	$\sigma^2 + ra\theta_B^2$
A·B	$(a - 1)(b - 1)$	SS_{AB}	$\frac{SS_{AB}}{(a - 1)(b - 1)}$	$\frac{MS_{AB}}{MS_{ERR}}$	$\sigma^2 + r\theta_{AB}^2$
Reziduali	$ab(r - 1)$	SS_{ERR}	$\frac{SS_{ERR}}{ab(r - 1)}$		σ^2

Testiraju se hipoteze o učinku faktora A, učinku faktora B i učinku interakcije (A*B)

Ako se dokaže interakcija – samo ona se interpretira

Testira se uvijek sa MS pogreške (reziduala)

Primjer – analiza varijance (2 prediktora)

Na 3 sorte lješnjaka je kroz 3 godine mjerena masa ploda. Potrebno je tretirati razlikuje li se broj plodova između sorti i između godina

```
> ljesnjak$god <- as.factor(ljesnjak$godina)
> brpl.anova <- lm(broj_plodova~god*sorta,data=ljesnjak)
> summary(brpl.anova)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
god	2	8114224	4057112	25.630	2.38e-09	***
sorta	2	7111054	3555527	22.462	1.74e-08	***
god:sorta	4	6178548	1544637	9.758	1.71e-06	***
Residuals	81	12821810	158294			

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

sorta	gram	godina	broj_plodova
HALESKI	1	2003	2711
HALESKI	2	2003	2565
HALESKI	3	2003	2431
HALESKI	4	2003	2320
HALESKI	5	2003	2516
HALESKI	6	2003	1748
HALESKI	7	2003	1891
HALESKI	8	2003	1254
HALESKI	9	2003	1802
HALESKI	10	2003	2298
ISTARSKI	1	2003	1393
ISTARSKI	2	2003	1400
ISTARSKI	3	2003	1517
ISTARSKI	4	2003	1277
ISTARSKI	5	2003	1843

```

> ljesnjak$god <- as.factor(ljesnjak$godina)
> brpl.anova <- lm(broj_plodova~god*sorta,data=ljesnjak)
> summary(brpl.anova)

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
god	2	8114224	4057112	25.630	2.38e-09	***
sorta	2	7111054	3555527	22.462	1.74e-08	***
god:sorta	4	6178548	1544637	9.758	1.71e-06	***
Residuals	81	12821810	158294			

```

---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

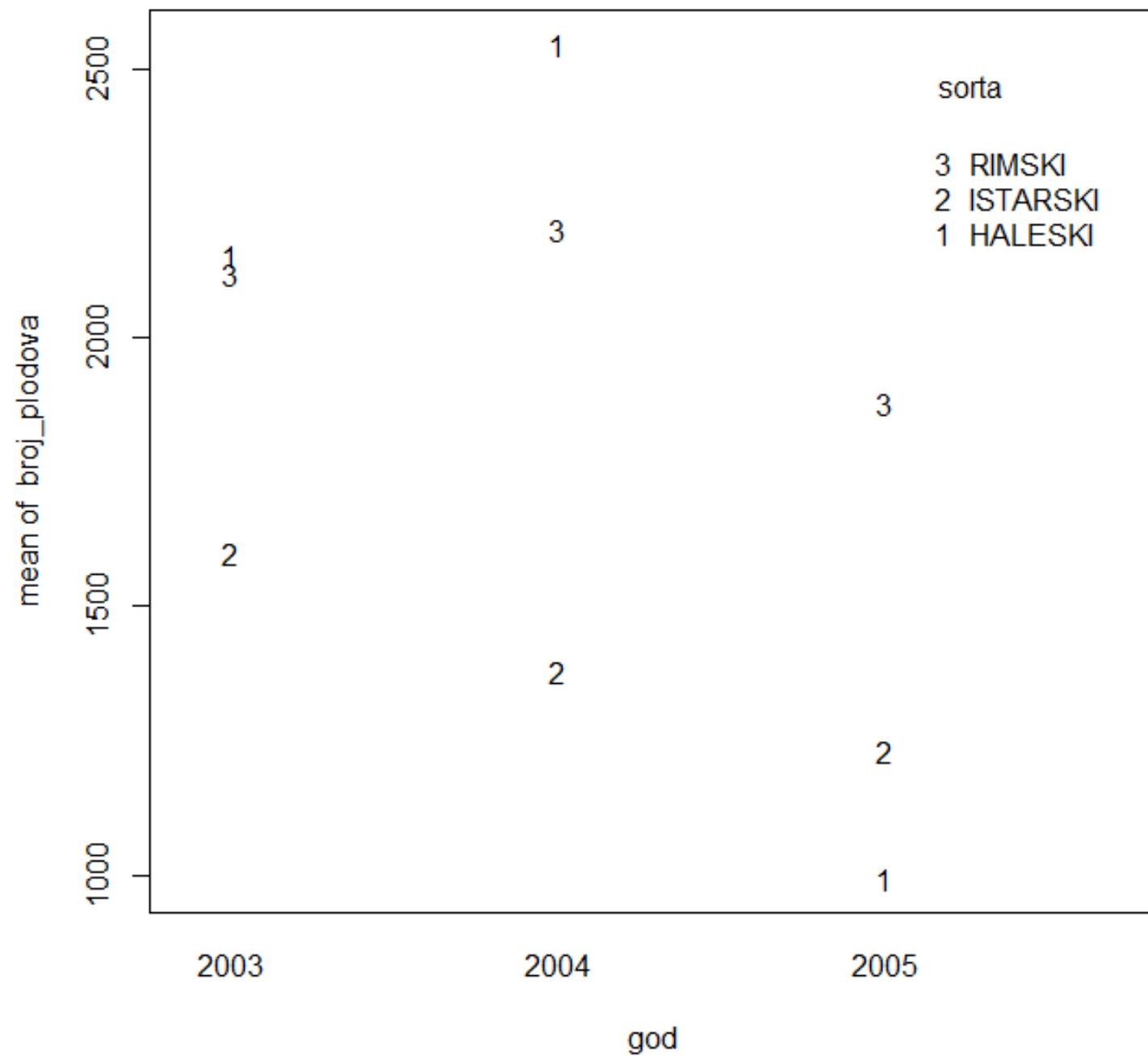
Interakcija je signifikantna – sorte ne daju jednaki broj plodova u različitim godinama (i redosljed im je možda drugačiji u različitim godinama)

Usporedbe prosjeka za interakciju – razlike prosjeka kombinacija razina faktora, često se rade samo usporedbe unutar jedne razine jednog faktora

Npr: uspoređuje se broj plodova na 3 sorte unutar svake od 3 godine

Interaction plot

Jednostavan prikaz
interakcije



```
library(emmeans)
brpl.lm <- lm(broj_plodova~god*sorta,data=ljesnjak)
emmip(brpl.lm, sorta~god)
```

```
emm_s.t <- emmeans(brpl.lm, pairwise ~ sorta | god)
emm_s.t
```

\$contrasts

god = 2003:

contrast	estimate	SE	df	t.ratio	p.value
HALESKI - ISTARSKI	553.5	178	81	3.111	0.0072
HALESKI - RIMSKI	36.6	178	81	0.206	0.9770
ISTARSKI - RIMSKI	-516.9	178	81	-2.905	0.0130

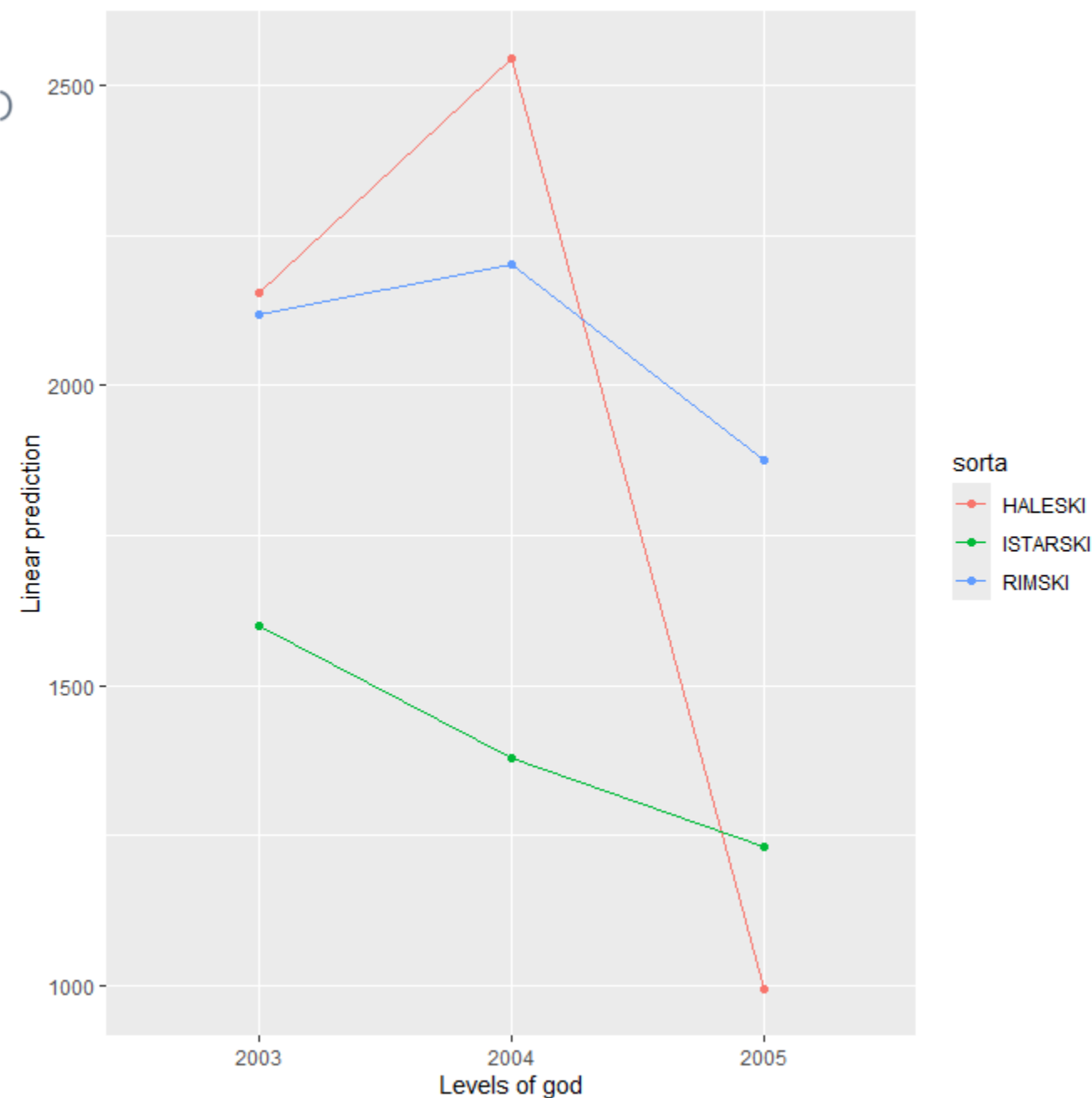
god = 2004:

contrast	estimate	SE	df	t.ratio	p.value
HALESKI - ISTARSKI	1167.5	178	81	6.562	<.0001
HALESKI - RIMSKI	343.8	178	81	1.932	0.1362
ISTARSKI - RIMSKI	-823.7	178	81	-4.629	<.0001

god = 2005:

contrast	estimate	SE	df	t.ratio	p.value
HALESKI - ISTARSKI	-237.0	178	81	-1.332	0.3817
HALESKI - RIMSKI	-882.7	178	81	-4.961	<.0001
ISTARSKI - RIMSKI	-645.7	178	81	-3.629	0.0014

P value adjustment: tukey method for comparing a family of 3 estimates



Model sa kontinuiranim i kategorijskim prediktorima – ANCOVA

Analiza kovarijance (kontinuirani prediktor – kovarijat)

Namjena – „uklanjanje” nepoželjnog efekta kontinuiranog prediktora (confoundera) koji utječe na vrijednost ishoda ali nije primarni interes istraživanja

Ili

Usporedba nagiba (slopes) i intercepata više regresijskih pravaca

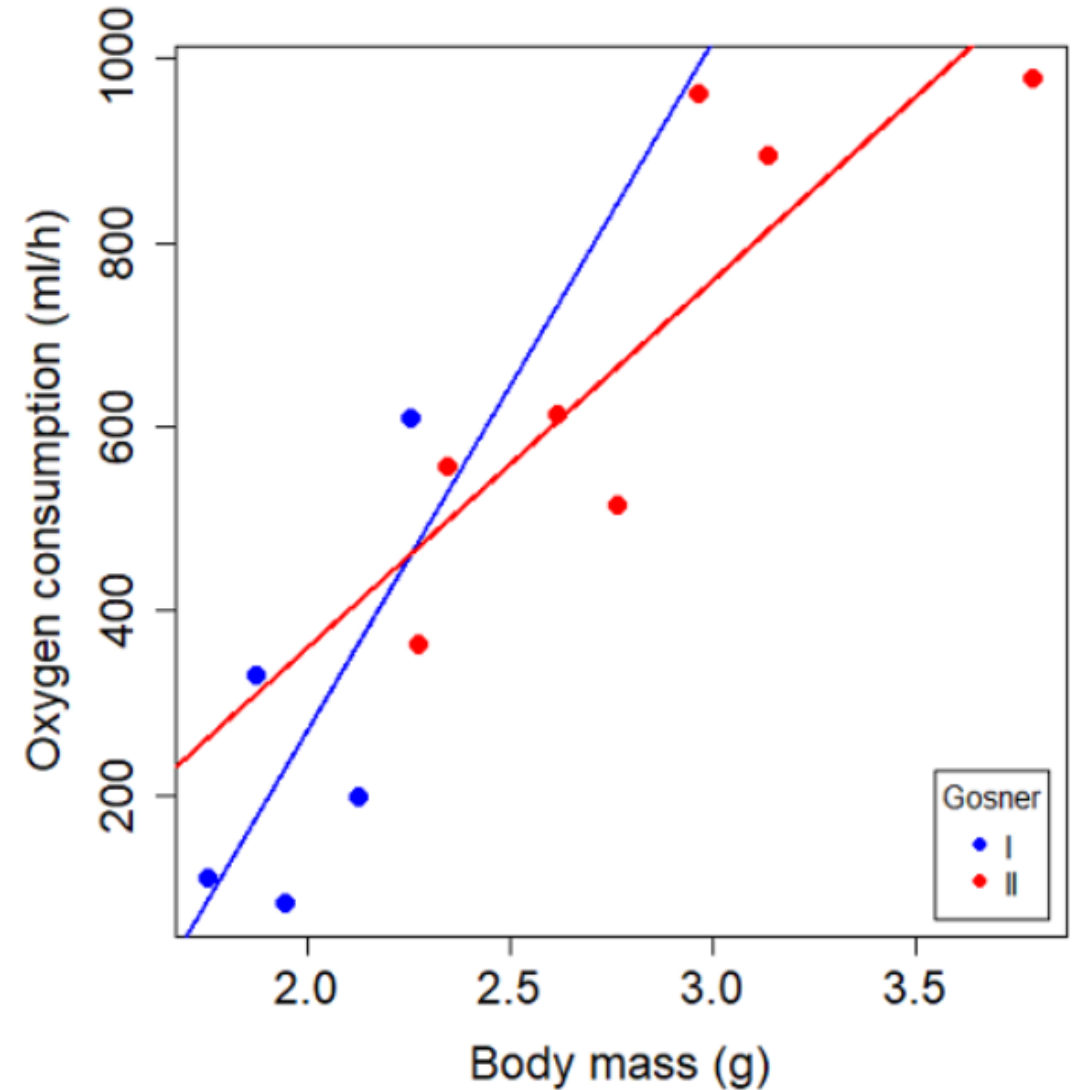
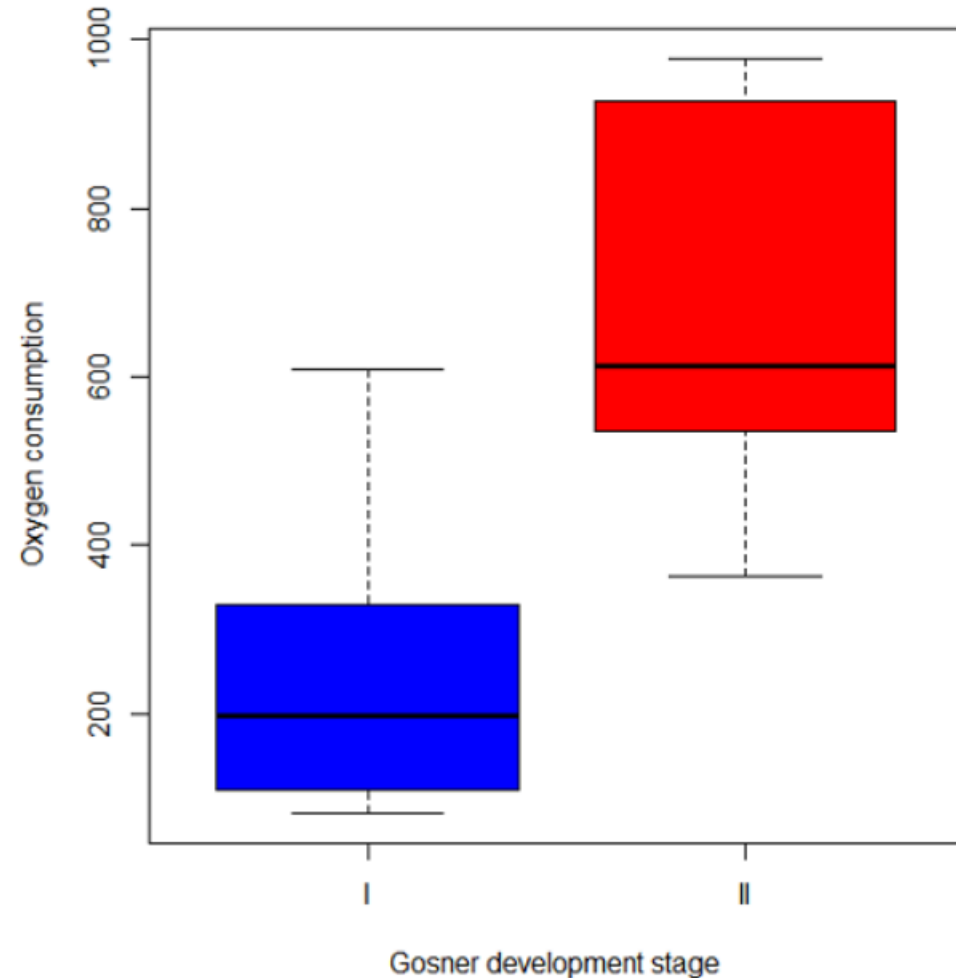
$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}$$

Gornji model ne uključuju interakciju, pretpostavlja jednake nagibe regresijskih pravaca za sve grupe prediktora A

Signifikantna interakcija kontinuiranog i kategorijskog prediktora B – nagibi regresijskih pravaca su nejednaki između grupa kategorijskog prediktora A

Primjer: Razlika u potrošnji kisika kod punoglavaca između 2 stadija

Masa bi mogla utjecati?



Primjer: vidre

<https://dzchilds.github.io/stats-for-bio/two-way-ancova-in-r.html>

mušice

<https://mjkeough.github.io/examples/partridge.nb.html>

Ježinci

<https://mjkeough.github.io/examples/constable.nb.html>