Razvoj statističnih modelov - po izvedbi poskusa

Milena Kovač

2. januar 2013

Preverimo prvih 8 korakov

- Je poskus potekal po načrtu?
- Preverimo strukturo podatkov:
  - število podatkov po razredih, število neznanih parametrov in stopinj prostosti
  - dodatni vplivi
  - spremenjeni pogoji ...
- Po potrebi popravimo osnovni in možni model!

9. Statistična ovrednotenje

Kriterij: verjetnost (p-vrednost), da drži ničelna hipoteza

Ničelna hipoteza: ni razlik med nivoju pri izbranem vplivu

Alternativna hipoteza: nivoji se med seboj razlikujejo
najmanj en nivo je različen od drugih

Verjetnost:

0.0	$\Leftrightarrow$	1.0
ničelna hipoteza ovržena	$\Leftrightarrow$	ničelna hipoteza drži
razlike so, vsaj ena	$\Leftrightarrow$	razlik ni, nobene

PRIMER A

Vpliv skupine: $P = 0.04$

- ničelna hipoteza drži samo v 4 % primerih
- alternativna hipoteza velja v 96 % primerih
- vpliv skupine je v poskusu najverjetneje precej pomemben in ga ne smemo izpustiti iz modela
- izjema: **je ni!**

PRIMER B

Vpliv krme: $P = 0.84$

- ničelna hipoteza drži kar v 84 % primerih
- alternativna hipoteza velja v 16 % primerih
- vpliv krme je v izvedenem poskusu najverjetneje popolnoma nepomemben in ga smemo izpustiti iz modela
- izjema: **vpliv krme je osrednji cilj naloge**

PRIMER C

Vpliv spola: $P = 0.45$

- ničelna hipoteza drži le v 45 %
- alternativna hipoteza velja v 55 %
- vpliv spola v izvedenem poskusu ni dovolj prepričljiv, rezultat je precej neodločen, in vpliv smemo izpustiti iz modela
- izjema: **vpliv spola je osrednji cilj naloge**

PRIMER D

Vpliv spola: $P = 0.65$

- ničelna hipoteza drži le v 65 % primerov
- alternativna hipoteza velja v 35 % primerov
- vpliv spola v izvedenem poskusu ni dovolj prepričljiv, rezultat je precej neodločen, in vpliv smemo izpustiti iz modela
- izjema: **vpliv spola je osrednji cilj naloge**

PRIMER E

Vpliv mase: $P = 0.08$

- ničelna hipoteza drži samo v 8 % primerov
- alternativna hipoteza velja v 92 % primerov
- vpliv mase v izvedenem poskusu ni dovolj prepričljiv, a je na meji. Razlik med nivoji ni ali niso dokazane. Morda smo imeli premalo opazovanj ...
- nakazuje se trend, da se lastnost z maso spreminja
- vpliva mase ne izpustitimo iz modela
- izjema: **zmanjkuje stopinj prostosti**

Kriteriji za izločanje vplivov

P-vrednost	Storimo:	Izjema, kadar
0.00 - 0.05	vedno obdržimo	—
0.05 - 0.40	gotovo obdržimo	zmanjkuje stopinj prostosti za ostanek
0.40 - 0.60	gotovo črtamo	vsi pričakujejo vpliv
0.60 - 1.00	vedno črtamo	je cilj raziskave

Kriteriji za izbor modela

- statistična značilnost vplivov
- strokovna presoja (število opazovanj in število parametrov, pričakovanja)
- primerjava modela z in brez posameznimi vplivi
- najprej izločamo "podrejene" vplive: interakcije, ugnezdene vplive
- obdržimo vse parametre, ki tvorijo celoto
  - regresije: presečišča z ordinato in regresijski koeficienti
  - laktacijske krivulje ...

Vplive izločamo postopoma in raje ponovimo izračun.

Postopek ponavljamo, dokler se model spreminja...

Če črtamo značilni sistematski vpliv ...

- rezultati so neočiščeni
- rezultati so pristranski in "frizirani", ponarejeni
- več nepojasnjene variance, pomembni rezultati so manj očitni
- **tega ne smemo nikdar storiti!**
- črtanje naključnega vpliva:
  - samo poveča varianco (neodvisni nivoji)
  - tudi pristranske ocene (odvisni nivoji - sorodstvo)

Če obdržimo nepomembni vpliv ...

- porabimo več stopinj prostosti
- rezultati niso napačni, se ne smejo razlikovati pomembno od rezultatov z modelom brez nepomembnega vpliva
- obrazložiti moramo, zakaj je v modelu nekaj nepomembnega - nejasni komentarji
- raje izključimo iz modela z besedami:
  - na osnovi predhodnih obdelav smo vpliv izključili iz nadaljnih analiz

Deduktivni postopek

možni model

$\Downarrow$

postopoma izločamo vplive z nepomembnimi vplivi

$\Downarrow$

optimalni model

$\Downarrow$

osnovni model

$$y_i = \mu + e_i$$

Induktivni postopek

osnovni model

$\Downarrow$

postopoma dodajamo vplive in jih preizkušamo

izpustimo nepomembne vplive

$\Downarrow$

optimalni model

$\Downarrow$

teoretični model

Pri tem postopku se nam lahko zgodi, da katerega od pomembnih, značilnih vplivov ali kombinacije ne preizkusimo.

Prilagodljivost modela

= kako dobro se model prilagaja podatkom (ang. fit of the model)

- več parametrov $\Rightarrow$večji delež pojasnjene variance
  $\Rightarrow$bolj se bo model prilegal

- število parametrov = število podatkov
  - model se prilagaja popolnoma
  - z modelom nismo nič pridobili
  - model ni uporaben (zakon skromnosti)

- skrbno načrtovan in izveden poskus
  $\Rightarrow$večji delež pojasnjene variance

10. Strokovna presoja

Vsak model mora biti

- statistično utemeljen in strokovno interpretiran

Interpretacija

- opravimo v istem vrstnem redu, kot so vplivi v modelu
- vse značilne vplive moramo prikazati in utemeljiti
- nepričakovani rezultati:
  - preverimo model
  - iščimo smiselno razlago (ni vse zanič, kar je v nasprotju z dosedanjim znanjem)
  - morda je možno razrešiti nejasnost samo z novim poskusom
- **potrebna ponovitev?**

11. Proces izgradnje modela

- izgradnja modela ni enkraten proces.

- Vračamo se:

  - na statistično ovrednotenje ali
  - povsem na začetek, na listo vplivov ali celo
  - iskanje dodatnih informacije v katerem od obstoječih informacijskih sistemov

POMNI! Pri presoji modela ne smemo absolutno zagovarjati svoje teorije - hipoteze. Poskusimo se vživeti v vlogo kritika, osvetlimo tudi druge plati medalje!

Krivulje: kompleti parametrov

- linearna enačba: parametri v parih

$$\cdots + M_j + b_j\,(x_{ijklm} - 100) + \cdots$$

$$\cdots + PM_{ij} + b_{ij}\,(x_{ijklm} - 100) + \cdots$$

- kvadratna enačba: trije parametri za vsako parabolo

$$\cdots + PM_{ij} + b_{Iij}x_{ijklm} + b_{IIij}x^2_{ijklm} + \cdots$$

- izpustimo lahko kvadratni člen, če ni značilen
  $\Rightarrow$linearna regresija

- za vse nivoje imamo iste enačbe

Nezaželene kombinacije

- črtamo presečišče premic z ordinato
  $\Rightarrow$ vse premice skozi isto točko na ordinati

$$\cdots + \cancel{M_j} + b_j x_{ijklm} + \cdots$$

- tudi v naslednjih enačbah parametri niso v kompletu

- kvadratnih členov je manj

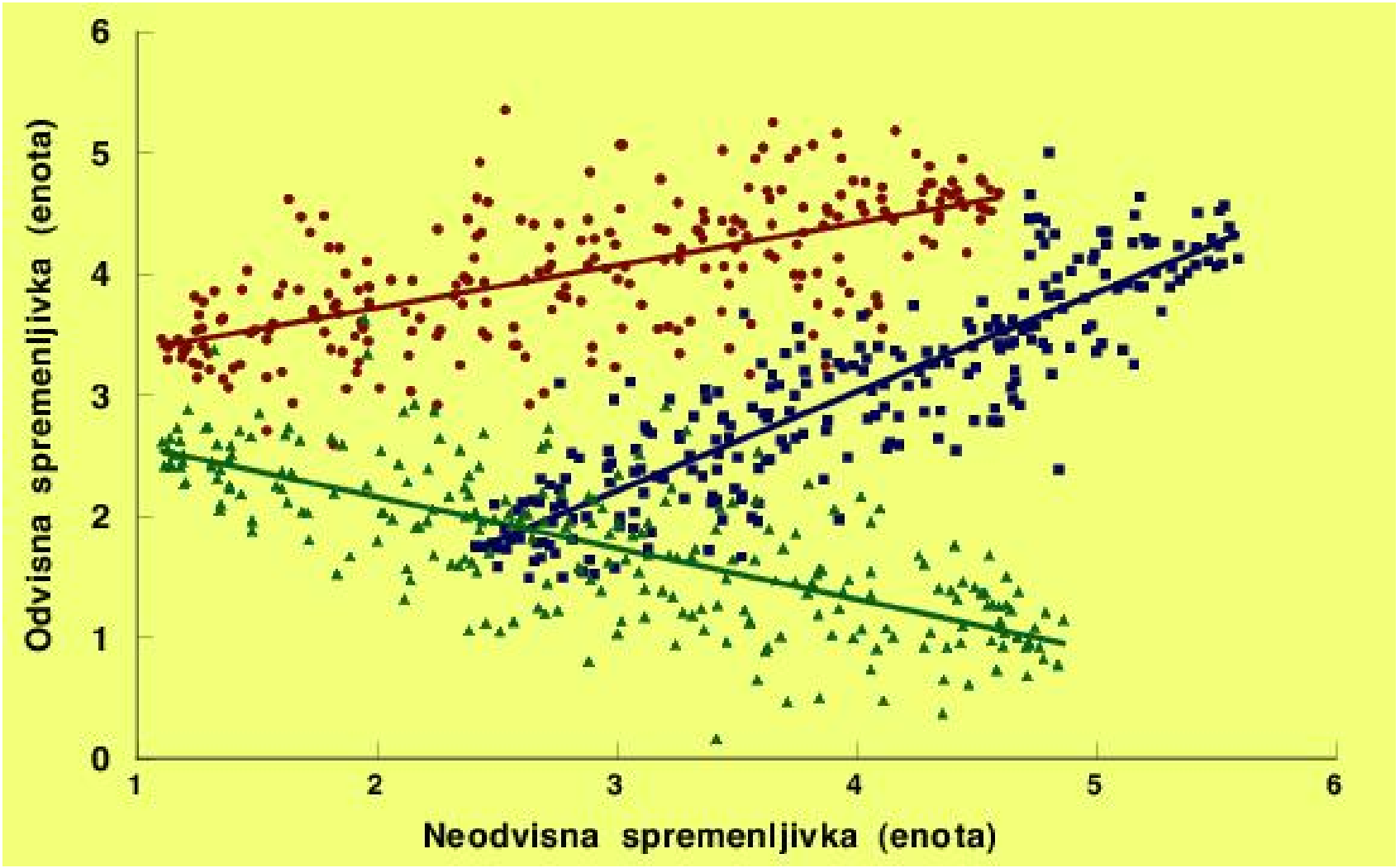
$$\cdots + PM_{ij} + b_{Iij} x_{ijklm} + b_{IIj} x^2_{ijklm} + \cdots$$

- linearnih členov je več

$$\cdots + M_j + b_{Iij} x_{ijklm} + b_{IIj} x^2_{ijklm} + \cdots$$

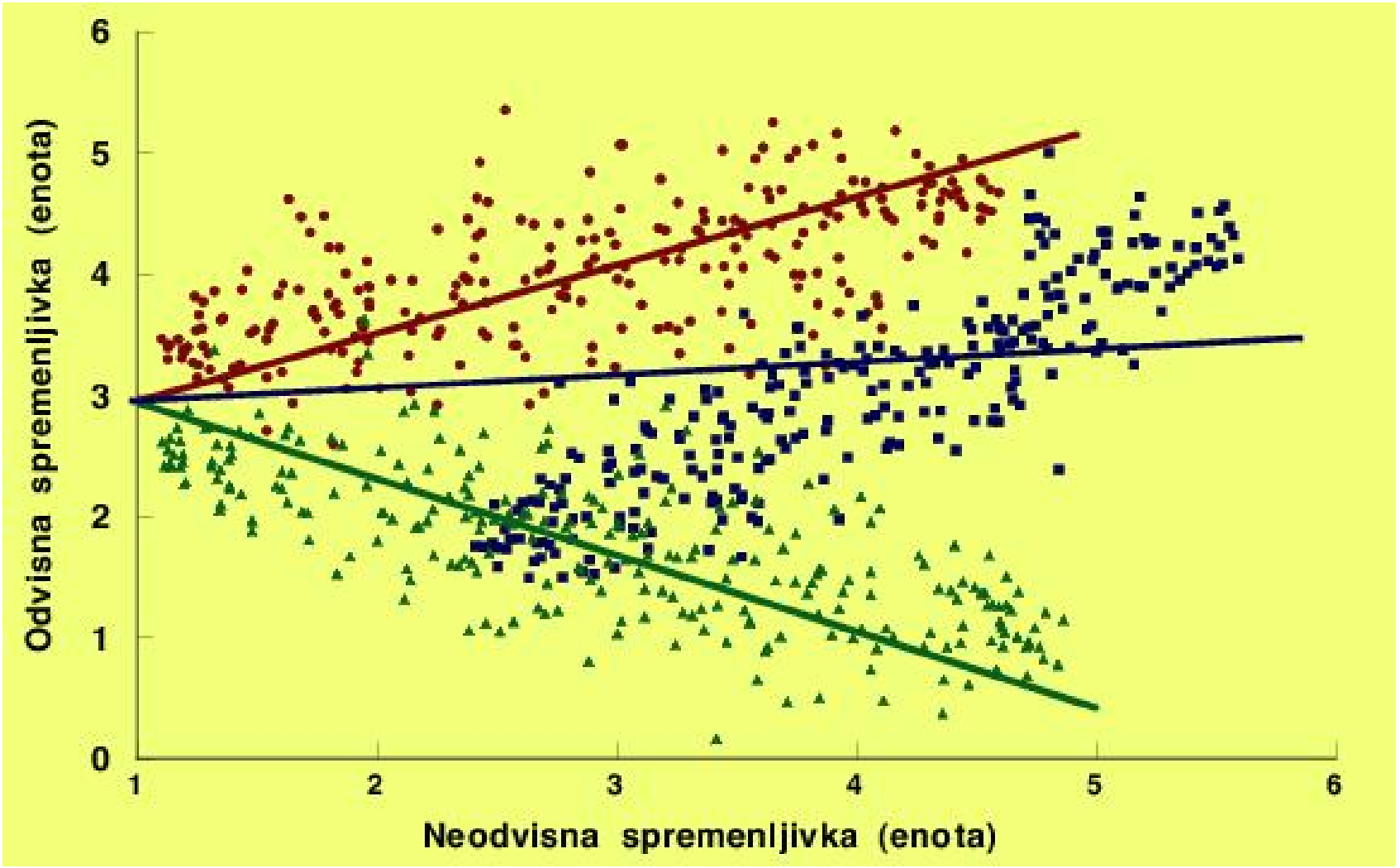
Primer 1

$$\cdots + M_j + b_j(x_{ijklm} - 1) + \cdots$$



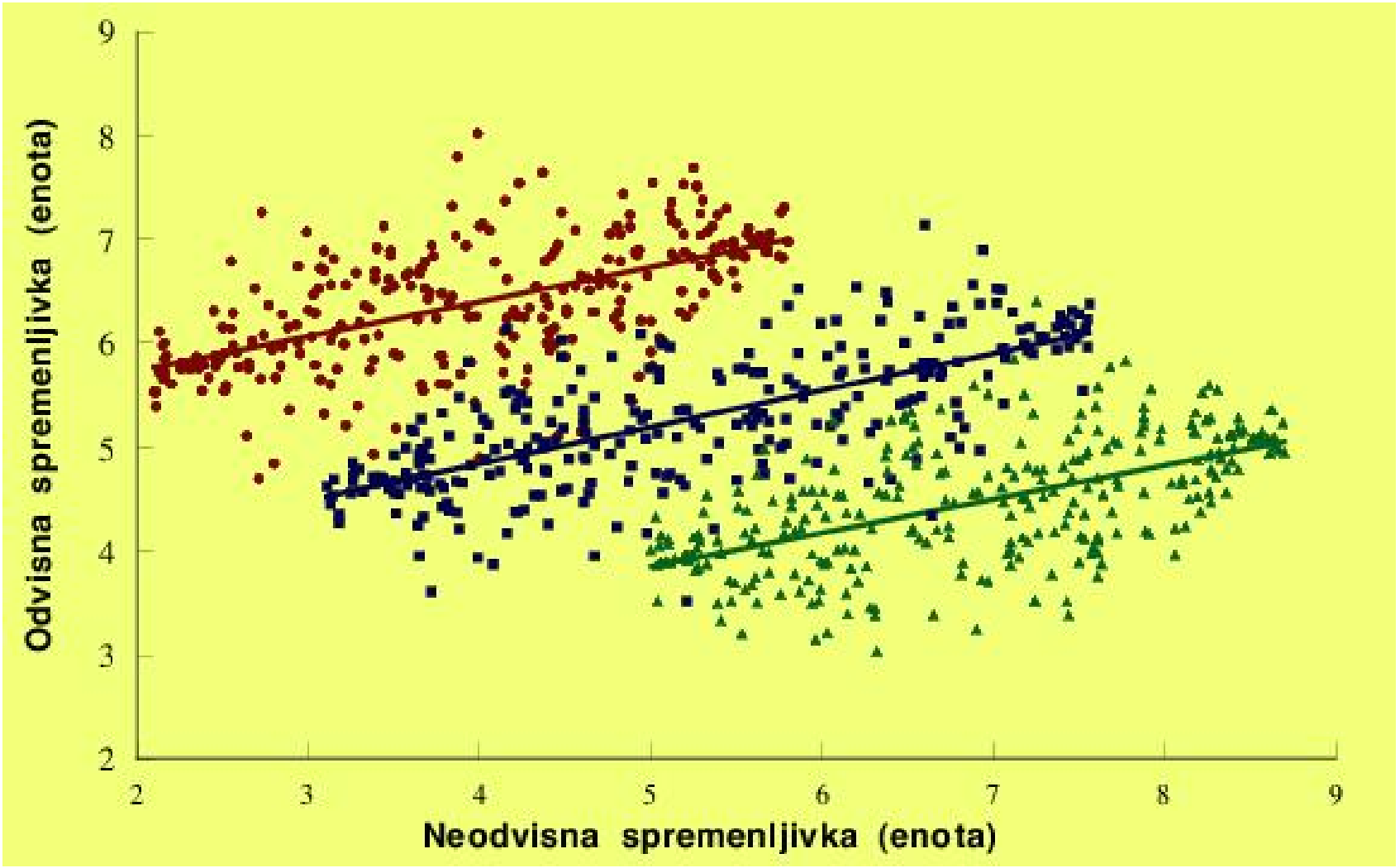
Primer 2

$$\cdots + \cancel{M}_j + b_j(x_{ijklm} - 1) + \cdots$$



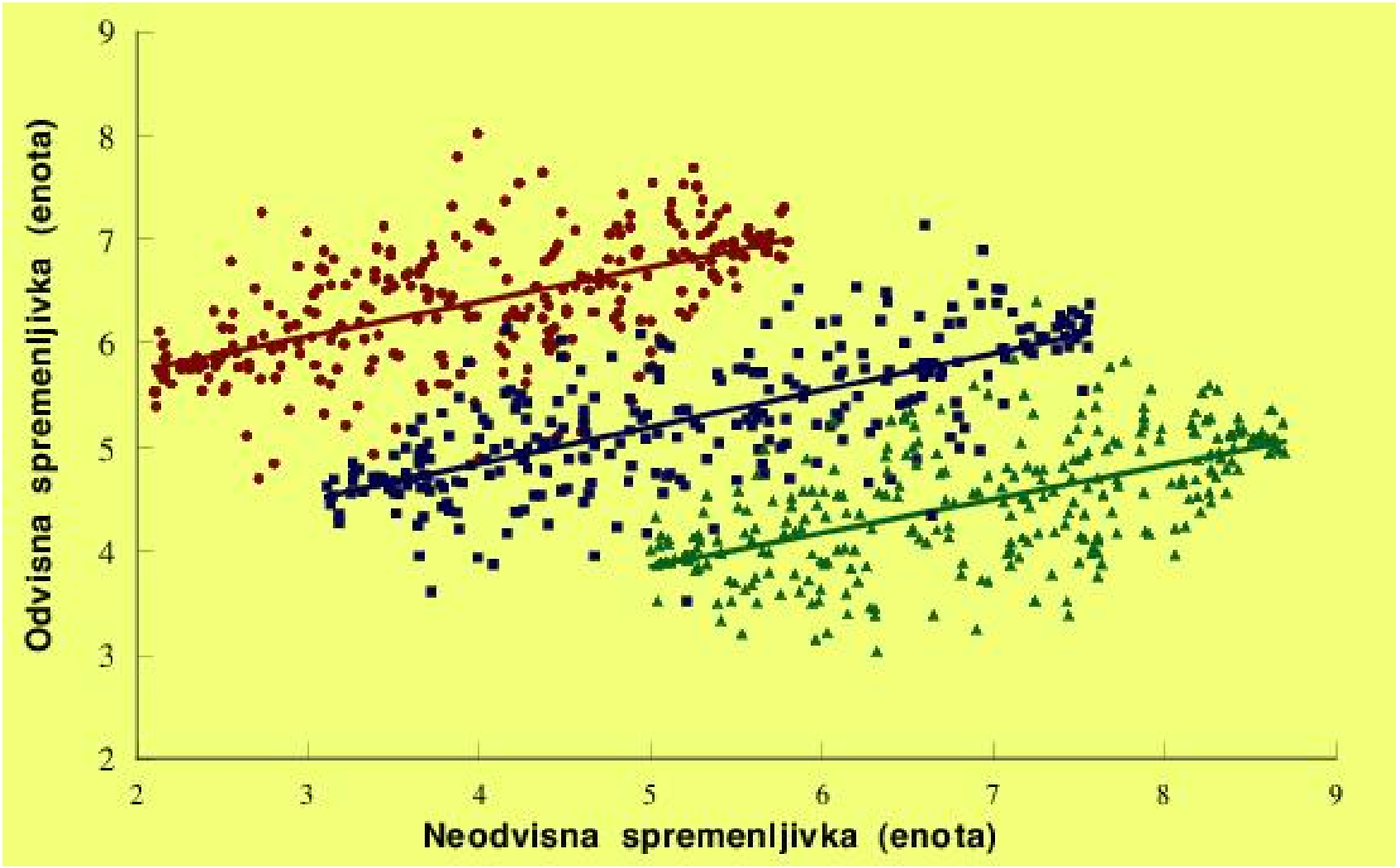
Primer 3

$$\cdots + M_j + b_j(x_{ijklm} - 2) + \cdots$$



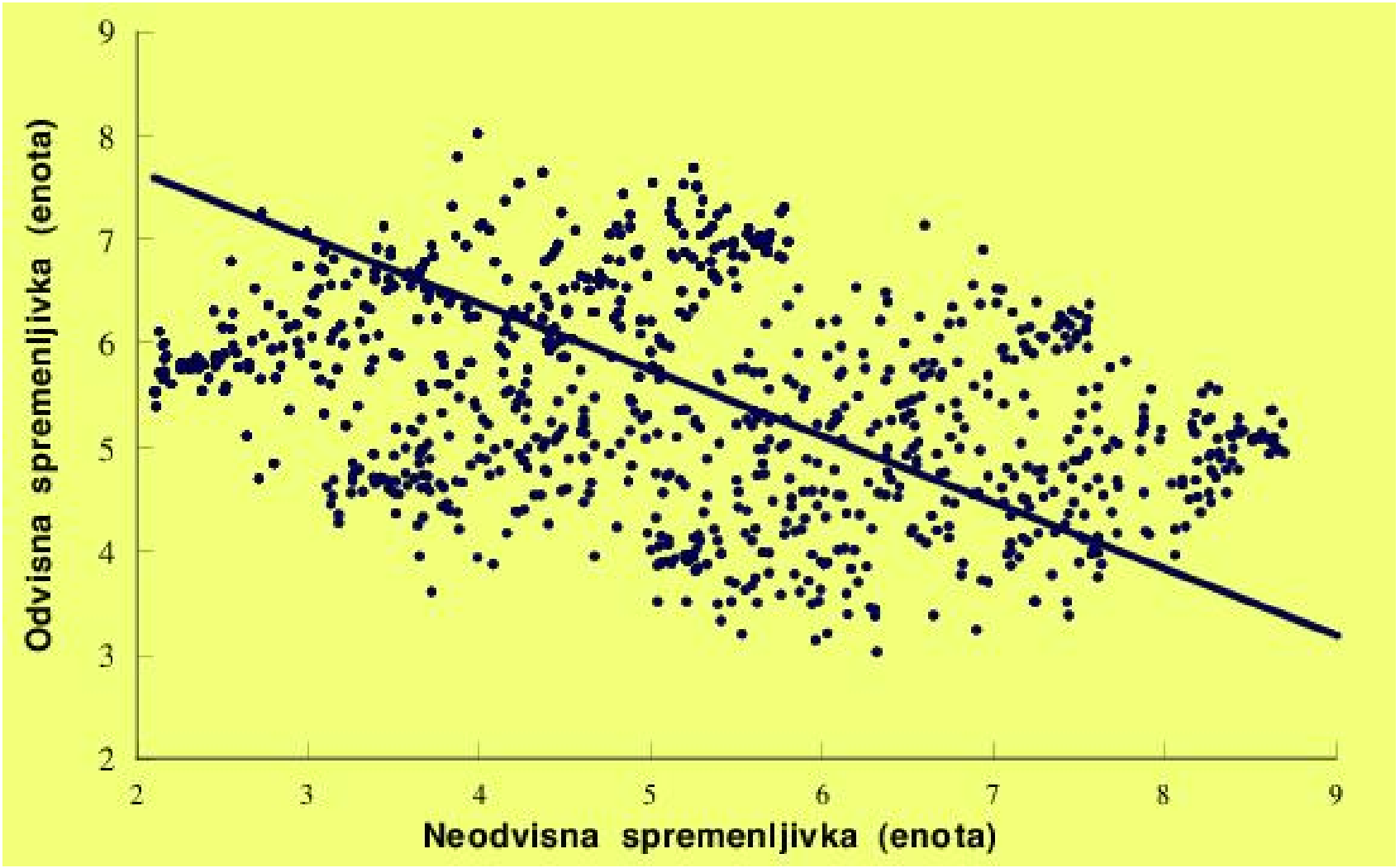
Primer 4

$$\cdots + M_j + b\left(x_{ijklm} - 2\right) + \cdots$$



Primer 5

$$\mu + \cdots + b\,(x_{ijklm} - 2) + \cdots$$



Primer 6: mladice

Žival	Gnezdo	Pasma	Mesec	Farma	Masa(kg)	DP (g/dan)	DHS(mm)	
1	1	SL	JAN	A	102	540	13	13
2	2	SL	JAN	B	98	550	16	14
3	1	SL	FEB	C	105	550	16	16
4	2	SL	FEB	D	102	580	15	
5	4	LW	JAN	A	95	520	20	
6	5	LW	FEB	B	101	500	24	24
7	4	LW	FEB	C	101	490	27	25
8	5	NL	JAN	A	97	560	26	
9	4	NL	JAN	B	100	550	22	
10	6	NL	FEB	C	97	600	23	25
11	7	NL	FEB	D	102	610	24	

Poreklo živali

Žival	Oče	Mati	Žival	Oče	Mati
1	15	—	2	15	10
3	15	—	4	15	10
5	14	10	6	14	12
7	14	10	8	13	12
9	—	10	10	15	12
11	—	—	12	—	—
13	14	—	14	16	—
15	16	—	16	—	—

- vsaki živali s podatki pripišemo starša
- nove starše vpišemo na konec seznama in jim poiščemo starše
- ponavljamo, dokler ne pridemo do neznanih prednikov
- žival ima samo eno vrstico

Število parametrov, stopinje prostosti in rang sistema

- Ponovimo osnovni model za dnevni prirast v preizkusu z mladicami.

- V preizkusu smo imeli skupno 11 meritev, vsaka žival je imela natanko eno meritev.

$$y_{ijklm} = \mu + P_i + M_j + F_k + g_{ijkl} + a_{ijklm} + e_{ijklm}$$

- Lokacijski parametri (ocene za sistematske vplive)

- Parametri disperzije (variance in kovariance za naključne vplive)

Število parametrov

Poskusimo najprej našteti vse parametre!

- Srednja vrednost μ
- Parametri za vpliv pasme: SL (P_1), LW (P_2) in NL (P_3)
- Parametri za vpliv sezone: januar (M_1), februar (M_2)
- Parametri za vpliv farme: A (F_1), B (F_2), C (F_3) in D (F_4)
- Parametri disperzije za naključne vplive:
  - varianca σ_g^2 za skupno okolje v gnezdu
  - genetska varianca σ_a^2 za vpliv živali

Seznam parametrov v modelu

Vpliv	Seznam parametrov	Obrazložitev
Srednja vrednost	μ	
Pasma	P_1, P_2, P_3	tri pasme
Mesec / sezona	M_1, M_2	dva meseca
Farma	F_1, F_2, F_3, F_4	štiri farme
Gnezdo	σ^2_g	varianca
Žival	σ^2_a	varianca

- Pri vplivu skupnega okolja v gnezdu napovemo napovedi za vse nivoje
- Pri vplivu živali napovemo napovedi plemenskih vrednosti za vsako žival
- Pri naključnih vplivih ne štejemo lokacijskih parametrov (= število nivojev)

Število parametrov in stopinj prostosti

Vplivi	Število parametrov	Število stopinj prostosti	
μ	1		1
Pasma	3	3-1 =	2
Mesec	2	2-1 =	1
Farma	4	4-1 =	3
Gnezdo	(7)	polni rang	—
Žival	(11 + 5)	polni rang	—
Za model	10		7
Za ostanek	—-	11 - 7 =	4
Red	10 + 7 + 16		
Rang			7 + 7 + 16

- **Red sistema** = število enačb za sistematske in naključne vplive
- **Rang sistema** = s.p. za model + število nivojev pri naključnih vplivih

Stopinje prostosti v možnem modelu - I

$$y_{ijklm} = \mu + P_i + M_j + F_k +$$
$$+PM_{ij} + PF_{ik} + MF_{jk} + PMF_{ijk} + g_{ijkl} + a_{ijklm} + e_{ijklm}$$

- ocenimo interakcijo, glavne vplive izpeljemo (kasneje izračunamo)

- Izpeljane parametre smo nakazali s kljukico

Parametri	A	B	C	D	Pasma
SL	PF_{11}	PF_{12}	PF_{13}	PF_{14}	√
LW	PF_{21}	PF_{22}	PF_{23}	PF_{24}	√
NL	PF_{31}	PF_{32}	PF_{33}	PF_{34}	√
Farma	√	√	√	√	√

$$P_i = \frac{PF_{i1}+PF_{i2}+PF_{i3}+PF_{i4}}{4}$$

Stopinje prostosti v možnem modelu - II

- Ocenimo lahko tudi srednjo vrednost, glavne vplive in interakcije
- Zadnji nivo pri glavnih vplivih in interakcijah izpeljemo
- Izpeljane parametre smo nakazali s kljukico

Parametri	A	B	C	D	Pasma
SL	PF_{11}	PF_{12}	PF_{13}	$\surd$	P_1
LW	PF_{21}	PF_{22}	PF_{23}	$\surd$	P_2
NL	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Farma	F_1	F_2	F_3	$\surd$	μ

$$P_i = \frac{PF_{i1}+PF_{i2}+PF_{i3}+PF_{i4}}{4}$$

$$4P_i = PF_{i1} + PF_{i2} + PF_{i3} + PF_{i4}$$

$$PF_{i4} = 4P_i - (PF_{i1} + PF_{i2} + PF_{i3})$$

Seznam parametrov in število stopinj prostosti v modelu Pasma, mesec, farma

Seznam parametrov	Število parametrov	Število stopinj prostosti
μ	1	1
P_1, P_2, P_3	3	3-1 = 2
M_1, M_2	2	2-1 = 1
F_1, F_2, F_3, F_4	4	4-1 = 3

Interakcije med P in M, med P in F, med M in F

Seznam parametrov	Štev. param.	Štev. s. p.
PM_{11}, PM_{12}, PM_{21}, PM_{22}, PM_{31}, PM_{32}	$3 \times 2 = 6$	$2 \times 1 = 2$
PF_{11}, PF_{12}, PF_{13}, PF_{14}, PF_{21}, PF_{22}, PF_{23}, PF_{24}, PF_{31}, PF_{32}, PF_{33}, PF_{34}	$3 \times 4 = 12$	$2 \times 3 = 6$
MF_{11}, MF_{12}, MF_{13}, MF_{14}, MF_{21}, MF_{22}, MF_{23}, MF_{24}	$2 \times 4 = 8$	$1 \times 3 = 3$

Interakcija med P, M in F in regresije*

Seznam parametrov	Štev. parametrov		Število s. p.	
PMF_{111}, PMF_{112}, PMF_{113}, ... PMF_{324}	3 x 2 x 4	= 24	2 x 1 x 3	= 6
b_{I111}, b_{I112}, b_{I113}, ... b_{I324}	3 x 2 x 4	= 24	3 x 2 x 4	= 24*
b_{II111}, b_{II112}, b_{II113}, ... b_{II324}	3 x 2 x 4	= 24	3 x 2 x 4	= 24*

* regresijo (polinom druge stopnje, ugnezden znotraj trojne interakcije) imamo v modelu za debelino hrbtne slanine:

$$y_{ijklmn} = \mu + P_i + M_j + F_k + PM_{ij} + PF_{ik} + MF_{jk} +$$
$$+ PMF_{ijk} + b_{Iijk}(x_{ijklm} - 100) + b_{IIijk}(x_{ijklm} - 100)^2 +$$
$$g_{ijkl} + a_{ijklm} + e_{ijklmn}$$

Vpliv skupnega okolja v gnezdu in vpliv živali

Seznam parametrov	Štev. parametrov	Štev. stopinj prostosti
$\boldsymbol{\sigma}_g^2$	(7)	(7)
$\boldsymbol{\sigma}_a^2$	(16)	(16)

- pri naključnih vplivih ne preverjamo števila stopinj prostosti
- ni linearno odvisnih enačb, razen pri enojajčnih dvojčkih ali klonih
- pričakovano vrednost poznamo $(= \mathbf{0})$ in tako ne porabimo stopinj prostosti

Povzetek

	Dnevni prirast	Debelina hrbtne slanine
Število opazovanj	11	17
Število parametrov za sistematske vplive	60	108
Število gnezd	7	7
Število živali	16	16
Red sistema	60+7+16 = 83	60+48+7+16 = 131
Rang sistema	24+7+16 = 47	24+48+7+16 = 95

- Model ima več parametrov kot opazovanj (overparameterization)
  - zato ni dober za to količino podatkov
  - hočemo vedeti veliko več, kot smo zbrali informacij
  - lokacijskih parametrov mora biti veliko manj kot podatkov, da je model primeren
- Za smiselno obdelavo podatkov potrebujemo veliko več podatkov