Del I

STATISTIČNI MODELI

Poglavje 1

Osnovni pojmi

Pred nami so podatki, ki smo jih zbrali bodisi v poizkusu bodisi v proizvodnih razmerah. Podatki so lahko številni, ali pa smo uspeli opraviti le nekaj meritev. Manjše število meritev je opravičljivo, kadar so meritve drage, ali pa so živali izpostavljene neugodnim pogojem. Če se le da, poskušamo opraviti zadostno število meritev. Potrebno število meritev lahko predvidimo pred začetkom preizkusa, v kolikor poznamo porazdelitev opazovanj in zanesljivost metode merjenja. Izogibamo se preizkusom z majhnim številom meritev. Le malo je še lastnosti, o katerih prav ničesar ne vemo in smo veseli, če najdemo povprečja in porazdelitve. Tako vedno poskrbimo že pri načrtovanju poskusa za hipoteze. Z njimi postavimo cilje in tako zagotovo vemo, kaj bomo s čim primerjali. Pri postavitvi hipotez gotovo ne bomo zadovoljni z najbolj preprosto, s katero dokažemo, če je srednja vrednost enaka 0 (ničelna hipoteza) ali pa se razlikuje od 0 (alternativna hipoteza). O načrtovanju poizkusov se bomo še pogovarjali, vendar pa se bomo najprej soočili s statističnim modelom in obdelavo podatkov.

V grobem ločimo dve vrsti modelov: deterministične in stohastične. Modele predstavimo z enačbami. ***Deterministični modeli*** (enačba 1.1) natančno določajo odvisno spremenljivko. Ko izberemo neodvisne spremenljivke (x_i), lahko odvisno spremenljivko (y_i) ***izračunamo***. Parametra β_0 in β_1 sta poznani konstanti. Deterministični model lahko ponazorimo z enačbo za izračun dnevnega prirasta (enačba 1.1) .

$$y_i = \beta_0 + \beta_1 * x_i; i = 1, 2, ..., n \quad [1.1]$$

$$dnevni\ prirast = \frac{prirast}{obdobje} \quad [1.2]$$

Stohastični modeli (enačba 1.3) vedno vsebujejo slučajno spremenljivko - napako (e_i). Zaradi te napake moramo meritve ponavljati in odvisno spremenljivko lahko na koncu samo bolj ali manj zanesljivo ***ocenimo***.

$$y_i = \beta_0 + \beta_1 * x_i + e_i \quad [1.3]$$

β_0 in β_1 sta ***parametra*** porazdelitve. $\hat{\beta}_0$ in $\hat{\beta}_1$ pa njuni ***oceni***. $\hat{y}_i$ je ocena odvisne spremenljivke pri določeni vrednosti neodvisne spremenljivke x (npr. $x = 0$, $x = \hat{x}$), to je tudi njena pričakovana vrednost. Izvrednotimo jo po enačbi 1.4.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i + \hat{e}_i \quad [1.4]$$

Spomnimo se, da smo ostanek poimenovali tudi odklon. Pri merah razpršenosti smo najprej začeli s povprečnim odklonom in ugotovili, da je njegova vrednost vedno 0. Tako lahko zatrdimo, da je pričakovana vrednost za ostanek enaka 0.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i \quad [1.5]$$

POMNI! Parameter ni enak oceni, torej $\beta_0 \neq \hat{\beta}_0$. Prav tako $\beta_1 \neq \hat{\beta}_1$ in $y_i \neq \hat{y}_i$. Parameter predstavlja dejansko, konstantno vrednost, ki je ne moremo izmeriti niti izračunati. Lahko jo le ocenimo ali napovemo. V tem primeru dobimo oceno oziroma napoved, ki bolj ali manj natančno (merilo je standardna deviacija) ter bolj ali manj pristransko (merilo je pristranost - bias) predstavlja omenjeni parameter.

Definirajmo še ***oceno napake***: $\hat{e}_i = y_i - \hat{y}_i$. To je razlika med opazovano vrednostjo y_i in njeno pričakovano vrednostjo $\hat{y}_i$.

1.1 Statistični modeli

Statistični model je abstrakcija realnosti in ne more nadomestiti kompletne slike. Z analizo podatkov želimo številne informacije "zgostiti" na predstavljivo raven. Človek obvlada le okrog 100 parametrov, podatkov pa je lahko na tisoče zapisov. Analiza nam torej nudi izvleček informacij. Izvlečki (parametri) naj bi zadovoljivo pojasnjevali podatke, vendar pa naj bi jih bilo le toliko, kot je nujno potrebno (***zakon skromnosti*** - parsimony). Z odvečnimi parametri izgubljamo učinkovitost statističnega preizkusa in s tem zmanjšamo zanesljivost ter uporabnost analize oziroma zaključkov. Model lahko dostikrat preuredimo tako, da zmanjšamo število parametrov. Npr. razrede pri kvantitativnih sistematskih vplivih lahko pogosto nadomestimo z regresijo. Drugi primer so interakcije, ki odvzamejo sorazmerno veliko stopinj prostosti. Kadar niso pomembne, jih velja izpustiti iz modela.

Opisi statističnih modelov so sestavljeni iz štirih delov - elementov:

1. ***Enačbe*** (equations)
2. ***Pričakovane vrednost***i (excepted values)
3. ***Strukture varianc in kovarianc*** (covariance structure, covariance matrices)
4. ***Predpostavk*** (assumptions) in ***omejitev*** (restrictions)

Najprej bomo porabili čas za opis enačbe statističnega modela, kasneje pa bomo obdelali tudi druge elemente modela. Pri običajnih poskusih, ki so načrtovani in temeljijo na naključnih vzorcih, so ostali trije elementi preprosti, z njimi pa se srečamo, kadar obdelujemo proizvodne podatke ali pa se ukvarjamo s selekcijo ali izločanjem živali.

Statistični model opisuje opazovanje, imenovano tudi meritev, podatek ali lastnost. Opazovanje ali meritev je funkcija sistematskih (fixed) ter naključnih (random) vplivov in ostanka (residual). V statistiki opazovanje ali lastnost poimenujejo tudi odvisna spremenljivka. Tako odvisno spremenljivko pojasnjujejo posamezni vplivi, ki jim zato rečemo tudi pojasnjevalne oziroma neodvisne spremenljivke.

$$\begin{bmatrix} opazovanje, \\ meritev, \\ lastnost\ ali \\ odvisna\ spremenljivka \end{bmatrix} = f \begin{bmatrix} sistematski\ vplivi \\ in\ naključni\ vplivi \end{bmatrix} + ostanek \qquad [1.6]$$

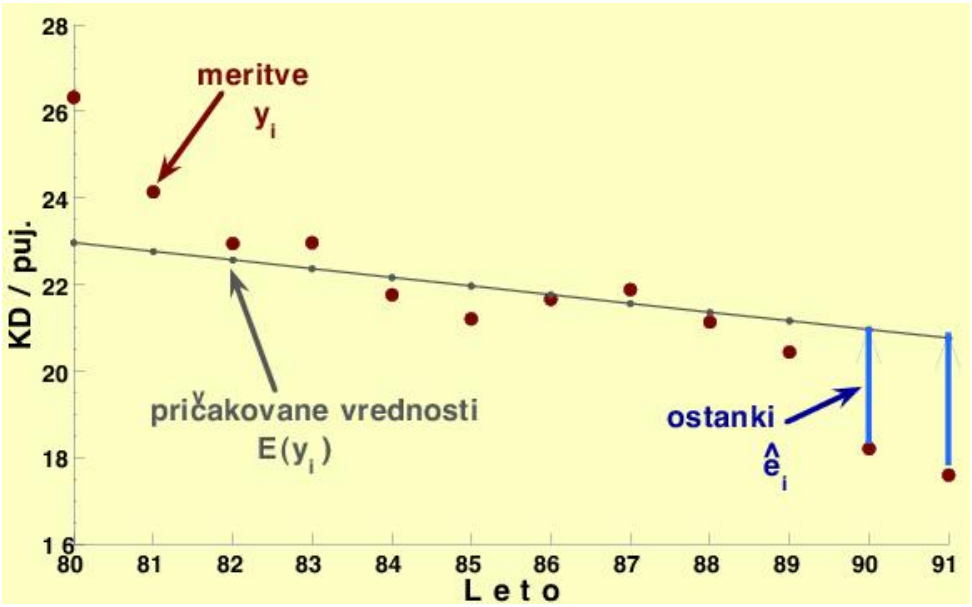
$$\begin{bmatrix} observation, \\ measurement, \\ trait, \\ dependent\ variable \end{bmatrix} = f \begin{bmatrix} fixed\ effects \\ random\ effects \end{bmatrix} + residual \qquad [1.7]$$

Spoznali smo se že z lastnostmi in vplivi, tudi razdelitev na sistematske in naključne vplive nam ni neznana, zato se bomo ukvarjali s funkcijami vplivov in ostankom.

Ostanek (slika 1.1) je tudi naključna spremenljivka, a jo bomo zaradi posebnega značaja in pomena vedno obravnavali ločeno od ostalih naključnih spremenljivk. Ostanek se pojavlja kot naključna napaka pri meritvah ali pa je posledica bolj ali manj zavestnih napak. Zavestne napake zagrešimo, ko iz modela izpustimo manj pomembne sistematske vplive. Izpustimo lahko naključna spremenljivko ali pa izpustimo vpliv, ki ga ne poznamo, ker ga npr. v poskusu nismo beležili. Te napake niso nujno katastrofalne, v danem poskusu samo malo ponagajajo, na zaključke morda niti ne vplivajo. Skupaj pa se jih le nabere za ostanek.

Že dober mizar večkrat premeri sobo in kote, ko se loti opremljanja sobe. Meritve ponavlja in se tako preverja, pa čeprav les, beton in kovina na spreminjajo veliko svoje oblike. V živinoreji pa delamo z biološkimi pojavi, ki so znatno bolj občutljivi na posamezne vplive, ker se biološki pojav (rast, plodnost) razvija dalj časa, nanj vplive več dejavnikov in med seboj sodelujejo ali se ovirajo. Da pojav spoznamo,

Slika 1.1: Ostanek



Slika 1.2: Meritve, primerjalna vrednost in ostanek

ne verjamemo eni sami meritvi, ampak jih ponovimo več hkrati pod istimi pogoji. Meritve so praviloma med seboj bolj različne kot mizarjeve. Da bi lahko presodili vplive dovolj zanesljivo, jih torej večkrat ponovimo. Povprečje meritev opravljenih pod istimi pogoji, predstavlja ***primerjalno vrednost***. Primerjalna vrednost se lahko razlikuje, če smo vzorčili iz več podmnožic - če smo opazovali več skupin. Primerjalna vrednost se lahko spreminja od meritve do meritve, kadar imamo opravka s kvantitativnim vplivom. Izraz primerjalna vrednost uporabimo, ko hočemo poudariti uporabnost vrednosti. Meritev primerjamo s primerjalno vrednostjo in vemo, ali je dobra ali slaba. Kadar pa za poudarimo vlogo v statistiki, pa primerjalni vrednosti rečemo pričakovana vrednost. Odstopanje meritev od primerjalne vrednosti imenujemo ostanek.

PRIMER : Na sliki 1.2 smo prikazali meritve za lastnost število krmnih dni na živorojenega pujska, kar predstavlja lastno ceno pujska, po letih. Izbrali smo primer z malo točkami in model z linearno regresijo, da se bodo jasno pokazala odstopanja. Rdeče točke na sliki predstavljajo naše meritve (y_i), premica predstavlja pričakovane vrednosti ($\hat{y}_i$).Tako za vsako meritev lahko izračunamo svojo pričakovano vrednost. Z modrimi črtami, ki povezujejo rdeče točke - meritve - s točkami na premici, smo označili ocene ostankov (e_i). Ocenjene ostanke imamo pri vsaki meritvi, čeprav smo na sliki prikazali samo dve oceni pri zadnjih meritvah. Kadar sta vrednosti za meritev in pričakovano vrednost enaka, je ocena ostanka enaka 0. Tudi v tem primeru imamo ostanek.

Tabela 1.1: Velikost paradižnikov (cm) po skupinah

Skupina 1			Skupina 2			Skupina 3			Skupina 4		
y_{1j}	$\hat{y}_1$	$\hat{e}_{1j}$	y_{2j}	$\hat{y}_2$	$\hat{e}_{2j}$	y_{3j}	$\hat{y}_3$	$\hat{e}_{3j}$	y_{4j}	$\hat{y}_4$	$\hat{e}_{4j}$
74	72	2	76	79	-3	87	90	-3	103	103	0
67	72	-5	80	79	1	91	90	1	99	103	-4
77	72	5	81	79	2	94	90	4	105	103	2
69	72	-3				88	90	-2	106	103	3
73	72	1							102	103	-1

Tabela 1.2: Velikost paradižnikov (cm) brez skupin

Skupina 1			Skupina 2			Skupina 3			Skupina 4		
y_i	$\hat{y}$	$\hat{e}_i$	y_i	$\hat{y}$	$\hat{e}_i$	y_i	$\hat{y}$	$\hat{e}_i$	y_i	$\hat{y}$	$\hat{e}_i$
74	86.59	-12.59	76	86.59	-10.59	87	86.59	0.41	103	86.59	16.41
67	86.59	-19.59	80	86.59	-6.59	91	86.59	4.41	99	86.59	12.41
77	86.59	-9.59	81	86.59	-5.59	94	86.59	7.41	105	86.59	18.41
69	86.59	-17.59				88	86.59	1.41	106	86.59	19.41
73	86.59	-13.59							102	86.59	15.41

PRIMER : V poskusu so študirali vpliv različnih količin gnojila na rast paradižnikov (pregl. 1.1). Merili so velikost osem tednov po presajanju. Velikost paradižnikov je odvisna spremenljivka. Izmerjene vrednosti pa njene realizacije, dogodki. Učinek gnojila naj bi se pokazal v velikosti paradižnikov. Neodvisna spremenljivka bi lahko bila količina gnojila, vendar pa se bomo tokrat raje odločili za razrede: imenovali jih bomo skupine. Ker pričakujemo razlike med skupinami, bomo za vsako skupino izračunali pričakovano vrednost - povprečje. Izračunajmo še vse ostanke in pričakovano vrednost za ostanek!

Ker smo različno gnojili parcele, ne pričakujemo enakega pridelka, ampak predvidevamo razlike med skipinami. Tako v statistični model (1.8) vključimo poleg srednje vrednosti (μ) tudi vpliv skupine (S_i). Ker smo po skupinah imeli posajenih več rastlin, pri opazovanju dodamo indeksu skupine (i) tudi indeks rastline (j).

$$y_{ij} = \mu + S_i + e_{ij} \qquad [1.8]$$

Pričakovana vrednost v posameznih skupinah je različna ($\hat{y}_i$). Ostanek (e_{ij}) ima iste indekse kot opazovanje, saj pri vsaki meritvi lahko naredimo tudi napako pri merjenju. Ostankov ne poznamo, ocenjujemo jih lahko le kot odstopanja meritve od povprečja (pregl. 1.1). Vsota ostankov za vsako skupino posebej je 0.

V primeru, da bi razlik med skupinami ne pričakovali, potem v modelu (1.9) ni drugega kot srednja vrednost (μ). Opazovanja (y_i) imajo samo en indeks, ki označuje rastline v poskusu.

$$y_i = \mu + e_i \qquad [1.9]$$

Ker za vse paradižnike pričakujemo enako rast, izvrednostimo samo eno pričakovano vrednost ($\hat{y}$), ki jo predstavlja povprečje višine vseh paradižnikov (86.59 cm). Ocenjeni ostanki so v tem primeru veliko večji. Vsota ostankov po skupinah nima vrednosti 0, ampak sta vsoti v prvih dveh skupinah negativni, v zadnjih dveh pa pozitivni. Tako že na oko vidimo, da ima skupina (gnojenje) pomembno vlogo pri rasti paradižnikov. Skupna vsota ostankov je enaka 0, odstopanje je le zaradi zaokroževanja vrednosti na dve decimalki.

Poglavje 2

Enačba modela

Datoteka s podatki o testiranju mladic na rast in mesnatost vsebuje 11 zapisov (pregl. 2.1). Izmerjenih je bilo 11 mladic (živali), treh pasem v mesecih januar in februar. Mase pri merjenju so bile med 96 in 105 kg. Slanino so merili z dvema ponovitvama, dnevni prirast pa je izračunan iz podatkov o starosti in masi pri merjenju.

Pri postavljanju modela si v besedilu, preglednici ali sliki najprej označimo lastnosti in nato še vplive. Oznake lahko izberete po svoje, v študijskem gradivu bomo pri prikazih dela lastnosti navedli z rdečim besedilom, vplive pa z zelenim. Predlagamo, da ločite tudi kvalitativne in kvantitativne vplive, npr. ene podčrtate. V našem poskusu imamo dve odvisni spremenljivki - **debelino slanine** in **dnevni prirast**. **Pasma** in **mesec** sta sistematska vpliva z nivoji oz. kvalitativna vpliva. **Masa** je kvantitativni vpliv, pomemben pri debelini hrbtne slanine, in ga vključimo v model z regresijo. Ker je dnevni prirast izračunan iz mase in trajanja pitanja, mase ne uporabljamo v modelu za dnevni prirast. Zadnji vpliv lahko predstavlja **žival**, kar predstavlja aditivni genetski vpliv, ki se prenaša od staršev na potomce. Aditivni genetski vpliv je naključni, vpliv z nivoji, med katerimi poznamo sorodstvo. Nivoji (razredi) niso neodvisni, ker imajo sorodne živali iste gene, ki izvirajo od prednikov.

V modelu so lahko prisotne vse tri skupine vplivov, vendar to sploh ni nujno. Tako imamo modele samo s sistematskimi vplivi z nivoji, samo z regresijo, z ali brez naključnih vplivov ter vse možne kombinacije. V model vključimo tiste vplive, ki jih želimo proučiti, in tiste, ki jih v poskusu nismo mogli kontrolirati in bi nam lahko pri rezultatih nagajali. Na splošno velja, naj bo model preprost, čeprav tega ne smemo doseči za vsako ceno. Kasneje bomo omenili, katere kriterije upoštevamo pri dokončni izbiri statističnega modela.

2.1 Oznake v statističnem modelu

Model pri obdelavi poskusa mora opisati zbrane podatke y_{ijk} (leva stran enačbe) z nizom sistematskih (*angl.* fixed) in naključnih (*angl.* random) vplivov na desni.

Napišimo enačbo za slanino iz preglednice 2.1 in vzemimo, da na debelino slanine vplivajo samo pasma, mesec, masa in žival. V poskusu so prisotne tri pasme, izvajali pa smo ga dva meseca. Vsak mesec smo preizkusili več živali pri vsaki pasmi, zato smo uporabili dodatni indeks k. Število živali smo označili

Tabela 2.1: Podatki o preizkusu mladic na rast in zamaščenost

Žival	Pasma	Mesec	Masa (kg)	Debelina slanine (mm)	Dnevni prirast (g/dan)
1	SL	JAN	102	13	13	540
2	SL	JAN	98	16	14	550
3	SL	FEB	105	16	16	550
4	SL	FEB	102	15	12	580
5	LW	JAN	95	20	17	520
6	LW	FEB	101	24	24	500
7	LW	FEB	101	27	25	490
8	NL	JAN	97	26	27	560
9	NL	JAN	100	22	19	550
10	NL	FEB	97	23	25	600
11	NL	FEB	102	24	22	610

z n_{ij}, saj je različno med skupinami.. Vsaka žival je imela dve meritvi, zato imamo pri opazovanju in ostanku dodaten indeks (l). Meritvi smo opravili pri isti masi. Pri preizkusu živali stehtamo in debelino slanine izmerimo s ponovitvami. V tem primeru ima neodvisna spremenljivka en indeks manj kot opazovanje. Indeksi so isti kot pri živali, saj ima vsaka žival le eno tehtanje, torej eno maso.

$$y_{ijkl} = \mu + P_i + M_j + b\left(x_{ijk} - \bar{x}\right) + a_{ijk} + e_{ijkl} \qquad [2.1]$$

kjer pomeni:

y_{ijkl}	- opazovanje (odvisna spremenljivka)	observation, trait
x_{ijk}	- neodvisna spremenljivka	independent variable
$\bar{x}$	- konstanta (povprečje neodvisne spremenljivke)	constant (average of x_{ijk})
μ	- srednja vrednost	mean
P_i	- sistematski vpliv pasme; i = 1, 2, 3	fixed effect
M_j	- sistematski vpliv meseca; j = 1, 2	
b	- linearni regresijski koeficient	linear regression coefficient
a_{ijk}	- naključni vpliv živali; k = 1, 2, ... n_{ij}	random effect
e_{ijkl}	- ostanek, z modelom nepojasnjen; l= 1, 2	residual

Prisotne morajo biti informacije:

1. o številu opazovanj
2. o številu razredov (nivojev) pri vsakem vplivu: npr. i = 1, 2, ... n_i
3. opis, kako so bila opazovanja merjenja
4. kateri vplivi so sistematski in kateri naključni.

Nomenklatura:

- opazovanje: mala črka "y". Če imamo več različnih lastnosti, dodamo kot prvi indeks arabsko številko (npr. $y_{1...}$, $y_{2...}$, $y_{t...}$) ali črko (npr. $y_{S...}$, $y_{D...}$, $y_{t...}$), ki nas spominja na lastnost. Izjemoma uporabimo tudi črko z. Pike, ki sledijo indeksu za lastnost, predstavljajo ostale indekse, ki jih določimo v modelu.
- sistematski vplivi: **velike** ali grške črke
- naključni vplivi: **male** črke; pogosto se uporablja "a" za aditivni genetski vpliv (plemenska vrednost), "u" in "z" na splošno
- ostanek: mala črka "e"
- glavni vplivi: **ena** črka z **enim indeksom** (primer: vpliv pasme P_i)
- vgnezdeni vplivi: **ena** črka z **več indeksi** (primer: vpliv živali znotraj pasme a_{ij}), za vsak nivo en indeks. Vgnezdeni vpliv nosi tudi indekse nadrejenih vplivov.
- interakcije: **več** črk z **več** indeksi. Praviloma najprej navedemo nadrejene vplive, potem pa interakcije. Interakcijo poimenujemo tako, da najprej sestavimo črke - oznake vplivom, med katerimi pričakujemo interakcijo, nato pa še indekse. Praviloma se držimo abecednega vrstnega reda indeksov. Če nadrejene vplive izpustimo, lahko obdržimo tudi kombinacijo črk. Tako poudarimo, kako smo do vpliva prišli. Bralcem bo morda tudi bolj jasno, če bomo potem z linearnimi kombinacijami ocen za interakcije poskušali oceniti nadrejene interakcije ali nadrejene glavne vplive.

2.2 Postopek izgradnje modela

Postopek izgradnje modela bomo spremljali na primeru mladic iz preglednice2.1. Na nekaterih mestih bomo za ilustracijo dodali še kakšno predpostavko. Uporabili bomo tudi druge primere, zato bodite pozorni na spremno besedilo.

2.2.1 Glavni vplivi

2.2.1.1 Seznam glavnih vplivov

Sestavimo seznam vseh vplivov, ki bi jih lahko opazovali v modelu. Izberemo vplive, ki jih bomo v poskusu proučevali. To so tako imenovani glavni vplivi. Kadar poskus izvedemo v kontroliranih pogojih, je lista vplivov končana. Pri večini poskusov v živinoreji pa v praksi "nagajajo" še drugi vplivi, na katere nimamo vpliva ali pa bi jih bilo težko kontrolirati. Tako na prirejo vpliva npr. sezona, ker so različni klimatski pogoji, različna prehrana itd. Tako med glavne vplive uvrstimo tudi tiste, ki bi nas v poskusu lahko motili, a jih ne moremo kontrolirati.

Število vplivov pa mora biti v poskusu obvladljivo, zato ne pretiravamo. Pri načrtu poskusa je potrebno poznati možne vplive, seznam katerih pripravimo iz pregleda literature, lahko pa dodamo seveda tudi kakšnega, ki še ni nikogar zanimalo. Razčlenjenost poskusa mora biti taka, da so končni rezultati uporabni.

Označimo si vplive, ki bi jih kazalo v poskusu načrtno spremljati in ostale zavržemo. To izbiro naredimo, ko poskus načrtujemo! Pri izvedbi poskusa pa je dobro zbirati tudi dodatne informacije, ki bi morda lahko vplivali na poskus. Če se že ne lotimo sistematskega zbiranja dodatnih informacij, pa pripišimo zadeve vsaj takrat, ko se med poskusom kaj nepričakovanega zgodi. Tako lahko žival zboli, pri opazovanju obnašanja pride do nepričakovanega ropota, v poletnem času sta bila dva dneva izredno mrzla in vetrovna, kar bi lahko vplivalo na opazovano lastnost itd.

PRIMER: Vzemimo primer za dnevni prirast pri mladicah Vplivi, ki smo jih v poskusu beležili so: pasma, mesec, žival. Sicer na prirast mladic lahko vpliva še veliko drugih vplivov, ki smo jih v poskusu držali konstantne ali pa jih nismo uspeli zabeležiti. Prirast bomo izvrednotili samo od rojstva do odbire, ne pa tudi na drugih intervalih rasti. Ker smo podatke zbrali v preizkusu mladic, bomo spremljali samo ženske živali, ne pa tudi kastratov ali merjascev. Ker nismo pričakovali, da bi rejec lahko vplival, ga nismo zapisali. V tem poskusu tudi nismo predvideli skupnega okolja v gnezdu, ki ima po dosedanjih izkušnjah pomemben vpliv. Kasneje bomo skupno okolje v gnezdu in čredi tudi vključevali, v prvem koraku pa se bomo zadovoljili s preprostejšim. Nekateri vplivi pa so združeni: prehrana se nekoliko spreminja s časom, ker se spreminjajo sestavine. Vpliv prehrane je torej močno povezan s časom - sezono in vpliva nista ločljiva. Tako bomo skupen vpliv še naprej združevali in skupno imenovali sezonski vpliv. Pri prašičih smo za sezono uporabili mesec. Da bi poudarili, da smo sezone naredili po mesecih, smo vpliv poimenovali "mesec".

Posebno mesto namenjamo neodvisnim spremenljivkam. Za opis teh vplivov je največkrat primerna regresija, ki jo ponazorimo s funkcijo, kriterije za izbor bomo obravnavali kasneje v poglavju o regresijskih modelih (2.3.2). Vpliv dobi ime po neodvisni spremenljivki, vedno pa ga označimo bodisi s črko b ali β**.**

2.2.1.2 Označimo glavne vplive

enolično, kot smo se dogovorili: sistematske z veliko črko, naključne z malo črko. Kot glavni vpliv razumemo vpliv, katerega razredi so identificirani z enim samim indeksom. Indeksi so nam pri gradnji modela dobrodošli pripomoček, zato jih velja pripisati. V tem koraku niso nujni, vendar se bomo odločili in jih navajali, da se jih navadimo in razumemo njihov pomen.

PRIMER: Nadaljujmo s primerom dnevnega prirasta pri mladicah
P_i - vpliv pasme; i = 1, 2, 3 (1=ŠL, 2=LW, 3=NL)
M_j - vpliv meseca oziroma sezone; j = 1, 2 (1=JAN, 2=FEB)
a_{ijk} - vpliv živali , aditivni genetski vpliv oziroma plemenska vrednost živali; k = 1, 2, ... n_{ij}

Dogovorili smo se, da glavne vplive označimo z eno črko, praviloma začetnico naziva. Pri sistematskih vplivih uporabimo veliko črko, pri naključnih pa malo črko. V omenjenem poskusu smo izmerili samo 11 živali, zato bi lahko bili celo v dilemi, kako obravnavati žival. Živali je malo, odlučojoče pa je dejstvo, da imamo po živali eno samo meritev. V resnici je poskus premajhen, da bi dal kakorkoli uporabne rezultate. Uporabljamo ga samo za ponazoritev postopka, pri tem bomo razmišljali, kot bi bilo v poskusu več 10000 živali.

Vpliv pasme in meseca sta križno klasificirana, kar označimo tako, da navedemo različna indeksa. Živali vseh treh pasem smo merili v obeh mesecih. Žival je lahko samo ene pasme. Ker je merjena samo enkrat, je bila meritev lahko opravljena samo v januarju ali februarju, nikakor pa ne v dveh mesecih. To ponazorimo tako, da navedemo indeks pasme in indeks meseca. Vsak mesec smo pri vsaki pasmi zagotovo izmerili več živali, saj smo že v našem vzorcu to nakazali. Tako potrebujemo dodatni indeks k, ki bo preštel živali (pravzaprav meritve) i-te pasme izmerjene v j-tem mesecu. Število živali bo v primeru iz prakse različno.

Ne pozabite! Vplive moramo opisati. K opisu pa sodi tudi opis indeksa. Če imamo manjše število nivojev, te nivoje naštejemo, pri daljšem nizu napišemo dva, nadaljujemo s tremi pikicami in končamo z največjo vrednostjo . Ta vrednost je lahko eksplicitno navedena, če je vrednost znana. V primernu, da se končni vrednosti želimo izogniti, navedemo oznako. Običajno, kadar ni dvoumno je to n ali m, ker nas to spomni na število. Zlasti kadar n uporabimo v indeksu, pa iščemo druge rešitve: p in q sta tudi sorazmeroma pogosto uporabljena, ker jih običajno uporabljamo za število parametrov. Kadar rabimo, lahko v indeksu navedemo tudi vpliv kot npr. n_P za število pasem, n_a za število živali. V našem primeru smo uporabili n_{ij}, kar pomeni, da pričakujemo različno število živali i-te pasme v j-tem mesecu. Pazite, da so opisi nazorni in nedvoumni!

2.2.1.3 Določimo naravo vplivov

O naravi vplivov odloča struktura podatkov in jo najbolje ponazorimo s postavitvijo indeksov. Če to storimo že ob poimenovanju, pri tem koraku še enkrat preverimo, da je delo korektno opravljeno. Da pa na to ne bi pozabili, smo dali še poseben korak.

Tako ločimo križno klasificirane (pregl. 2.2) in vgnezdene vplive (pregl. 2.3). Križno klasificirani modeli so zaželjeni, ker opisujejo splošna pravila in omogočajo prepoznavanje tudi posebnosti (specifičnih vplivov). Da sta dva (ali več) vpliva križno klasificirana (pregl. 2.2), moramo imeti meritve pri vseh možnih kombinacijah teh dveh (ali več) vplivov. Po drugi strani so vgnezdeni vplivi (pregl. 2.3) tisti, pri katerih skupina nivojev enega vpliva pripada drugemu vplivu in sicer samo enemu nivoju tega nadrejenega vpliva. Nivoji vgnezdenega vpliva so hierarhično razporejeni znotraj nadrejenega vpliva. Tako najprej navedemo nadrejeni in nato vgnezdeni vpliv. V našiem primeru iz preglednice smo vsako pasmo preizkusili v dveh zaporednih, a različnih mesecih. Pri prvi pasmi smo preizkus opravili v prvih dveh mesecih (zimskih), pri drugi pasmi začetek pomladi, pri tretji pa konec pomladi in začetek poletja. Razlike med meseci so za večino proizvodnih lastnosti med prašiči tolikšne, da primerjava razlik med pasmami in meseci ni mogoča.

Prikazujemo pa še ponesrečen načrt poskusa (pregl. 2.4). Če bi hoteli uvrstiti ta načrt glede na naravo vplivov, bi bili vplivi križno klasificirani, vendar pa je obdelava takih podatkov zelo problematična zaradi praznih celic. Prazne celice so tisti meseci pri posamezni pasmi (oznaka PM_{ij}), ko nismo opravili nobene meritve. Načeloma bi zadoščala samo ena meritev, čeprav sedaj že dobro vemo, da bi bilo povprečje slabo ocenjeno. V tem primeru bi verjetno najbolje storili, če bi tri celice (PM_{13}, PM_{26} in PM_{31}) izločili ali pa jih obdelali v ločenih obdelavah s podatki iz tistih celic, kjer so primerjave mogoče. Pri poskusu

Tabela 2.2: Križno klasificirana vpliva pasme in meseca

	M_1	M_2	$\sum P$
P_1	n_{11}	n_{12}	$n_{1.}$
P_2	n_{21}	n_{22}	$n_{2.}$
P_3	n_{31}	n_{32}	$n_{3.}$
$\sum M$	$n_{.1}$	$n_{.2}$	$n_{..}$

Tabela 2.3: Vpliv meseca vgnezden znotraj pasme

	M_1	M_2	M_3	M_4	M_5	M_6	$\sum P$
P_1	n_{11}	n_{12}	0	0	0	0	$n_{1.}$
P_2	0	0	n_{23}	n_{24}	0	0	$n_{2.}$
P_3	0	0	0	0	n_{35}	n_{36}	$n_{3.}$
$\sum M$	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	$n_{.6}$	$n_{..}$

smo opravili več dela in morda več analiz, kot bomo uspeli izluščiti iz poskusa.Čeprav si takih podatkov ne želimo, pa se nam včasih le zgodi, da moramo pristati na take podatke. Kadar rejec ne redi vseh pasem ali genotipov, ki bi jih radi opazovali, nimamo dosti izbire. Tudi pri nabiranju vzorcev v trgovinah ne moremo kupiti iste izdelke vseh prodajalcev v vseh trgovinah. V večjih trgovinah prodajajo izdelke številnih proizvajalcev, manjše pa se z izdelkom oskrbujejo pri enem proizvajalcu. Izberejo lahko celo manjšega proizvajalca, ki zagotavlja posebno kakovost in ne sodeluje z večjimi trgovci.Tako je nemogoče priti do optimalnega načrta poskusa.

PRIMER: **Nadaljujmo primer za dnevni prirast pri mladicah** Križno klasificirane in vgnezdene vplive smo že določili. Tako sta pasma in mesec križno klasificirana vpliva, žival pa je vgnezdena znotraj pasme in tudi meseca, ker so bile mladice samo enkrat merjene.

Tabela 2.4: Ponesrečen načrt poskusa

	M_1	M_2	M_3	M_4	M_5	M_6	$\sum P$
P_1	n_{11}	n_{12}	n_{13}	0	0	0	$n_{1.}$
P_2	0	0	n_{23}	n_{24}	0	n_{26}	$n_{2.}$
P_3	n_{31}	0	0	0	n_{35}	n_{36}	$n_{3.}$
$\sum M$	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	$n_{.6}$	$n_{..}$

Tabela 2.5: Določanje možnih interakcij za dnevni prirast pri mladicah (korak 1)

Vpliv			Interakcija	Novi vpliv
P_i				
M_j	:	P_i	možna	PM_{ij}
a_{ijk}	:	P_i	ni možna	

***PRIMER*: Vpliv leta in pasme pri molznicah** Ponazorimo še en križno klasificirani model (enačba 2.2). Tako npr. opažamo, da se količina mleka na kravo v standardni laktaciji iz leta v leto polagoma povečuje pri vseh pasmah. V modelu lahko imamo dva glavna vpliva: leto in pasmo, ki sta križno klasificirana. V vseh letih smo opazovali vse pasme krav in lahko primerjamo med seboj leta ali pasme. Tako dobimo splošne razlike med pasmami in med leti. Toda v sušnem letu, ko je bila pridelana krma slabše kakovosti in še ta v manjših količinah, pa je bila količina mleka manjša kot običajno. Izboljševanje iz leta v leto pa je splošni trend in bi ga lahko opisali s krivuljo. Kadar pa lahko pride do slabe letine, bomo raje uporabili vplive z nivoji, čeprav je tudi to splošni trend, saj je poslabšanje opazno pri vseh pasmah. Če pa je bil padec večji pri pasmi z večjo povprečno količino mleka, pa je to specifični vpliv pasme. V mislih smo imeli krave črno-bele pasme. Vedeti tudi moramo, da so se najbrž rejci črno-bele pasme drugače (specifično) obnašali: svojim kravam so, čeprav je bila krma draga, kupili vse, kar so potrebovale. Tako v resnici sploh niso občutile slabe letine na svoji koži in so obržale običajno količino mleka. Tudi to je specifični odziv pri črno-beli pasmi krav. V obeh primerih so reagirale črno-bele krave drugače, kot smo "na splošno pričakovali". V tem primeru imamo prisotno interakcijo (PL_{ij}). Interakcija ni razvidna iz samega poskusa. Kasneje bomo omenjali postopek, kako jih določimo in preverimo, če so pomembne.

$$y_{ijk} = \mu + P_i + L_j + PL_{ij} + e_{ijk} \qquad [2.2]$$

2.2.1.4 Določimo vgnezdene vplive

Vgnezdeni vplivi so vplivi, katerih skupina nivojev tega vpliva pripada enemu nivoju nadrejenega vpliva. Vgnezdeni vplivi niso v celoti identificirani, dokler niso opisani glavni vplivi (ali kombinacije vplivov), znotraj katerega so vgnezdeni. Tako vgnezdene vplive običajno navajamo v modelu za nadrejenimi vplivi. V seznamu jim dodamo oznako za "nadrejene" vplive (npr. DA: D je vgnezden znotraj A) ali pa pripišemo indekse (D_{ij}). Več o vgnezdenih vplivih in primere si lahko preberete v poglavju o hierarhičnih modelih (2.3.3.1).

2.2.1.5 Določitev možnih interakcije

Glavne vplive razvrstimo v stolpec (pregl. 2.5), pri regresiji navedemo regresijske koeficiente. Prvi vpliv izpustimo in za ostalimi vplivi napišemo dvopičje ter pripišemo prvi vpliv na desni strani dvopičja od vključno druge vrstice dalje. Interakcija je možna, če se črki v nazivu vpliva ali v indeksu na desni in levi strani ne ponovita. Poimenujemo jo tako, da sestavimo najprej črke sodelujočih vplivov po abecednem zaporedju indeksov. Indekse navedemo skupaj, za nizom črk. Kjer se na levi in desni strani dvopičja pojavita bodisi ista črka za vpliv bodisi isti indeks, interakcija ni možna. Le pri regresijskih koeficientih je drugače. Vse regresijske koeficiente označujemo z isto črko *b*, zato upoštevamo predvsem indekse. Zadnja dva stolpca nista zaželjena, ker pri celotnem postopku zmanjšata preglednost. Prvega lahko nadomestimo s tem, da tam, kjer je možna interakcija, pri desnem stolpcu vpliv obkrožimo ali obkljukamo, drugače pa pustimo nespremenjeno ali prečrtamo. Zadnji pa tudi ni potreben, ker bomo nastalo interakcijo dopisali spodaj.

Možne interakcije dopišemo na konec liste (pregl. 2.6), se pomaknemo v tretjo vrstico, na koncu pripišemo podpičje in drugi vpliv in nato storimo isto pri vseh naslednjh vplivih, tudi pri tistih, ki so bili

Tabela 2.6: Določanje možnih interakcij za dnevni prirast pri mladicah (korak 2)

Vpliv	Interakcija	Novi vpliv
P_i		
$M_j : P_i$		
$a_{ijk} : P_i : M_j$	ni možna	
$PM_{ij} : M_j$	ni možna	

Tabela 2.7: Določanje možnih interakcij za dnevni prirast pri mladicah (korak 3)

Vpliv		nov vpliv
P_i		
$M_j : P_i$		
$a_{ijk} : P_i : M_j$		
$PM_{ij} : M_j : a_{ijk}$	ni možna	

pravkar dodani. Lepo podpisujemo, da ne bi česa pozabili ali preskočili. Ko bomo določali možne interakcije, gledamo prvi stolpec na levi in zadnji stolpec na desni. Kriteriji in postopek so isti kot pri prvem koraku.

V drugem koraku ni bilo nobene nove možne interakcije, zato se lista možnih vplivov v preglednici 2.5 ni nič podaljšala. Postopek moramo nadaljevati z vplivom a_{ijk}. Ostala nam je samo ena kombinacija, ki tudi ni možna, ker na obeh straneh najdemo kar dva ista indeksa i in j.

V našem primeru je postopek končan. Praviloma bi nadaljevali z interakcijo PM_{ij}, vendar se lista pri tem vplivu konča. Torej nismo končali, ker smo se znašli na koncu osnovne liste glavnih vplivov. Če bi sledile dodatne interakcije, bi postopek nadaljevali.

2.2.1.6 Označitev sistematskih ali naključnih vplivov

Glavni vplivi so praviloma že določeni. Pri razvoju modela smo morda spoznali, da smo naredili napako. M To pač popravimo in postopek ponovimo. Interakcije prevzamejo naravo (značaj ali tip) vplivov. Tako so interakcije med samimi sistematskimi vplivi so sistematske. Če pa v interakciji nastopa vsaj eden naključni vpliv, pa bo interakcija naključna. Označimo ga lahko s kombinacijo velikih in malih črk (aT_{ijk}) ali samo z malimi črkami (at_{ijk}), da poudarimo naravo parametra. Tako kot za ostale naključne vplive, tudi za take interakcije potrebujemo znane variance ali pa jih računamo.

PRIMER: **Nadaljujemo primer za dnevni prirast pri mladicah** Interakcija PM_{ij} je sistematska, ker sta oba sodelujoča vpliva sistematska.

7. Izločitev podvojenih vplivov Podvojeni vplivi se nam lahko pojavijo iz različnih vzrokov. Eden od vzrokov bi lahko bil pri vgnezdenih vplivih, kjer ima vsak ali samo kateri od nivojev pri nadrejenem vplivu le po en nivo vgnezdenega vpliva. Za primer omenimo situacijo, ko ima vsak rejec samo eno pasmo. Ko naredimo poskus samo na eni farmi ali na eni pasmi, tudi ne moremo vključiti vpliv farme oziroma pasme. Na ta način je podvojena srednja vrednost μ, ker vpliv z enim nivojem predstavlja populacijo.

Do odvečnih interakcij bi lahko prišlo, če se ne bi držali abecednega zaporedja indeksov. Drugačen vrstni red sicer ni napaka, je pa težje preveriti, če imamo podvojene vplive. Podvojen je vpliv, če se nahajajo iste črke v nazivu ali indeksih. Trojno interakcijo med vplivi A_j, B_i in C_k lahko poimenujemo ABC_{jik} ali BAC_{ijk}. Vplive bi lahko razvrstili tudi drugače, skupno je možnih šest poimenovanj. Vse te oznake

pa predstavljajo isti vpliv: interakcijo pravilno poimenovano BAC_{ijk}. Če se držimo abecednega vrstnega reda indeksov, se bomo že pri določanju možnih interakcij izognili podvojitvam.

Izločimo tudi vplive brez zadostnih dodatnih informacij! Parametri pri sistematskih vplivih ne smejo imeti popolnoma iste indekse kot ostanek. To bi pomenilo, da smo za posamezni nivo pri takem sistematskem vplivu opravili samo eno meritev, kar pa je znatno premalo, da bi ga zadovoljivo ocenili. Podvojitev indeksov je možna pri naključnih vplivih. Za napoved je zadosti ena sama meritev za posamezni nivo (enoto, žival), dodatne informacije lahko dobimo še od koreliranih nivojev (sorodnih živali) ali koreliranih lastnosti, vir informacij pa so tudi znane variance in kovariance. Omenili smo pa že, da lahko naključni vpliv napovemo tudi, ko nimamo nobenih informacij. Se pač odločimo, da ne odstopa od pričakovane vrednosti (npr. plemenska vrednost je 0). Nekoliko bolj pozorni moramo biti na strukturo podatkov, ko ocenjujemo komponente (ko)varianc. Imeti moramo ustrezno količino informacij za posamezno komponento, vendar pa to presega okvir našega predmeta.

POMNI! Viri informacij so lahko različni. Najbolje je, da meritev izmerimo na živali sami, vendar pa to ni vedno mogoče. Če opravljamo preizkus na mlečnost pri govedu ali velikost gnezda pri prašičih, podatkov pri moških živalih ni mogoče izmeriti. Da pa bi napovedali plemensko vrednost pri moških živalih, uporabimo meritve na sorodnicah. To pa je že drugi vir informacij, ki ga povezujemo s poreklom. Tretji vir pa so korelirane lastnosti. Korelacija je lahko pozitivna ali negativna. Pri izbiri koreliranih lastnosti pa pazimo na to, da je korelacija med lastnostima visoka, pri merjeni lastnosti pa mora biti visoka tudi heritabiliteta (dednostni delež). Kot primer naj predstavimo meritve debeline podkožnega maščobnega tkiva pri prašičih. Če smo hudo natančni, nas maščobno podkožno tkivo sploh ne zanima, izboljšali bi radi samo mesnatost. Ker pa smo prepričani in vsi poskusi potrjujejo, da sta mesnatost in debelina podkožnega maščobnega tkiva tesno povezani in je merjenje podkožnega maščobnega tkiva priročnejše, pač merimo podkožno maščobno tkivo. Vir informacij so tudi znane variance in kovariance ali razmerja med njimi. Če vemo, da predstavljata naključni del modela le žival in ostanek in sta njuni varianci v razmerju $0.4 : 0.6$, potem lahko "ostanek", ki ostane po odstanitvi sistematskih vplivov, radelimo na dve komponenti. Brez razmerja pa delitev ne bi bila možna.

PRIMER: **Nadaljujmo primer za dnevni prirast pri mladicah** V tem primeru ni podvojenih vplivov, pri živali in ostanku pa imamo iste indekse, ker ima žival samo eno meritev. Da bi plemensko vrednost živali lahko ločili, potrebujemo še najmanj en vir informacij. Zadostuje že razmerje med aditivno genetsko varianco (σ^2_a) in varianco za ostanek (σ^2_e), zadovoljni pa smo lahko tudi s poreklom.

Pri regresiji moramo paziti, da modelov ne okrnimo. Linearno premico določata dva parametra: regresijski koeficient b in presečišče z ordinato, osjo y. Pri enostavni regresiji ni nevarnosti, medtem ko bi se pri vgnezdeni regresiji kaj hitro lahko pri statistični presoji črtali nadrejeni vpliv ali interakcijo. Vzemimo nekaj pravilnih kombinacij. V prvi enačbi (2.3) imamo regresijo b_j vgnezdno znotraj meseca M_j. Prisotnost vpliva M_j v modelu zagotavlja, da dobimo presečišče z osjo y za vsako premico posebej (slika 2.1).

$$\cdots + M_j + b_j \left(x_{ijklm} - 100\right) + \cdots \qquad [2.3]$$

Tudi v drugi enačbi (2.4) je regresija vgnezdena, tokrat znotraj interakcije PM_{ij}. Prav tako je prisotnost interakcije potrebna, dokler nismo prepričani, če lahko regresijo poenostavimo.

$$\cdots + PM_{ij} + b_{ij} \left(x_{ijklm} - 100\right) + \cdots \qquad [2.4]$$

V slednji enačbi (2.5) imamo kvadratno regresijo z dvema členoma, prav tako kot premica, tudi parabola potrebuje presečišče z osjo y. V našem primeru to zagotavlja prisotnost interakcije. Vsi trije členi imajo

pri vseh treh členih iste indekse, vendar niso tretirani kot podvojeni vplivi. Prvi člen zagotavlja presečišča z osjo y, drugi predstavlja naklon in tretji je povezan z ukrivljenostjo parabole.

$$\cdots + PM_{ij} + b_{Iij}\left(x_{ijklm} - 100\right) + b_{IIij}\left(x_{ijklm} - 100\right)^2 + \cdots \quad [2.5]$$

Sedaj pa še nekaj nezaželenih kombinacij. V naslednjem modelu (2.6) smo črtali mesec kot glavni vpliv. Vse premice so vodene skozi eno točko na osi y. Premicam smo tako vsilili potek skozi eno točko (slika 2.2).

$$\cdots + \not{M} + b_j\left(x_{ijklm} - 100\right) + \cdots \quad [2.6]$$

V naslednji enačbi (2.7) imamo regresijski koeficient za parabolo vgnezden samo znotraj meseca, linearni člen pa znotraj interakcije. Kot vemo, mora imeti vsaka parabola tri člene. Če mora biti vgnezdena znotraj interakcije, moramo to storiti tudi za kvadratni člen. Ko pa je zadostno vgnezdenje znotraj samo enega vpliva, pa to naredimo tudi pri linearnem členu. Tudi v tem primeru pa lahko v modelu ostane interakcija. Tudi tretji primer (2.8) je napačen, komentar pa je zelo podoben, kot v drugem napačnem primeru. Poskusite ga sami opisati.

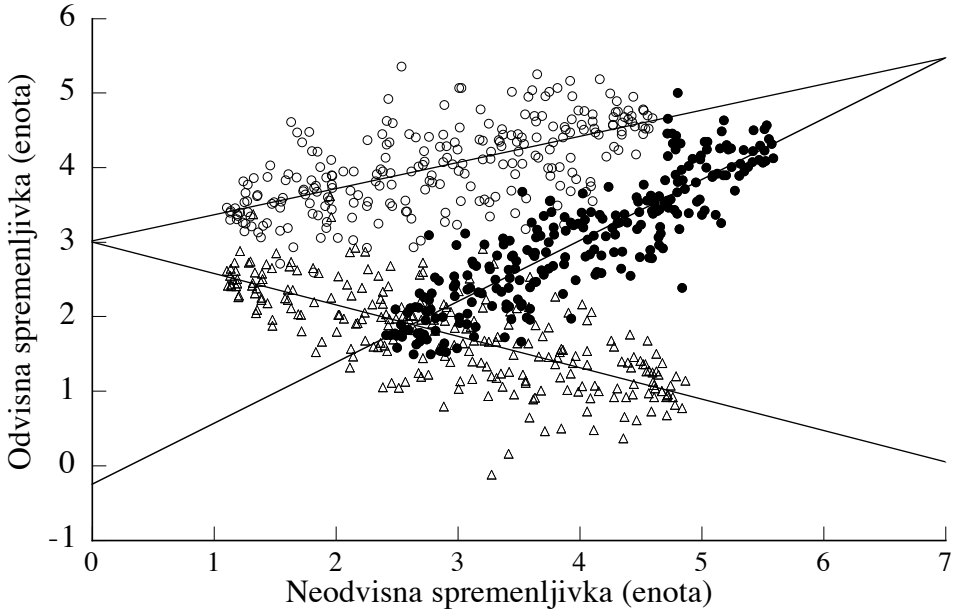
$$\cdots + PM_{ij} + b_{Iij}\left(x_{ijklm} - 100\right) + b_{IIj}\left(x_{ijklm} - 100\right)^2 + \cdots \quad [2.7]$$

$$\cdots + M_j + b_{Iij}\left(x_{ijklm} - 100\right) + b_{IIj}\left(x_{ijklm} - 100\right)^2 + \cdots \quad [2.8]$$

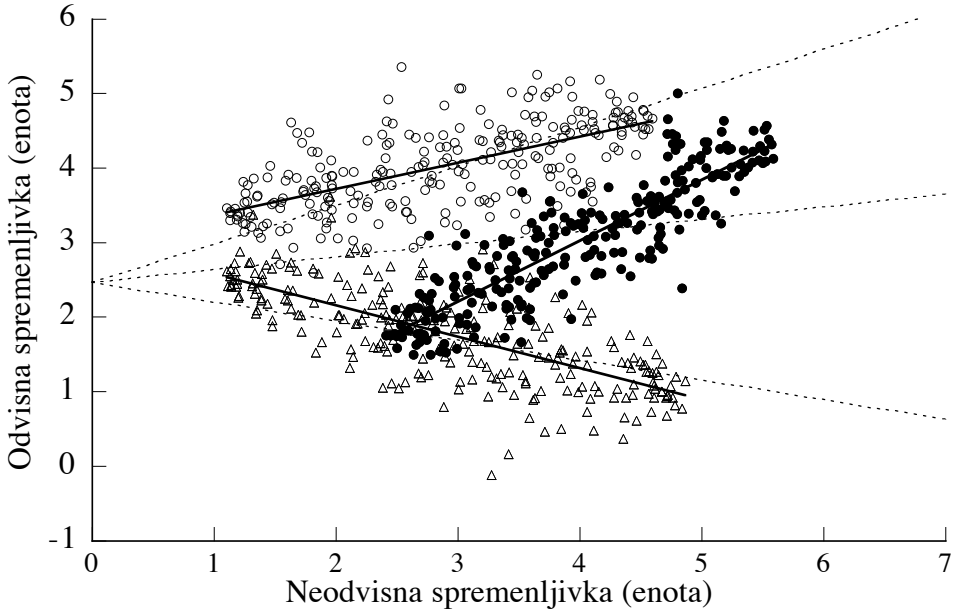
PRIMER: Oglejmo si primer še na sliki (2.1), kjer imamo tri nivoje nekega vpliva, npr. meseca M. Kot običajno je neodvisna spremenljivka oziroma vliv na osi x in odvisna spremenljivka na osi y. Podatke pri prvem nivoju smo prikazali z belimi krogci, drugega s črnimi in tretjega s trikotniki. Pri prvih dveh je regresijski koeficient pozitiven in pri zadnjem pa negativen. Očitno je, da moramo oceniti tri regresijske koeficiente. Če poiščemo presečišča regresijskih premic (polna črta) z osjo y, boste ugotovili, da so različna. Regresija je torej vgnezdena znotraj nekega glavnega vpliva, npr. meseca. Model bo vseboval člena iz enačbe 2.4.
Sedaj pa še poglejmo, kaj se zgodi, če iz modela črtamo "nadrejeni" vpliv mesec (2.2). Ponovno dobimo tri regresijske premice, ki izhajajo iz ene točke na osi y kot šop. Na naši sliki so prikazane črtkano. Kot vidimo iz slike, šop regresijskih premic ne opisuje dobro nobenega oblaka podatkov. Pri oblaku s črnimi krogci še posebej močno odstopa od prave vrednosti.

PRIMER: Vzemimo podatke iz neke druge simulacije. Podatke smo nanesli na sliko 2.3. Ponovno imamo neodvisno spremenljivko x_i in odvisno spremenljivko y_i. Poskusite drug drugemu opisati, kako bi uredili model, da bi opisal kar najbolje situacijo na sliki!
V resnici so zopet predvideni trije snopi točk. Če ste dovolj pozorni, lahko te snopiče tudi prepoznate. Da pa bi jih bolj nazorno videli, smo jih ponazorili še s različnimi znaki na sliki 2.4. V tem primeru skupno presečišče z osjo y zadovoljivo odgovarja in lahko izjemoma po statistični obdelavi nadrejeni vpliv S izpustimo brez škode. Model 2.9bi v tem primeru lahko bil poenostavljen, vendar tudi v takem primeru, če je le dovolj podatkov, obdržimo nadrejeni vpliv v modelu.

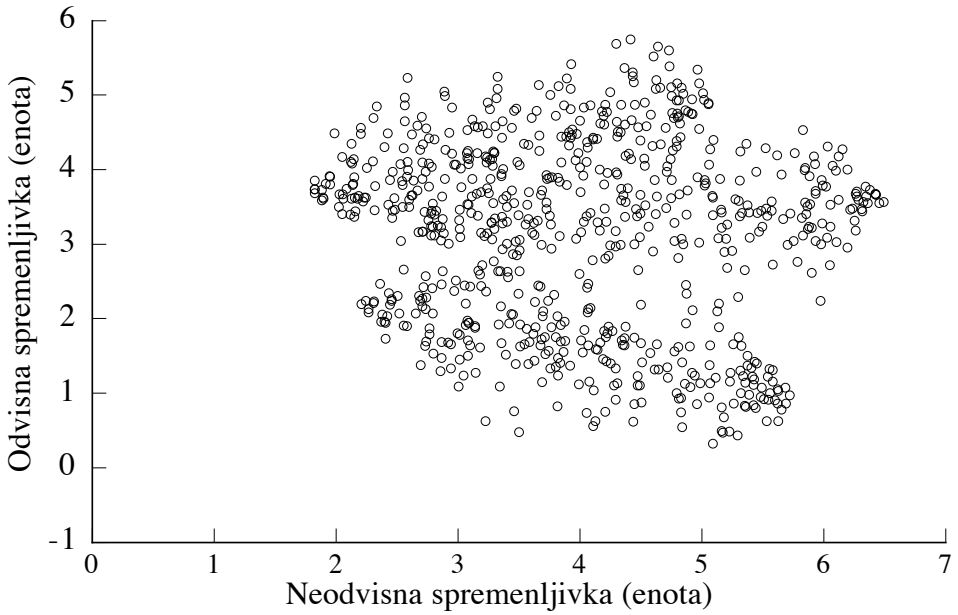
$$y_{ij} = \mu + S_i + b_i x_{ij} + e_{ij} \quad [2.9]$$



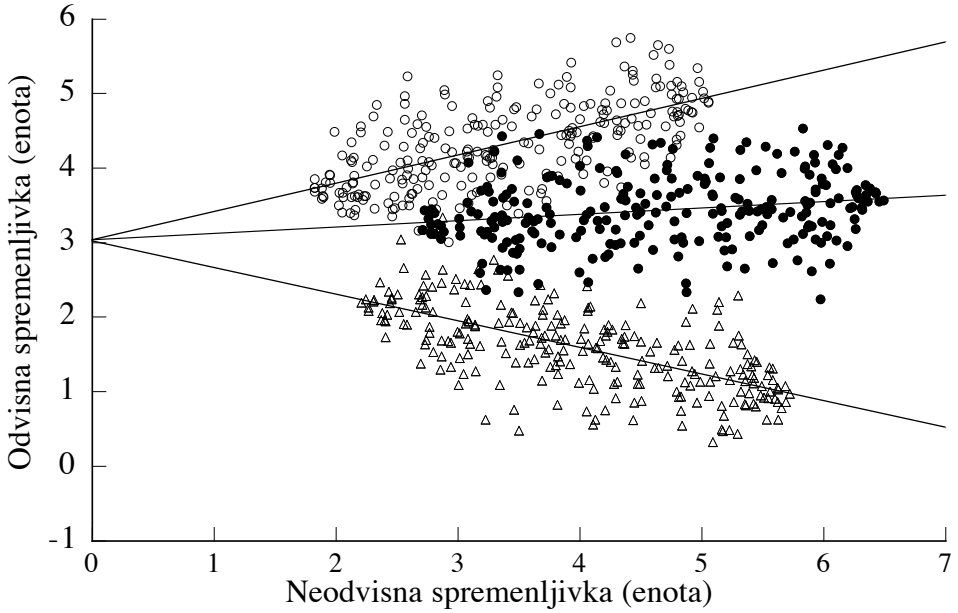
Slika 2.1: Opazovanja pripadajo trem nivojem glavnega vpliva *M* v primeru 2.2.1.6



Slika 2.2: Premicam vsiljen potek skozi skupno presečišče na osi *y* v primeru 2.2.1.6



Slika 2.3: Vpliv neodvisne na odvisno spremenljivko v primeru 2.2.1.6



Slika 2.4: Snopič regresijskih premic za primer 2.2.1.6

8. Napišemo osnovni in možni model. Model je potrebno jasno in v celoti opisati. Tu bomo opisali le prvi del modela - enačbo, vendar pa ne smemo pozabiti na ostale tri komponente, ki pa smo jih do sedaj le omenili. Katere so že?

PRIMER : **Nadaljujemo zgornji primer za dnevni prirast pri odbiri mladic.** Osnovni model 2.10 vsebuje samo glavne vplive, ki smo jih napisali v levi stolpec v preglednici 2.5. V model za dnevni prirast (y_{ijk}) smo vključili srednjo vrednost populacije (μ), sistematska vpliva pasma (P_i) in mesec (M_j) ter naključni vpliv živali (a_{ijk}). Zadnji člen e_{ijk} predstavlja ostanek. V poskusu smo imeli tri pasme, dva meseca in 11 živali.

$$y_{ijk} = \mu + P_i + M_j + a_{ijk} + e_{ijk} \qquad [2.10]$$

Možni model 2.11 vključuje poleg vplivov iz osnovnega modela 2.10 še interakcijo med pasmo in mesecem (PM_{ij}).

$$y_{ijk} = \mu + P_i + M_j + PM_{ij} + a_{ijk} + e_{ijk} \qquad [2.11]$$

POMNI! Pri krajših modelih lahko opis modela navedemo v nekoliko zgoščeni obliki kar v besedilu, kot smo to naredili v zgoraj.

Do sem postopek opravimo prvič že pred pričetkom poskusa. Poleg tega preverimo število opazovanj, število (željenih) parametrov in število stopinj prostosti. Zaradi preglednosti bomo obdelali najprej vse korake pri določanju modelov, se nato vrnili k določanju omenjenih vrednosti in naredili še kakšen primer. Po opravljenem preizkusu vsekakor preverimo, če sta osnovni in možni model postavljena na začetku poskusa še vedno primerno izhodišče. Lahko bi sicer ugotovili, da so se pri izvedbi poskusa spreminjali pogoji, ki jih sicer nismo predvideli. Zaradi dobro vodenega dnevnika pa informacij nismo izgubili in jih lahko upoštevamo.

9. Statistično ovrednotenje opravimo s statistično analizo po izvedbi preizkusa. To delo boste lahko vadili na vajah, tu pa se bomo seznanili samo s pravili. Imamo deduktivni in induktivni postopek. Pri obeh metodah je kriterij verjetnost, da drži ničelna hipoteza. O hipotezah se bomo kasneje pogovarjali. Tule si bomo zapomnili le, da pri ničelni hipotezi predpostavimo, da ni razlik med nivoji pri izbranem vplivu. Vse ostale možnosti so pa predstavljene z alternativno hipotezo. Takrat, ko je bolj verjetna ničena hipoteza in je P-vrednost bliže 1.00 (pregl. 2.8), vpliva praviloma ne upoštevamo. Le v izjemnih primerih, ko je to najpomembnejši rezultat raziskave in je vključen v naslov naloge, ga bomo obdržali. V nasprotnem primeru bi pač zgrešili naslov. Ko pa so P-vrednosti bliže 0.00, pa je vpliv potrebno obravnavati, zlasti pri prvih preizkusih. Zaradi pomanjkanja stopinj prostosti lahko črtamo tudi vplive, kjer sta ničelna in alternativna hipoteza enakovredni. V majhnih a dragih preizkusih, ko smo uspeli opraviti le nekaj analiz oziroma meritev, morda izpustimo tudi druge vplive, da le niso značilni.

PRIMER: Vzemimo, da je P-vrednost 0.04. Verjetnost, da ničelna hipoteza drži, je zelo majhna, samo 4 %. Alternativna hipoteza je veliko bolj verjetna in sicer velja v 96 % primerov. Torej je vpliv najverjetneje precej pomemben in ga ne smemo zanemariti, čtrati iz modela.

***Deduktivni postopek*:** Obdelavo lahko začnemo s teoretičnim, možnim modelom in izločimo vplive, ki nimajo pomembnega vpliva v našem poskusu (P vrednost večja kot 0.60). Pri vrednostih med 0.60 in 0.40 se odločimo na osnovi strokovne presoje (število opazovanj in število parametrov, pričakovanja). Presodimo tudi, koliko slabši je model brez posameznih vplivov. Statistično značilne vplive vedno

Tabela 2.8: Kriteriji za izločanje vplivov

P-vrednost	Storimo:	Izjema:
0.00 - 0.05	vedno obdržimo	—
0.05 - 0.40	skoraj gotovo obdržimo	zmanjkuje stopinj prostosti
0.40 - 0.60	skoraj gotovo črtamo	vsi bodo čudno gledali, če vpliva ne bo
0.60 - 1.00	skoraj gotovo črtamo	je cilj raziskave

obdržimo v modelu in vplive s P vrednostjo pod 0.40 pa praviloma tudi. Praviloma izločamo najprej interkacije z največ sodelujočimi vplivi, nadrejene črtamo, ko v modelu ni več podrejenih. Vplive izločamo postopoma in raje ponovimo izračun. Postopek ponavljamo, dokler se model spreminja ... Zelo tudi pazimo, da v modelih z regresijo, dovolimo presečišča z ordinato (y - osjo).

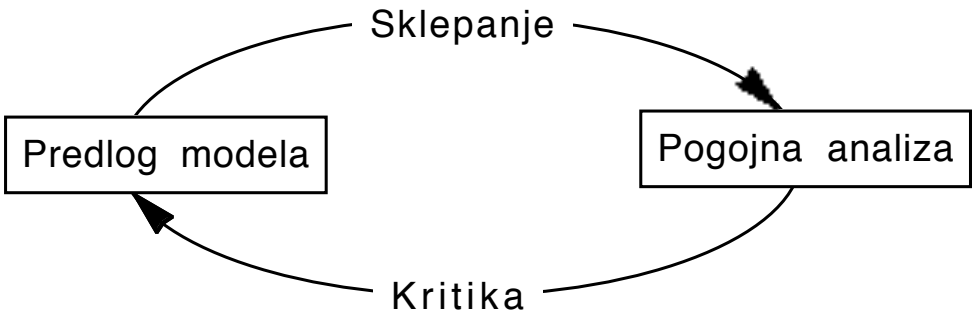
***Induktivni postopek*:** V primeru, ko imamo z možnim modelom težave, lahko začnemo z manjšim številom vplivov. Izberemo jih lahko tudi na osnovi osnovnih statističnih parametrov po vplivih, slikah, literaturi, itd. Model očistimo po zgornjem postopku in nato poizkušamo modele z dodanimi vplivi in spremljamo, koliko boljši je razširjeni model. Pri tem postopku se nam lahko zgodi, da katerega od pomembnih, značilnih vplivov ali kombinacije ne preizkusimo.

***Prilagajanje modela*:** Več bo v modelu parametrov, večji bo delež pojasnjene variance, bolj se bo model prilegal. Če izberemo toliko parametrov, kot imamo meritev, se bo model prilegal popolnoma. Ne bo pa uporaben. Vsekakor strmimo za enostavnejšim modelom (zakon skromnosti). To lahko zagotovimo samo s skrbno načrtovanim poskusom. Prav lahko pa se nam ob pridnem beleženju podatkov zgodi, da šele pri obdelavi odkrijemo vpliv, ki ga nismo načrtovali. Poruši se nam uravnotežena struktura podatkov, pojavijo se nivoji brez meritev, itd. Po drugi strani pa bi črtanje statistično (skoraj) značilnega sistematskega vpliva signifikantno povečalo varianco ostanka ter pripomoglo tudi k pristranski oceni. Ocene bi bile "onesnažene" z izpuščenim sistematskim vplivom. Črtanje slučajnega vpliva praviloma poveča varianco ostanka. V primeru, ko so nivoji izpuščenega vpliva med seboj korelirani, pa so lahko ocene v reduciranem modelu tudi pristransko ocenjene.

10. Strokovna presoja Model, ki se pri statistični presoji izkaže za najbolj primernega, moramo še temeljito strokovno presoditi. Težko bi se izognili strokovni razlagi statistično značilnega vpliva, pa čeprav nas sploh ni zanimal. Moramo ga vsaj na kratko odpraviti, kar je možno, ko je rezultat povsem pričakovan. Malo več težav bodo povzročali vplivi, ki niso v skladu s pričakovanji. Potolažite se lahko z dejstvom, da so predvsem taki rezultati spodbujali znanstvenike k razmišljanju in odkrivanju novih spoznanj. Temeljito se lotite študija narave tako vplivov kot lastnosti v modelu. Če ste se vam je porodila nova ideja, ponovite postopek od začetka... Morda pa je zaplet mogoče pojasniti le z novim poskusom. Tudi ugotovitev, da nam ni šlo vse po sreči, ni katastrofa, če vem, da smo uporabili vse do sedaj pridobljeno znanje iz literature. Tudi informacija, zakaj je nekaj šlo po nepričakovani poti, je lahko dragocena.

11. Proces izgradnje modela Kot vidimo, izgradnja modela ni enkraten proces. Vračati se moramo nazaj. Morda je dovolj, da nekajkrat ponovimo statistično ovrednotenje, prav nič nenavadno pa ni, da bomo morda spremenili listo vplivov. Morda bomo iskali ponovne informacije, ki bi se znali skrivati 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!



Slika 2.5: Proces izgradnje modela

Tabela 2.9: Krmni preizkus

Vrsta obroka	Žival	Čas (T)	
(D)	(a)	1	2
1	1	y_{111}	y_{112}
	2	y_{121}	y_{122}
2	3	y_{231}	y_{232}
	4	y_{241}	

VAJA 1:

Opazujemo štiri živali v dveh dneh. V preizkusu smo imeli tudi dva obroka, z vsakim smo krmili po dve živali. Merili smo lastnost "y" (npr. količino mleka).

a) Razvijte možni model, ko uporabite čas kot razred!

$$y_{ijk} = \mu + D_i + T_j + DT_{ij} + a_{ik} + aT_{ijk} + e_{ijk} \quad [2.12]$$

b) Razvijte možni model, ko uporabite čas kot neodvisno spremenljivko in jo vključite v model kot linearno regresijo!

$$y_{ijk} = \mu + D_i + \beta_i \left(x_{ijk} - \bar{x}\right) + a_{ij} + e_{ijk} \quad [2.13]$$

Čeprav imate rezultat nakazan, prikažite celoten postopek in model tudi pravilno opišite!

VAJA 2:

Vzemimo preizkus mladic (2.10). Napišite osnovni model za debelino hrbtne slanine in razvijte možni model!

Med glavne vplive bomo uvrstili pasmo (P_i), mesec (M_j), farmo (F_k), maso ob odbiri kot kvadratno regresijo in žival (a_{ijkl}).

$$y_{ijklm} = \mu + P_i + M_j + F_k + b_I(x_{ijkl} - 100) + b_{II}(x_{ijkl} - 100)^2 + a_{ijkl} + e_{ijklm} \quad [2.14]$$

$$y_{ijklm} = \mu + P_i + M_j + F_k + PM_{ij} + PF_{ik} + MF_{jk} + PMF_{ijk} + b_{Iijk}(x_{ijkl} - 100) + b_{IIijk}(x_{ijkl} - 100)^2 + a_{ijkl} + e_{ijklm} \quad [2.15]$$

V obeh modelih pomeni:

y_{ijklm} - opazovanje za debelino hrbtne slanine

Tabela 2.10: Podatki o preizkusu mladic na rast in zamaščenost na različnih treh farmah

Žival	Pasma	Mesec	Farma	Masa (kg)	Debelina slanine (mm)	Dnevni prirast (g/dan)
1	SL	JAN	A	102	12	13	506
2	SL	JAN	B	98	15	14	550
3	SL	FEB	C	105	15	16	532
4	SL	FEB	A	102	14	12	577
5	LW	JAN	B	95	19	17	512
6	LW	FEB	B	101	23	24	499
7	LW	FEB	C	101	26	24	466
8	NL	JAN	A	97	25	27	545
9	NL	JAN	C	100	21	19	549
10	NL	FEB	C	97	22	23	600
11	NL	FEB	B	102	23	15	610

Tabela 2.11: Določitev možnih interakcij za debelino hrbtno slanino pri preizkusu mladic

Vplivi							
P_i							
M_j	$: P_i$						
F_k	$: P_i$	$: M_j$					
b_I	$: P_i$	$: M_j$	$: F_k$				
b_{II}	$: P_i$	$: M_j$	$: F_k$	$: b_I$			
a_{ijkl}	$: P_i$	$: M_j$	$: F_k$	$: b_I$	$: b_{II}$		
PM_{ij}		$: M_j$	$: F_k$	$: b_I$	$: b_{II}$	$: a_{ijkl}$	
PF_{ik}		$: M_j$	$: F_k$	$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{Ii}		$: M_j$	$: F_k$	$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{IIi}		$: M_j$	$: F_k$	$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
MF_{jk}			$: F_k$	$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{Ij}			$: F_k$	$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{IIj}			$: F_k$	$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
PMF_{ijk}			$: F_k$	$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{Ik}				$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{IIk}				$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{Iik}				$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{IIik}				$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{Ijk}				$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{IIjk}				$: b_I$	$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{Iij}					$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{Iijk}					$: b_{II}$	$: a_{ijkl}$	$: PM_{ij}$
b_{Iij}						$: a_{ijkl}$	$: PM_{ij}$
b_{Iijk}						$: a_{ijkl}$	$: PM_{ij}$

Tabela 2.12: Seznam parametrov v modelu 2.16

Vpliv	Seznam parametrov	Obrazložitev
Srednja vrednost populacije	μ	
Pasma	P_1, P_2, P_3	tri pasme
Mesec / sezona	M_1, M_2	dva meseca
Žival	σ_a^2	naključni vpliv opišemo z varianco

μ - srednja vrednost populacije

P_i - vpliv pasme; i = 1, 2, 3

M_j - vpliv meseca; j = 1, 2

F_k - vpliv farme; k = 1, 2, 3, 4

PM_{ij}, PF_{ik}, MF_{jk} - dvojne interakcije med posameznimi vplivi

PMF_{ijk} - trojna interakcija med posameznimi vplivi

b_I, b_{Iijk} - linearni regresijski koeficienti

b_{II}, b_{IIijk} - kvadratni regresijski koeficienti

x_{ijkl} - masa živali ob odbiri

a_{ijkl} - vpliv živali, aditivni genetski vpliv, plemenska vrednost; l = 1,2, ..., n_{ijk}

e_{ijklm} - ostanek; m = 1, 2

n_{ijk} - število živali pasme i, v mesecu j in na farmi k

2.2.2 Število parametrov, stopinje prostosti in rang sistema

Ponovimo osnovni model za dnevni prirast v preizkusu z mladicami. V poizkusu smo imeli skupno 11 meritev, vsaka žival je imela natanko eno meritev.

$$y_{ijk} = \mu + P_i + M_j + a_{ijk} + e_{ijk} \qquad [2.16]$$

Število parametrov. Parametri so vrednosti, ki jih želimo oceniti, da bi poznali posamezne nivoje. Tokrat nas zanimajo predvsem lokacijski parametri pri sistematskih vplivih (2.12). Poskusimo najprej našteti vse parametre! Srednja vrednost populacije je ena sama, označena z μ. V preizkusu smo imeli tri pasme: švedska landrace (P_1), large white (P_2) in nemška landrace (P_3). Merili smo samo januarja (M_1) in februarja (M_2). Žival je naključni vpliv, zato je opisan s parametrom disperzije, z varianco (σ_a^2), in ne z lokacijskimi parametri.

Sedaj moramo samo prešteti število parametrov pri posameznemu vplivu in jih vpisati v preglednico 2.13. Parametri so neznanke, ki jih iščemo, in za vsako neznanko moramo v sistemu nastaviti eno enačbo. Prav tako bi radi izvrednotili plemenske vrednosti, torej so udi neznanke in moramo nastaviti po eno enačbo za vsako žival. Sistem enačb ima v našem primeru 17 enačb, ***red sistema*** je torej 17.

Tabela 2.13: Določanje števila parametrov in stopinj prostosti v modelu 2.16

Vplivi	μ	P_i	M_j	a_{ijk}	V modelu	V ostanku	Red	Rang
Število parametrov	1	3	2	(11)	6	ne določamo	6+11	
		3-1	2-1	ne določamo		11-4		
Število stopinj prostosti	1	2	1	-	4	7		4+11

Stopinje prostosti predstavljajo število ocenljivih parametrov. Ti parametri povedo vse o podatkih, če je izbrani model pravilen. Ostale ocene so le linearne kombinacije ocenljivih parametrov. Ko imamo npr. ocenjeno srednjo vrednost, je dovolj, da izvemo še oceni za dve pasmi, oceno tretje pa lahko dobimo s sklepanjem - lahko jo izračunamo iz omenjenih treh ocen.

Za izhodišče vzamemo število meritev. Vsaka meritev prinese eno stopinjo prostosti, torej imamo v našem preizkusu z dnevnim prirastom 11 stopinj prostosti. Dogovorimo se, da ste opravili preizkus vi. Torej so informacije vaše. Jaz pa bi rada napisala članek in bi rada od vas kupila informacije, ki zbrane podatke zadostno opišejo. Postavili ste mi visoko ceno, torej bom dobro premislila, kaj bom kupovala.

Največ o populaciji mi bo gotovo povedala kar srednja vrednost μ in je tudi najbolj zanesljiva ocena, saj smo za izračun porabili kar vse razpoložljive podatke. Ko sem plačala ceno, je ta informacija moja, vi pa imate eno manj.

Potem poskusimo s kupčijo pri vplivu pasme. Prav gotovo moram kupiti informacijo za prvo pasmo (P_1). Tudi pri drugi pasmi mi ne pomaga nič drugega kot plačati novo informacijo (P_2). Pri zadnjem nivoju pa z nekaj sklepanja lahko sama pridem do rezultata, saj velja enačba 2.17. Ker imam že tri informacije, lahko izračunam še četrto neznanko (P_3). S pasmami sem kupila dve informaciji, ki sem jih pridobila od vas.

$$\mu = \frac{P_1 + P_2 + P_3}{3} \qquad [2.17]$$

Pri mesecu lahko postopamo enako. Zadnjega mi ni treba kupiti, ostale pa moram. Ker sta bila samo dva meseca, potrebujem sam eno novo informacijo.

Živali pa je preveč in posamezne plemenske vrednosti so neprimerne za objavo, najti moramo drug način. Uporabili bomo varianco. Lokacijskih parametrov ne izračunavamo za naključne vplive, zato model ne dobi novih stopinj prostosti, tiste, ki so ostale, pa so ***stopinje prostosti za ostanek*** (2.19). Število ***stopinj prostosti za model*** (2.18) je vsota vseh stopinj prostosti za model.

$$s.p.\,za\,model = \sum p_i \qquad [2.18]$$

$$s.p.\,za\,ostanek = tevilo\,opazovanj - s.p.\,za\,model \qquad [2.19]$$

V možnem, teoretičnem modelu 2.23 imamo tudi interakcije. Vzemimo interakcijo med farmo (F_k) in pasmo (P_i), rezultate razmišljanja bomo zbrali v preglednici 2.14. Neznane parametre, ki jih moramo pridobiti iz podatkov, bomo navedli z nazivom. Tiste, ki jih lahko izvrednotimo iz že pridobljenih vrednosti, pa bomo predstavili kar z enačajem (=). Ker imamo v modelu tudi srednjo vrednost μ, lahko zadnjo farmo (F_4) in zadnjo pasmo (P_3) izvrednotimo na način predstavljen v enačbi 2.17. Na vsaki farmi imamo v poskusu tri pasme in nič drugega, prva dva parametra za moram pridobiti iz podatkov, tretjega pa lahko izvrednotimo, ker poznamo rezultat na farmi. To velja za prve tri farme. Pri četrti farmi pa lahko na rezultate že sklepamo (enačba 2.20). Rezultate pasme P_1smo že pridobili za prve tri, ostane nam edino še neznana proizvodnost pasme 1 na farmi 4 (PF_{14}), saj rezultat za prvo pasmo skupaj že imamo. Uganemo lahko, da dobimo stopinje prostosti pri interakcijah tako, da pomnožimo stopinje prostosti pri sodelujočih glavnih vplivih.

$$P_1 = \frac{PF_{11} + PF_{12} + PF_{13} + PF_{14}}{3} \qquad [2.20]$$

Tabela 2.14: Ocenljivi parametri in število stopinj prostosti pri križno klasificiranih vplivih

μ	F_1	F_2	F_3	=
P_1	PF_{11}	PF_{11}	PF_{11}	=
P_2	PF_{11}	PF_{11}	PF_{11}	=
=	=	=	=	=

Red in rang sistema. Red sistema predstavlja število neznank v modelu. Za vsak nivo pri vsakem vlivu, tako sistematskem kot naključnem, imamo po eno neznanko. Za vsako od neznank bomo rabili po eno enačbo, zato je število neznank in število enačb enako.

Sistem enačb lahko vsebuje odvisne enačbe. Nekatere enačbe so linearne kombinacije drugih. V naših sistemih so linearno odvisne tiste enačbe, kjer je število stopinj prostosti različno od števila parametrov. Pri sistematskem delu modela je rang enak številu stopinj prostosti, pri naključnem pa številu parametrov. Za mešani model dobimo rang tako, da seštejemo obe vrenosti.

$$rang \;=\; tevilo\ s.p.\ za\ model \;+\; tevilo\ parametrov\ za\ nakl ju£ni\ del\ modela \qquad [2.21]$$

Sedaj pa vajo ponovimo še za možni model za debelino hrbtne slanine (model 2.15). V modelu bomo tako imeli tri glavne vplive, tri dvojne in eno trojno interkacijo. Kvadratna regresija je vgnezdena znotraj trojne interakcije. Število parametrov pri glavnih vplivih je enako številu nivojev, pri interakciji in vgnezdeni regresiji (2.22) pa je enako produktu števila nivojev pri vseh sodelujočih vplivih.

$$p_{PMF} = p_P \bullet p_M \bullet p_F \qquad [2.22]$$

Najprej bomo napisali seznam vseh parametrov (2.15), ki naspopajo v modelu, in nato še ocenili število parametrov in stopinj prostosti. Poleg 48 regresijskih koeficientov je ocenljivih še 24 ostalih lokacijskih parametrov, ostali so linearne kombinacije teh 24. Pravzaprav bi zadoščal na videz preprostejši model 2.23. Možni (2.15) in spodnji model (2.23) sta tako ekvivalentna in dajeta iste rezultate. Pri spodnjem modelu sta red in rang sistema enačb enaka. Vsi dobimo iste rešitve, kar ni garantirano pri sistemih z nepolnim rangom, kot je to model 2.15. Model je polnega ranga in zato ima manj numeričnih problemov, je pa manj prikladen za interpretacijo.

$$y_{ijklm} = PMF_{ijk} + b_{Iijk}(x_{ijklm} - 100) + b_{IIijk}(x_{ijkl} - 100)^2 + a_{ijkl} + e_{ijklm} \qquad [2.23]$$

Tabela 2.15: Seznam parametrov in število stopinj prostostiv modelu 2.15

Vpliv	Seznam parametrov	Štev. parametrov		Štev. stopinj prostosti	
Srednja vrednost	μ	1	1		1
Pasma	P_1, P_2, P_3	3	3	3-1	2
Mesec	M_1, M_2	2	2	2-1	1
Farma	F_1, F_2, F_3, F_4	4	4	4-1	3
Interakcija med P in M	$PM_{11}, PM_{12}, PM_{21}, PM_{22}, PM_{31}, PM_{32}$	3 x 2	6	(3-1) x (2-1)	2
Interakcija med P in F	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 x 4	12	(3-1) x (4-1)	6
Interakcija med M in F	MF_{11}, MF_{12}, MF_{13}, MF_{14}, MF_{21}, $MF_{22}, MF_{23}, MF_{24}$	2 x 4	8	(2-1) x (4-1)	3
Interakcija med P, M in F	$PMF_{111}, PMF_{112}, PMF_{113}, \ldots PMF_{324}$	3 x 2 x 4	24	(3-1) x (2-1) x (4-1)	6
Regresijski koeficient za linearni člen	$b_{I111}, b_{I112}, b_{I113}, \ldots b_{I324}$	3 x 2 x 4	24	(3-1) x (2-1)	24
Regresijski koeficient za kvadratni člen	$b_{II111}, b_{II112}, b_{II113}, \ldots b_{II324}$	3 x 2 x 4	24	(3-1) x (2-1)	24
Žival	σ_a^2				-

VAJA 1: Pri miškah smo delali poskus s petimi različnimi krmami. Poskus smo izvajali v treh ponovitvah. V poskus smo vsakič vzeli 50 gnezd, kjer je bilo ob odstavitvi med 8 in 12 potomcev. V celoten poskus so bili vključeni trije očetje, ki so bili v vseh treh poskusih enakomerno zastopani. Odstavljene miške smo enakomerno porazdelili v skupine s krmo. Pri uvrščanju v skupine smo gledali tudi na to, da sta bila spola čim bolj enakomerno zastopana in da je bila povprečna odstavitvena masa med skupinami čim bolj enaka. Zanimala nas je rast mišk od odstavitve naprej. Živali smo tehtali vsak teden. Poskus je trajal mesec in pol.

- Razvijte osnovni in možni model in ju napišite!
- Določite število parametrov, stopinje prostosti, določite število podatkov, red in rang sistema!
- Kritično presodite model!

VAJA 2: Pri miškah smo delali poskus s petimi različnimi krmami. Poskus smo izvajali v treh ponovitvah. V poskus smo vsakič vzeli 50 gnezd, kjer je bilo ob odstavitvi med 8 in 12 potomcev. V vsakem poskusu so bili vključeni po trije različni očetje, ki so bili v poskusu enakomerno zastopani. Odstavljene miške smo enakomerno porazdelili v skupine s krmo. Pri uvrščanju v skupine smo gledali tudi na to, da sta bila spola čim bolj enakomerno zastopana in da je bila povprečna odstavitvena masa med skupinami čim bolj enaka. Zanimala nas je rast mišk od odstavitve do pubertete. Živali smo tehtali na začetku in koncu poskusa.

- Razvijte osnovni in možni model in ju napišite!
- Določite število parametrov, stopinje prostosti, določite število podatkov, red in rang sistema!
- Kritično presodite model!

VAJA 3: Naredili bomo poskus na pujskih. Iz vsakega gnezda bomo vzeli po 4 svinjke in 4 kastrate. Pri vsaki svinji bomo vzeli po 3 gnezda. Svinja je bila vedno parjena z istim merjascem, merjasec je bil parjen s 6 svinjami. V poskusu je bilo 5 merjascev. Pujske smo naključno razdelili enakomerno razdelili v 4 skupine, ki so dobivale različno vsebnost lizina v krmi in sicer 50 ng/kg, 60 ng/kg, 70 ng/kg in 80 ng/kg (vrednosti niso priporočila za prehrano prašičev). Stehtali jih bomo pri starosti 180 dni.

- Razvijte osnovni in možni model in ju napišite!
- Določite število parametrov, stopinje prostosti, določite število podatkov, red in rang sistema!

REŠITEV2.2.2:

Kot rešitev navajamo osnovni 2.36 in možni 2.37 model, izpustili smo opis parametrov, odvisnih in neodvisnih spremenljivk ter indeksov. Vrstni red vplivov je lahko različen, potem bodo lahko tudi indeksi pri vplivih različni, kar ni nič napačnega. V preglednici 2.18 povzemamo izračun števila parametrov in stopinj prostosti za možni model, za osnovnega preverite rezultate tako, da upoštevate samo tiste vplive, ki nastopajo v osnovnem, izhodiščnem modelu. Število podatkov ne moremo natančno predvideti, ker ni povsem jasno, koliko mladičev vzamemo iz gnezda. Vemo pa, da imamo 3 poskuse, pri vsakem imamo 50 gnezd, število mladičev v gnezdu pa variira med 8 in 10. Velikost gnezda nomo označili z n_{jkl}. Skupno število (2.26) torej dobimo tako, da seštejemo mladiče iz vseh gnezd. Pričakujemo, da bo v 150-tih gnezdih najbrž v poskusu nekako 1350 mladičev, lahko pa skupno število variira med 1200 do 1500 mladičev. V primeru možnega modela bi radi ocenili 270 parametrov. To nekako pomeni, da podatkov ni prav veliko, zato zelo upamo, da se bomo lahko znebili nekaterih členov in poenostavili model.

$$y_{ijklmn} = \mu + K_i + P_j + O_k + S_l + b_{Iijkl}\left(x_{ijklmn} - \bar{x}\right) + b_{IIijkl}\left(x_{ijklmn} - \bar{x}\right)^2 + g_{jkm} + e_{ijklnm} \qquad [2.24]$$

$$\begin{aligned} y_{ijklmn} = \quad & \mu + K_i + P_j + O_k + S_l + KP_{ij} + KO_{ik} + KS + PO_{jk} + PS_{jl} + OS_{kl} + \\ & + KPO_{ijk} + KPS_{ijl} + KOS_{ikl} + POS_{jkl} + KPOS_{ijkl} + b_{Iijkl}\left(x_{ijklmn} - \bar{x}\right) + \\ & + b_{IIijkl}\left(x_{ijklmn} - \bar{x}\right)^2 + g_{jkm} + Kg_{ijkm} + Sg_{jklm} + e_{ijklnm} \end{aligned} \qquad [2.25]$$

$$N = \sum_{j=1}^{3}\sum_{k=1}^{50} n_{jkm} \qquad [2.26]$$

REŠITEV 2.2.2:

V osnovni model 2.27 vključimo najprej spol (S_i) in krmo. Krme se razlikujejo zaradi različne vsebnosti lizina, vsebnosti so znane in jih bomo zato tretirali kot kvantitativni vpliv, ki jo predstavlja neodvisna spremenljivka $\left(x_{ijklm}\right)$. Brez lizina naj ne bi bilo krme, saj je aminokislina neobhodno potrebna. Presečišče z osjo y tako nima pravega pomena in bomo rezultate težko presojali. Bolje je, da ordinatno os y prestavimo na eno od vsebnosti, ki jih krma vsebuje. V tem primeru je lahko najmanjša vsebnost (x_{min})povsem primerna. Lizin je draga sestavina in bi bili kar zadovoljni, če ga lahko dodamo v manjših količinah. Torej se ima smisel osredotočiti na to točko. Rezultati niso zaradi tega prav nič potvorjeni, ne glede, kam bomo premaknili ordinato, morajo biti zaključki enaki. Tule bomo predpostavili kar linearno regresijo, imamo samo en regresijski koeficient (b), čeprav je potrebno to pri obdelavi preveriti. Na rezultate lahko vpliva tudi izbor merjascev $\left(M_j\right)$, ker niso naključni vzorec iz populacije moških prašičev. Dobro so se izkazali v preizkusu na proizvodne lastnosti. Podobno, čeprav manj intenzivno, so bile odbrane tudi svinje. Meritev po svinji je kar nekaj, saj ima v preizkusu tri gnezda z 8 pujski in lahko ocenimo oziroa odstranimo njen vpliv. Svinje $\left(F_{jk}\right)$ so vgnezdene znotraj merjasca. Po gnezdu imamo veliko meritev, le po 8 pujskov s po eno meritvijo. Pri navadnem krmnem poskusu bi ga nadvse radi izpustili, morda pa bi ga uspeli nadomestiti s čim drugim. Poskrbeli smo tudi za enakomerno zastopanost spolov in enakomerno porazdelitev po drugih vplivih. V takih uravnoteženih poskusih izpustitev gnezda kot vpliva ni napaka. V našem poskusu ga bomo torej z veseljem zavrgli, bomo ga pa najprej preverili. Ker je malo podatkov po gnezdu in kar nekaj gnezd v poskusu, bomo vpliv gnezda obravnavali kot naključni vpliv $\left(g_{jkl}\right)$. Gnezdo je vgnezdeno znotraj svinje in merjasca, a križno klasificirano s spolom.

$$y_{ijklm} = \mu + S_i + b\left(x_{ijklm} - x_{min}\right) + M_j + F_{jk} + g_{jkl} + e_{ijklm} \qquad [2.27]$$

Možni model 2.28 vključuje glavne vplive iz osnovnega modela 2.27. Pridobili smo tudi interakcije med spolom in očetom $\left(SM_{ij}\right)$ oziroma svinjo $\left(SF_{ijk}\right)$. Pri regresiji je možna vgnezditev znotraj interakcije SF_{ijk}, čeprav je povsem nepraktična. Če bi to res obstajalo, bi za vsako svinjo morali narediti krmni poskus in sicer v kombinaciji z vsako krmo. Za nameček bi jo morali svinjo krmiti z dvema krmama:

Tabela 2.16: Izračun stopinj prostosti za model 2.37

Vpliv	Seznam parametrov	Štev. parametrov		Štev. stopinj prostosti	
Srednja vrednost	μ	1	1		1
Krma	K_1, K_2, K_3, K_4, K_5	5	5	5-1	4
Poskus	P_1, P_2, P_3	3	3	3-1	2
Oče	O_1, O_2, O_3	3	3	3-1	2
Spol	S_1, S_2	2	2	2-1	1
Interakcija med K in P	$KP_{11}, \cdots KP_{ij}, \cdots, KP_{53}$	5 x 3	15	(5-1) x (3-1)	8
Interakcija med K in O	$KO_{11}, \cdots KO_{ik}, \cdots, KO_{53}$	5 x 3	15	(5-1) x (3-1)	8
Interakcija med K in S	$KS_{11}, \cdots KS_{il}, \cdots, KS_{52}$	5 x 2	10	(5-1) x (2-1)	4
Interakcija med P in O	$PO_{11}, \cdots PO_{jk}, \cdots, PO_{33}$	3 x 3	9	(3-1) x (3-1)	4
Interakcija med P in S	$PS_{11}, \cdots PS_{jk}, \cdots, PS_{32}$	3 x 2	6	(3-1) x (2-1)	2
Interakcija med O in S	$OS_{11}, \cdots OS_{jk}, \cdots, OS_{32}$	3 x 2	6	(3-1) x (2-1)	2
Interakcija med K, P in O	$KPO_{111}, \cdots KPO_{ijk}, \cdots, KPO_{533}$	5 x 3 x 3	45		16
Interakcija med K, P in S	$KPS_{111}, \cdots KPS_{ijl}, \cdots, KPS_{532}$	5 x 3 x 2	30		8
Interakcija med K, O in S	$KOS_{111}, \cdots KOS_{ijl}, \cdots, KOS_{532}$	5 x 3 x 2	30		8
Interakcija med P, O in S	$POS_{111}, \cdots POS_{ijl}, \cdots, POS_{332}$	3 x 3 x 2	18		4
Interakcija med K, P, O in S	$KPOS_{1111}, \cdots KPOS_{ijkl}, \cdots, KPOS_{5332}$	5x3x3x2	90	(5-1)x (3-1)x (3-1) x (2-1)	16
Regresijski koeficient za linearni člen	$b_{I1111}, \cdots, b_{Iijkl}, \cdots .. b_{I5332}$	5x3x3x2	90		90
Regresijski koeficient za kvadratni člen	$b_{I1111}, \cdots, b_{Iijkl}, \cdots .. b_{I5332}$	5x3x3x2	90		90
Gnezdo	σ^2_g		(150)		
Interakcija med K in g	σ^2_{Kg}		(750)		
Interakcija med S in g	σ^2_{Sg}		(300)		

Tabela 2.17: Ocenljivi parametri in stopinje prostosti pri vgnezdenih vplivih

	Merjasci				
Svinja	M_1	M_2	M_3	M_4	=
1	F_{11}	F_{21}	F_{31}	F_{41}	F_{51}
2	F_{12}	F_{22}	F_{32}	F_{42}	F_{52}
3	F_{13}	F_{23}	F_{33}	F_{43}	F_{53}
4	F_{14}	F_{24}	F_{34}	F_{44}	F_{54}
5	F_{15}	F_{25}	F_{35}	F_{45}	F_{25}
6	=	=	=	=	=

ena bi bila bolj primerna za svinjke, druga za kastrate. Ko bi poskus za silo končali, bi bila svinja že ostarela in primerna za izločitev. Kaj bi nam rezultati sploh pomagali, če rezultati ne veljajo tudi za druge svinje?

Čemu torej tako kompleksen model na začetku? Z veseljem bomo vsak sestavljen vpliv in tudi kakšnega od glavnih črtali, ko bomo potrdili, da niso značilni. Model ima smisel le, če je obvadljiv, če ga lahko razložimo. Toda ne smemo se vnaprej odločiti, katere vplive bomo vključili v model. V postopku izgradnje modela preverimo, če je poskus potekal, kot smo si zastavili. Celo vpliva svinj in merjascev bi se radi iznebili. Reja prašičev bi bila prava nočna mora za gospodarja, če bi vsaki svinji moral pripraviti poseben obrok! Lahko pa bi se izkazalo, da je stara mati svoji svinji prinašala posebke. Taka ljubljenka ali bolna žival bi v poskusu nagajala. V takšni analizi bi jo našli in odstranili iz analize.

V enem povsem resnem poskusu pitanja bikov se je zgodil nenavaden primer, ki dobro kaže na to, da se je potrebno lotiti analize sistematično. V poskusu je najlepše, že kar neverjetno, priraščal bik, ki je pojedel najmanj krme, celo tako malo, da bi moral beljakovine graditi iz dušika iz zraka. Raziskovalci smo napenjali možgane, a nismo našli problema. Klepet s oskrbnikom hleva pa je pokazal, da gre za sila navihanega bika. Najprej je pomagal pospraviti krmo v jaslih svojega desnega in levega soseda, nekaj njegove krme pa mu je celo ostalo v jaslih, saj kasneje pač ni bil več pripravljen deliti krme s kolegoma. Tudi pregled podatkov je pokazal, da je bila pripoved resnična.

Dobili smo še eno interakcijo in sicer med spolom in gnezdom $\left(S\,g_{ijkl}\right)$. Vpliv je naključni: gnezdo smo razdelili na dve skupini in sicer po spolu. V eni skupinici je le polovica opazovanj v primerjavi z gnezdom, skupin pa je še enkrat toliko. Interakcija, ki ima vsaj en naključni vpliv, je naključna.

$$\begin{aligned} y_{ijklm} = \quad & \mu + S_i + M_j + F_{jk} + S\,M_{ij} + S\,F_{ijk} + \\ & + b_{ijk}\left(x_{ijklm} - x_{min}\right) + g_{jkl} + S\,g_{ijkl} + e_{ijklm} \end{aligned} \qquad [2.28]$$

Število opazovanj (2.15) lahko natančno izračunamo. Opravili smo 720 tehtanj.

$$N = 8\ pujskov\ x\ 3\ gnezda\ x\ 6\ svinj\ x\ 5\ merjascev\ = 720 \qquad [2.29]$$

Najprej poskusimo ugotoviti število stopinj prosti pri vgnezdenih vplivih.

2.2.3 Pričakovane vrednosti

Pričakovano vrednost ali matematično upanje bomo označevali z veliko črko E, za njo pa v oklepaju navedemo spremenljivko, matematični izraz ali statistični model. Pričakovano vrednost opazovanja y_{ijk} bomo dobili tako, da opazovanje v izrazu nadomestimo z modelom [2.30]. Operacije za izračun pričakovane vrednosti so podobne operacijam za vsoto.

Tabela 2.18: Izračun stopinj prostosti za model 2.37

Vpliv	Seznam parametrov	Štev. parametrov		Štev. stopinj prostosti	
Srednja vrednost	μ	1	1		1
Spol	S_1, S_2	2	2	2-1	1
Merjasec	M_1, M_2, M_3, M_4, M_5	5	5	5-1	4
Svinja	$F_{11}, \cdots, F_{jk}, \cdots, F_{56}$	5 x 6	30	5 x (6-1)	25
Interakcija med S in M	$SM_{11}, \cdots SM_{ij}, \cdots, SM_{25}$	2 x 5	10	(2-1) x (5-1)	4
Interakcija med S in F	$SF_{111}, \cdots SF_{ijk}, \cdots, SF_{256}$	2 x 5 x 6	60	(2-1) x 5 x (6-1)	25
Regresijski koeficient za linearni člen	$b_{111}, \cdots, b_{ijk}, \cdots .. b_{256}$	2 x 5 x 6	60	2 x 5 x 6	60
Gnezdo	σ^2_g		(150)		
Interakcija med S in g	σ^2_{Sg}		(300)		

Torej pri primeru mladic:

$$\begin{aligned} E\left(y_{ijk}\right) &= E\left(\mu + P_i + b\left(x_{ijk} - \bar{x}\right) + a_{ij} + e_{ijk}\right) \\ &= E\left(\mu\right) + E\left(P_i\right) + E\left(b\left(x_{ijk} - \bar{x}\right)\right) + E\left(a_{ij}\right) + E\left(e_{ijk}\right) \\ &= E\left(\mu\right) + E\left(P_i\right) + \left(x_{ijk} - \bar{x}\right) E\left(b\right) + E\left(a_{ij}\right) + E\left(e_{ijk}\right) \\ &= \mu + P_i + b\left(x_{ijk} - \bar{x}\right) + 0 + 0 \end{aligned} \quad [2.30]$$

Pričakovane vrednosti za konstante in neslučajne spremenljivke so konstante oziroma parametri sami po sebi. Konstante imajo določeno vrednost (npr. 1, 10, -2.4, 0, včasih jih označujemo s črkami na začetku abecede).

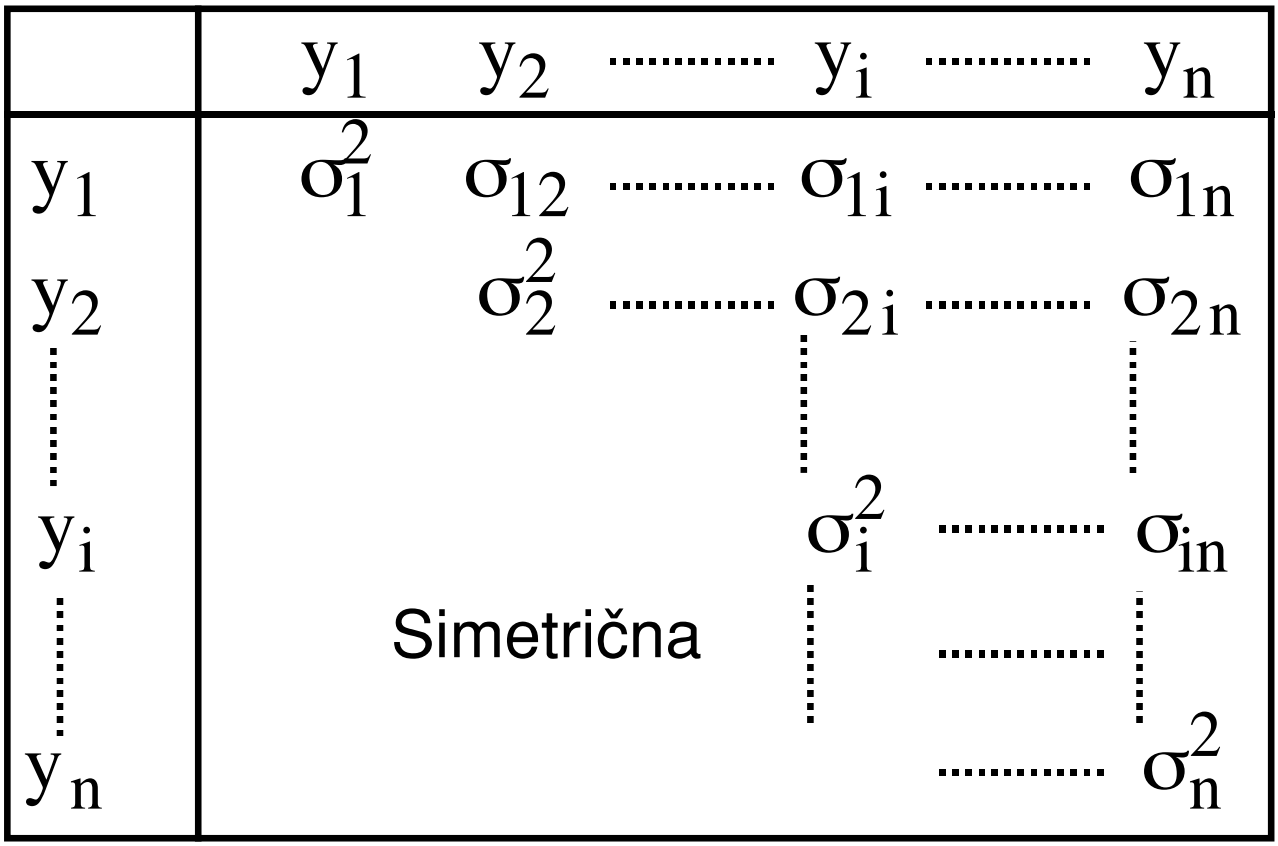
Sistematske (neslučajne) spremenljivke se obnašajo kot konstante, vendar pa imamo na voljo večjo zalogo vrednosti. Vsaka od teh vrednosti pa je konstanta. Kot primer si lahko ogledamo vpliv pasme. V populaciji ihanskih prašičev imamo štiri pasme: švedska landrace, large white, duroc in nemška landrace. Za naključno izbranega prašiča duroc pričakujemo, da bo njegova proizvodnja enaka povprečni proizvodnji pasme duroc. To velja, dokler nimamo dodatnih informacij, npr. meritve na prašiču ali sorodnikih. Vendar pa s temi dodatnimi informacijami govorimo o pogojih. Torej je nivo proizvodnje vnaprej določen - sistematski (sistematski vplivi). Niso variabilni in med njimi ne obstaja podobnosti (korelacij, kovarianc). Če poznamo vrednosti nekaterih razredov sistematskih spremenljivk (pasem), ne moremo sklepati na lastnosti drugih. K sistematskim spremenljivkam štejemo tudi neodvisne spremenljivke, ki nastopajo v modelu kot kovariable. Spremenljivke kot so pasma, sezona, hlev, itd., imajo končno zalogo vrednosti. Tako smo omenili, da imajo na farmi v Ihanu le štiri pasme. Vrednosti so nezvezne in pogosto kvalitativne. Neodvisne spremenljivke so lahko zvezne ali nezvezne, vedno pa so kvantitativne (telesna teža, količina uporabljenega zdravila, čas, velikost, koncentracije, itd.). Neodvisne spremenljivke bomo označevali z malimi črkami proti koncu abecede (x, u, w), ostale sistematske spremenljivke z velikimi črkami, običajno z začetnico naziva. Za glavne spremenljivke uporabljamo le po eno črko.

$$E\left(\mu + P_i + a_{ij} + e_{ijk}\right) = \mu + P_i \quad [2.31]$$

Pričakovana vrednost pri pogoju:

$$E\left(y_{ijk} | \, a_{ij} = \text{"}06 - 4965 - 23\text{"}\right) = \mu + P_i + a_{ij} \quad [2.32]$$

Za slučajne spremenljivke bomo privzeli, da je njihova pričakovana vrednost nič. V nasprotnem primeru, je model možno preurediti tako, da naša predpostavka drži. Pričakovana vrednost je tako in tako posledica nekega sistematskega vpliva. Namesto, da si otežujemo obdelavo, je primerneje ta sistematski vpliv



Slika 2.6: Matrika kovarianc med opazovanji

vključiti v model (npr. genetske skupine). Slučajne spremenljivke povzroča individualna odstopanja od pričakovane vrednosti. So variabilne in med njimi lahko nastopa podobnost (sorodne živali). Slučajna spremenljivka je plemenska vrednost živali, ostanek, skupno okolje v gnezdu. Zaloge vrednosti so lahko končne ali neskončne množice, imajo pa (poznano) porazdelitev. Ko poznamo nekaj spremenljivk, lahko napovedujemo (prediction) o lastnostih ostalih elementov množice (ostalih živali iz populacije).

2.2.4 Struktura varianc in kovarianc

Zanima nas variabilnost opazovanj in podobnost (kovarianca) med njimi. Najenostavneje se problema lotimo na ta način, da nanizamo opazovanja v stolpični in vrstični vektor (slika 2.6) in jih potem po parih primerjamo.

Na sliki 2.6 dovoljujemo, da imajo opazovanja različno varianco (različno napako meritev) in tudi kovariance so lahko različne. Če poznamo variance in kovariance, potem bomo lahko ocenili pričakovane vrednosti. V primeru pa, da želimo oceniti iz podatkov tudi variance (napake meritev) in kovariance, pa bi pri taki strukturi imeli več enačb kot parametrov. Da lahko ocenimo varianco, rabimo ponovitve - poiskuse pod istimi pogoji. To pa pomeni, da je struktura varianc bolj urejena.

Običajne predpostavke so:

- ostanki (napake meritev) so med seboj nekorelirani, če niso merjeni na isti enoti (živali)
- živali v poskusu so sorodne (imamo podatke o poreklu), niso sorodne, (ni)so inbridirane...
- živali iz istega gnezda imajo skupno okolje: iz različnih gnezd pa si niso podobne, če niso v sorodu

- živali od istega rejca imajo skupno okolje: živali različnih rejcev si niso podobne, če niso v sorodu, itd. (matrika sorodstva).

Opazovanje nadomestimo z modelom. Operacije so podobne množenju.

$$var\left(y_{ijk}\right) = var\left(\mu + P_i + b\left(x_{ijk} - \bar{x}\right) + a_{ij} + e_{ijk}\right) \quad [2.33]$$

Enačbo razrešimo.

$$\begin{aligned} = \quad & var\left(\mu\right) + var\left(P_i\right) + var\left(b\left(x_{ijk} - \bar{x}\right)\right) + var\left(a_{ij}\right) + var\left(e_{ijk}\right) + \\ & 2cov\left(\mu, P_i\right) + 2cov\left(\mu, b\left(x_{ijk} - \bar{x}\right)\right) + 2cov\left(\mu, a_{ij}\right) + 2cov\left(\mu, e_{ijk}\right) + \\ & 2cov\left(P_i, b\left(x_{ijk} - \bar{x}\right)\right) + 2cov\left(P_i, a_{ij}\right) + 2cov\left(P_i, e_{ijk}\right) + \\ & 2cov\left(b\left(x_{ijk} - \bar{x}\right), a_{ij}\right) + 2cov\left(b\left(x_{ijk} - \bar{x}\right), e_{ijk}\right) + \\ & 2cov\left(a_{ij}, e_{ijk}\right) \end{aligned} \quad [2.34]$$

Varianca parametra in neodvisne spremenljivke je enaka vrednosti 0. To si je lahko predstavljati pri kvalitativnih lastnostih (npr. pasme). Ne moremo jih razvrstiti po velikosti. Ko poznamo eno, ne vemo nič o drugi. Pri vsaki meritvi je nivo poznan in določen. Tako ni nič naključnega (varianca=0) in nivoji niso sorodni (kovarianca = 0) Kovarianca med sistematskima vplivoma in med sistematskim ter naključnim je tudi enaka vrednosti 0 zaradi značaja sistematskih vplivov (obnašajo se kot konstante).

$$= var\left(a_{ij}\right) + var\left(e_{ijk}\right) + 2cov\left(a_{ij}, e_{ijk}\right) \quad [2.35]$$

2.2.5 Predpostavke in restrikcije

Pri kovariancah pogosto predpostavimo, da so enake nič ali pa so vezane na neko skupno okolje:

- soroden genotip (genetske kovariance)
- skupno okolje v gnezdu
- skupno okolje v čredi.

Pogosto predpostavimo, da obstajajo kovariance le med nivoji, razredi naključnih spremenljivk. Če pa imamo v modelu več aditivnih genetskih vplivov (neposredni, maternalni), pa si te predpostavke ne smemo privoščiti!

Pri poskusu predpostavimo, katere kovariance so lahko nič, ostale pa moramo bodisi poznati bodisi zagotoviti primerno strukturo podatkov za analizo.

Predpostavke o porazdelitvi, ki niso opisane z zgornjimi parametri.

ID - identično porazdeljene spremenljivke (vse smo izmerili z istim ostankom)

IID- identično in neodvisno (independent - ostanki med nivoji niso korelirani) porazdeljene spremenljivke

IIDN - IID normalno porazdeljene spremenljivke

1. Opiši konstantne pogoje (standardizirano okolje), npr. vse merjene živali so istega spola in starosti. Drugi vplivi morajo biti v modelu.
2. Opiši časovne variable (sezona)
3. Pomni, da so neomenjeni vplivi tretirani od bralcev kot konstantni ali pa vzbujajo sume o zasnovi poskusa.

Lahko je tudi več točk - odvisno od modela. Model ni popoln, če niso opisani prav vsi deli modela. Le tako je možno poznati vse omejitve pri modelu ali poskusu. Model se lahko vedno izboljša, poskus pa bi morda bil predrag in prekasen, če bi ga hoteli spremeniti (dopolniti) ali ponoviti. Torej model bi morali napisati pred postavitvijo poskusa. S tem bi morda uspeli izboljšati plan poskusa, hkrati pa bi si zagotovili, da bi pravilno postavili in testirali hipoteze.

Vaje

1. Napišite teoretični model za dnevni prirast iz preglednice **??** in pojasnite odločitve.
2. Napišite teoretični model za debelino hrbtne slanine iz preglednice **??** in pojasnite odločitve.
3. Izvrednotite pričakovane vrednosti in ugotovite strukturo variance za model

$$y_{ijklmn} = \mu + K_i + P_j + O_k + S_l + b_{Iijkl}\left(x_{ijklmn} - \bar{x}\right) + b_{IIijkl}\left(x_{ijklmn} - \bar{x}\right)^2 + g_{jkm} + e_{ijklnm} \quad [2.36]$$

Izvrednotite pričakovane vrednosti, varianco in kovarianco za naslednje izraze (O vsebini, pomenu in smislu modela ali posameznih členov ne razmišljajte! O linearnih kombinacijah in njihovih interpretacijah se bomo ukvarjali v nadaljevanju.)

$\left(y_{ijklmn}\right) =$

$\left(y_{ijklmn}|g_{jkm}\right) =$

$\left(S_l + b_{Iijkl}\left(x_{ijklmn} - \bar{x}\right)\right) =$

$\left(S_l + b_{Iijkl}\left(x_{ijklmn} - \bar{x}\right)|g_{jkm}\right) =$

$\left(\mu + 3P_j + \frac{1}{2}O_k + \frac{1}{4}g_{jkm} + e_{ijklnm}\right) =$

$\left(\mu + 3P_j + \frac{1}{2}O_k + \frac{1}{4}g_{jkm} + e_{ijklnm}|g_{jkm}\right) =$

$\left(\frac{1}{4}g_{jkm} + e_{ijklnm}\right) =$

$\left(\frac{1}{4}g_{jkm} + e_{ijklnm}|g_{jkm}\right) =$

4 Izvrednotite pričakovane vrednosti in ugotovite strukturo variance za model

$$\begin{aligned} y_{ijklmn} = \quad & \mu + K_i + P_j + O_k + S_l + KP_{ij} + KO_{ik} + KS + PO_{jk} + PS_{jl} + OS_{kl} + \\ & + KPO_{ijk} + KPS_{ijl} + KOS_{ikl} + POS_{jkl} + KPOS_{ijkl} + b_{Iijkl}\left(x_{ijklmn} - \bar{x}\right) + \\ & + b_{IIijkl}\left(x_{ijklmn} - \bar{x}\right)^2 + g_{jkm} + Kg_{ijkm} + Sg_{jklm} + e_{ijklnm} \end{aligned} \quad [2.37]$$

Izvrednotite pričakovane vrednosti, varianco in kovarianco za naslednje izraze (O vsebini, pomenu in smislu modela ali posameznih členov ne razmišljajte! O linearnih kombinacijah in njihovih interpretacijah se bomo ukvarjali v nadaljevanju.):

$\left(y_{ijklmn}\right) =$

$\left(y_{ijklmn}|g_{jkm}\right) =$

$\left(y_{ijklmn}|Kg_{ijkm}\right) =$

$\left(KPS_{ijl} + KOS_{ikl} + POS_{jkl} + KPOS_{ijkl} + b_{Iijkl}\left(x_{ijklmn} - \bar{x}\right)\right) =$

$\left(KPS_{ijl} + KOS_{ikl} + POS_{jkl} + KPOS_{ijkl} + b_{Iijkl}\left(x_{ijklmn} - \bar{x}\right)|Kg_{ijkm}\right) =$

$\left(\mu + K_i + P_j + O_k + g_{jkm} + Kg_{ijkm} + Sg_{jklm} + e_{ijklnm}|Kg_{ijkm}\right) =$

$\left(KPS_{ijl} + KOS_{ikl} + POS_{jkl} + KPOS_{ijkl} + b_{Iijkl}\left(x_{ijklmn} - \bar{x}\right)|g_{jkm}, Kg_{ijkm}\right) =$

$\left(\mu + K_i + P_j + O_k + g_{jkm} + Kg_{ijkm} + Sg_{jklm} + e_{ijklnm}|g_{jkm}, Kg_{ijkm}\right) =$

2.2.6 Linearni, pogojno linearni in nelinearni modeli

Modele razdelimo na linearne, pogojno linearne in (pogojno) nelinearne modele. Razdelitev opravimo na osnovi prvih parcialnih odvodov odvisne spremenljivke z ozirom na vse parametre. Če se v nobenem parcialnem odvodu ne pojavi parameter, je model linearen. Nekateri modeli po naravi niso linearni, najdemo pa tako transformacijo, da transformiran model zadosti pogojem. V nekaterih primerih pa se lahko poslužimo tudi aproksimacije zahtevne nelinearne funkcije. To velja še posebej, če opazujemo spremenljivke le na manjšem intervalu.

Linearni modeli so zaželjeni, ker so računsko manj zahtevni in enostavnejši za interpretacijo. Pogosto pa pojava ne moremo poenostaviti tako preprosto. Rast organizmov od rojstva do odrasle velikosti in razgradnja hrane od zaužitja do neprebavljivih ostankov, koagulacija mleka, laktacijske krivulje so pogosto proučevane lastnosti v živinoreji, ki jih moremo v celoti opisati z linearnimi modeli. Le ti zadostujejo samo na zadostno majhnih intervalih.

PRIMER:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk} \quad [2.38]$$

kjer pomeni:

y_{ijk} - opazovanje

μ - srednja vrednost

α_i - sistematski vpliv α; i = 1, 2, ... p

β_j - sistematski vpliv β; j = 1, 2, ... q

$\alpha\beta_{ij}$ - interakcija med vplivoma α in β

e_{ijk} - ostanek; k = 1, 2, ... n_{ij}

V modelu 2.38 smo za sistematske vplive uporabili male grške črke, kar bomo v skalarni obliki bolj redko uporabljali. Uporabimo jih predvsem, kadar iz matrične oblike izpeljujemo skalarno obliko. Zanima nas, če je model linearen ali nelinearen. Poiskati moramo vse parcialne odvode z ozirom na vse nezname parametre.
Pri odvajanju se moramo navaditi, da so neznanke parametri iz modela. Namesto opazovanja y_{ijk} vnesemo desno stran modela 2.38, ki pojasnjuje variabilnost v modelu. Izraz lahko razstavimo na člene. Vplivi so neodvisni, zato so odvodi členov, ki jih odvajamo na drug parameter kot ga vsebuje v izrazu, nadomestimo z vrednostjo nič. Tudi ostanki so neodvisni od parametrov in člene z ostanki nadomestimo z vrednostjo odvoda, ki je enak 0. Le členi, ki vključujejo parametre, na katere odvajamo, dobijo od 0 različno.
Najprej odvajajmo na parameter μ (enačba 2.39). V prvem členu dobimo vrednost 1, pri ostalih pa 0, ker so parametri neodvisni od μ.

$$\frac{\partial y_{ijk}}{\partial \mu} = \frac{\partial \left(\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}\right)}{\partial \mu} = \frac{\partial \mu}{\partial \mu} + \frac{\partial \alpha_i}{\partial \mu} + \frac{\partial \beta_j}{\partial \mu} + \frac{\partial \alpha\beta_{ij}}{\partial \mu} + \frac{\partial e_{ijk}}{\partial \mu} = 1 + 0 + 0 + 0 + 0 \quad [2.39]$$

Nadaljujmo z vplivom α. V tej skupini imamo p med seboj neodvisnih parametrov. Za vsak parameter moramo poiskati parcialni odvod. Enačbe bodo podobne enačbi 2.40, kjer bi menjali le indekse i' in sicer od 1, 2 do p. Tam, kjer sta i in i' enaka, dobimo vrednost različno od 0. V vsaki enačbi je le en člen tak, ostali pa vsebujejo druge parametre.

$$\frac{\partial y_{ijk}}{\partial \alpha_{i'}} = \frac{\partial \mu}{\partial \alpha_{i'}} + \frac{\partial \alpha_i}{\partial \alpha_{i'}} + \frac{\partial \beta_j}{\partial \alpha_{i'}} + \frac{\partial \alpha\beta_{ij}}{\partial \alpha_{i'}} + \frac{\partial e_{ijk}}{\partial \alpha_{i'}} = 0 + \left\{ \begin{array}{ll} 1; & i = i' \\ 0; & i \neq i' \end{array} \right\} + 0 + 0 + 0 \quad [2.40]$$

Pazite, da vas ne zavede interakcija. Oznaka za interakcijo vsebuje sicer oznaki glavnih vplivov, vendar pa ne predstavlja produkta! Je le oznaka, ki bi jo lahko zamenjali s povsem drugo oznako. Torej interakcije ne moremo izraziti kot funkcijo glavnih vplivov, je torej od njiju (ali njih) neodvisna. Vrednost odvodov, kjer nastopa interakcija in eden od glavnih vplivov, je 0.
Tudi odvoda na parametre za vpliv β (2.41) in interakcijo $\alpha\beta$ (2.42) dobimo po istem zgledu.

$$\frac{\partial y_{ijk}}{\partial \beta_{j'}} = \frac{\partial \mu}{\partial \beta_{j'}} + \frac{\partial \alpha_i}{\partial \beta_{j'}} + \frac{\partial \beta_j}{\partial \beta_{j'}} + \frac{\partial \alpha\beta_{ij}}{\partial \beta_{j'}} + \frac{\partial e_{ijk}}{\partial \beta_{j'}} = 0 + 0 + \left\{ \begin{array}{ll} 1; & j = j' \\ 0; & j \neq j' \end{array} \right\} + 0 + 0 \quad [2.41]$$

$$\frac{\partial y_{ijk}}{\partial \alpha\beta_{i'j'}} = \frac{\partial \mu}{\partial \alpha\beta_{i'j'}} + \frac{\partial \alpha_i}{\partial \alpha\beta_{i'j'}} + \frac{\partial \beta_j}{\partial \alpha\beta_{i'j'}} + \frac{\partial \alpha\beta_{ij}}{\partial \alpha\beta_{i'j'}} + \frac{\partial e_{ijk}}{\partial \alpha\beta_{i'j'}} = 0+0+0+\left\{ \begin{array}{ll} 1; & i = i' \wedge j = j' \\ 0; & i \neq i' \vee j \neq j' \end{array} \right\} + 0 \quad [2.42]$$

Vsi prvi odvodi so brez parametrov, torej je model 2.38 linearen.

PRIMER:

Vzemimo model iz vaje 2.2.2. Prepišimo model in ga razširimo še za kvadratni člen pri regresiji.

$$y_{ijklm} = \mu + S_i + b_1 \left(x_{ijklm} - x_{min}\right) + b_2 \left(x_{ijklm} - x_{min}\right)^2 + M_j + F_{jk} + g_{jkl} + e_{ijklm} \quad [2.43]$$

kjer pomeni:

y_{ijklm} - opazovanja

μ - srednja vrednost

S_i - vpliv spola; i = 1, 2

b_1 - regresijski koeficient za linearni člen

b_2 - regresijski člen za kvadratni člen

x_{ijklm} - neodvisna spremenljivka za vsebnost lizina v krmi

x_{min} - minimalna količina lizina v krmi

M_j - vpliv merjasca; j = 1, 2, 3, 4, 5

F_{jk} - vpliv svinje; k = 1, 2, ... 6

g_{jkl} - vpliv skupnega okolja v gnezdu; l = 1, 2, 3

e_{ijklm} - ostanek; m = 1, 2, ... 4

Sedaj pa poskusimo odvajati. Prva neznanka je μ. Prvi parcialni odvod prikazujemo v enačbi 2.44. Samo pri prvem členu je odvod različen od 0.

$$\frac{\partial y_{ijk}}{\partial \mu} = \frac{\partial \mu}{\partial \mu} + \frac{\partial S_i}{\partial \mu} + \frac{\partial b_1 \left(x_{ijklm} - x_{min}\right)}{\partial \mu} + \frac{\partial b_2 \left(x_{ijklm} - x_{min}\right)^2}{\partial \mu} + \frac{\partial M_j}{\partial \mu} + \frac{\partial F_{jk}}{\partial \mu} + \frac{\partial g_{jkl}}{\partial \mu} + \frac{\partial e_{ijk}}{\partial \mu} = 1+0+\cdots \quad [2.44]$$

$$\frac{\partial y_{ijk}}{\partial S_{i'}} = \frac{\partial \mu}{\partial S_{i'}} + \frac{\partial S_i}{\partial S_{i'}} + \frac{\partial b_1 \left(x_{ijklm} - x_{min}\right)}{\partial S_{i'}} + \frac{\partial b_2 \left(x_{ijklm} - x_{min}\right)^2}{\partial S_{i'}} + \frac{\partial M_j}{\partial S_{i'}} + \frac{\partial F_{jk}}{\partial S_{i'}} + \frac{\partial g_{jkl}}{\partial S_{i'}} + \frac{\partial e_{ijk}}{\partial S_{i'}} = \quad [2.45]$$

$$= 0 + \left\{ \begin{array}{l} 1;\ i = i' \\ 0;\ i \neq i' \end{array} \right\} + 0 + 0 + 0 + 0 + 0 + 0 \quad [2.46]$$

$$\frac{\partial y_{ijk}}{\partial b_1} = \frac{\partial \mu}{\partial b_1} + \frac{\partial S_i}{\partial b_1} + \frac{\partial b_1 \left(x_{ijklm} - x_{min}\right)}{\partial b_1} + \frac{\partial b_2 \left(x_{ijklm} - x_{min}\right)^2}{\partial b_1} + \frac{\partial M_j}{\partial b_1} + \frac{\partial F_{jk}}{\partial b_1} + \frac{\partial g_{jkl}}{\partial b_1} + \frac{\partial e_{ijk}}{\partial b_1} = \quad [2.47]$$

$$= 0 + 0 + \left(x_{ijklm} - x_{min}\right) + 0 + 0 + 0 + 0 + 0 \quad [2.48]$$

$$\frac{\partial y_{ijk}}{\partial b_2} = \frac{\partial \mu}{\partial b_2} + \frac{\partial S_i}{\partial b_2} + \frac{\partial b_1 \left(x_{ijklm} - x_{min}\right)}{\partial b_2} + \frac{\partial b_2 \left(x_{ijklm} - x_{min}\right)^2}{\partial b_2} + \frac{\partial M_j}{\partial b_2} + \frac{\partial F_{jk}}{\partial b_2} + \frac{\partial g_{jkl}}{\partial b_2} + \frac{\partial e_{ijk}}{\partial b_2} = \quad [2.49]$$

$$= 0 + 0 + 0 + \left(x_{ijklm} - x_{min}\right)^2 + 0 + 0 + 0 + 0 \quad [2.50]$$

$$\frac{\partial y_{ijk}}{\partial M_{j'}} = \frac{\partial \mu}{\partial M_{j'}} + \frac{\partial S_i}{\partial M_{j'}} + \frac{\partial b_1 \left(x_{ijklm} - x_{min}\right)}{\partial M_{j'}} + \frac{\partial b_2 \left(x_{ijklm} - x_{min}\right)^2}{\partial M_{j'}} + \frac{\partial M_j}{\partial M_{j'}} + \frac{\partial F_{jk}}{\partial M_{j'}} + \frac{\partial g_{jkl}}{\partial M_{j'}} + \frac{\partial e_{ijk}}{\partial M_{j'}} = \quad [2.51]$$

$$= 0 + 0 + 0 + 0 + \left\{ \begin{array}{l} 1;\ j = j' \\ 0;\ j \neq j' \end{array} \right\} + 0 + 0 + 0 \quad [2.52]$$

$$\frac{\partial y_{ijk}}{\partial F_{j'k'}} = \frac{\partial \mu}{\partial F_{j'k'}} + \frac{\partial S_i}{\partial F_{j'k'}} + \frac{\partial b_1 \left(x_{ijklm} - x_{min}\right)}{\partial F_{j'k'}} + \frac{\partial b_2 \left(x_{ijklm} - x_{min}\right)^2}{\partial F_{j'k'}} + \frac{\partial M_j}{\partial F_{j'k'}} + \frac{\partial F_{jk}}{\partial F_{j'k'}} + \frac{\partial g_{jkl}}{\partial F_{j'k'}} + \frac{\partial e_{ijk}}{\partial F_{j'k'}} = [2.53]$$

$$= 0 + 0 + 0 + 0 + 0 + \left\{ \begin{array}{l} 1;\ j = j' \wedge k = k' \\ 0;\ j \neq j' \vee k \neq k' \end{array} \right\} + 0 + 0 \qquad [2.54]$$

$$\frac{\partial y_{ijk}}{\partial g_{j'k'l'}} = \frac{\partial \mu}{\partial g_{j'k'l'}} + \frac{\partial S_i}{\partial g_{j'k'l'}} + \frac{\partial b_1 \left(x_{ijklm} - x_{min}\right)}{\partial g_{j'k'l'}} + \frac{\partial b_2 \left(x_{ijklm} - x_{min}\right)^2}{\partial g_{j'k'l'}} + \frac{\partial M_j}{\partial g_{j'k'l'}} + \frac{\partial F_{jk}}{\partial g_{j'k'l'}} + \frac{\partial g_{jkl}}{\partial g_{j'k'l'}} + \frac{\partial e_{ijk}}{\partial g_{j'k'l'}} = [2.55]$$

$$= 0 + 0 + 0 + 0 + 0 + 0 + \left\{ \begin{array}{l} 1;\ j = j' \wedge k = k' \wedge l = l' \\ 0;\ j \neq j' \vee k \neq k' \vee l \neq l' \end{array} \right\} + 0 \qquad [2.56]$$

V nobenem od parcialnih odvodov ni nobenega parametra. Model je linearen.

V nadaljevanju poskusimo nekoliko spremenjene modele. Modeli so izpeljani iz modela 2.43. Spremembe so nerejene namenoma za ilustracijo pogojno linearnih modelov. Za analizo odbire pri mladicah, za katerega smo razvili izvorni model, niso priporočljivi.

$$y_{ijklm} = \mu + S_i + b_1 \left(x_{ijklm} - x_{min}\right) + b_2^2 \left(x_{ijklm} - x_{min}\right)^2 + M_j + F_{jk} + g_{jkl} + e_{ijklm} \qquad [2.57]$$

$$y_{ijklm} = \mu + S_i + b_1 sin \left(x_{ijklm} - x_{min}\right) + M_j + F_{jk} + g_{jkl} + e_{ijklm} \qquad [2.58]$$

$$y_{ijklm} = \mu + S_i + sin\left(b_1\right) \left(x_{ijklm} - x_{min}\right) + M_j + F_{jk} + g_{jkl} + e_{ijklm} \qquad [2.59]$$

Vaje:

1. Določi, ali je model linearni, pogojno-linearni oziroma nelinearni model!

$y_{ijk} = \mu + A_i + B_{ij} + e_{ijk}$

$y_{ijk} = \mu + K_i + S_j + KS_{ij} + e_{ijk}$

$y_{ijk} = \mu + F_i + B_j + b_i \left(x_{ijk} - 100\right) + e_{ijk}$

$y_{ijk} = \mu + F_i + B_j + b_{|i} \left(x_{ijk} - 100\right) + b_{||i} \left(x_{ijk} - 100\right)^2 + e_{ijk}$

$y_{ijk} = \mu + F_i + B_j + b_i\, sin\left(x_{ijk}\right) + e_{ijk}$

$y_{ijk} = \mu + F_i + B_j + sin^b \left(x_{ijk}\right) + e_{ijk}$

$y_{ijk} = \mu + F_i + B_j + b_i\, exp\left(c\left(x_{ijk} - 100\right)\right) + e_{ijk}$

2.3 Klasifikacija modelov

2.3.1 Z ozirom na število lastnosti

Enolastnostni modeli

$$y_{ijk} = \mu + P_i + b\left(x_{ijk} - \bar{x}\right) + e_{ijk} \qquad [2.60]$$

Opazujemo in merimo samo eno odvisno spremenljivko. Lastnosti merjenih v poskusu je lahko tudi več, vendar jih obravnavamo kot nekorelirane, nepovezane. Dobimo povsem iste izračune, če jih analiziramo ločeno ali pa skupaj, istočasno.

Večlastnostni modeli

Opazujemo več lastnosti hkrati. Lastnosti so med seboj korelirane (podobne ali nasprotne). Iz ene lastnosti sklepamo na drugo lastnost in ta dodatni vir informacij tudi koristimo. Lastnosti so lahko povsem različno porazdeljene (različne porazdelitvene funkcije), lahko se razlikujejo v modelih, lahko so merjene v različnih okoljih. Podatki lahko celo manjkajo, ki lahko izpadejo naključno (npr. ko odstranimo osamalce) ali pa načrtno, sistematsko (selekcija).

Pri večlastnostnih modelih moramo opozoriti na lastnosti s popolno ali skoraj popolno podobnostjo. Korelacija med njima je 1 ali blizu 1 oziroma -1 ali blizu 1.

2.3.2 Z ozirom na rang sistema enačb

Modeli s polnim rangom

Vsi parametri so ocenljivi. Imajo eno samo rešitev.

Regresijski modeli

Regresijski modeli (enostavna, polinomska regresija, multipla regresija - regresija z več kot eno pojasnjevalno spremenljivko). Z regresijo proučujemo odnose med merjenimi spremenljivkami. Običajno so to modeli s polnim rangom. V primerih, ko je eden od regresijskih koeficientov enak nič (visoka korelacija med neodvisnimi spremenljivkami, nepotreben element v polinomu), je lahko model tudi z nepolnim rangom. Toda v takih primerih poenostavimo - izpustimo parameter, ki je enak nič in ponovimo analizo.

$opazovanalastnost = f(pojasnjevalnespremenljivke)$

$[response = f(predictors)]$

Linearna regresija je poseben razred povezanosti, opisana z ravno črto - regresijsko premico [2.61], kadar je v modelu ena neodvisna spremenljivka oziroma z ravninami pri regresijah z več neodvisnimi spremenljivkami [2.62]. Modeli tega tipa se lahko uporabljanjo v več namenov, zelo razširjeni pa sta opisovanje odnosov med spremenljivkami in napovedovanje vrednosti. Odvisno spremenljivko lahko pojmujemo kot funkcijo lastnosti, določenih (konstantnih) pri poskusu: npr., debelina slanine se povečuje s težo.

PRIMER: Linearna regresija

$$y_i = \beta_0 + \beta x_i + e_i \quad [2.61]$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i \quad [2.62]$$

Pri ***polinomski regresiji*** za povezavo neodvisne in odvisne spremenljivke uporabljamo polinom druge ali višje stopnje [2.63]. Polinom druge stopnje ponazorjujemo s parabolo, pri več neodvisnih spremenljivkah, če je pri vseh spremenljivkah primerna kvadratna regresija, pa paraboloid. Toda v splošnem so povezave različne [2.64]. Povezavo s prvo neodvisno spremenljivko dobro opisuje parabola, za drugo je zadostna linearna regresija, pri tretji pa moramo prirediti polinom tretje stopnje. V primeru z več spremenljivkami lahko nastopajo tudi produkti [2.65].

PRIMER: Polinomska regresija

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \quad [2.63]$$

$$y_i = \beta_0 + \beta_{11} x_{1i} + \beta_{12} x_{1i2} + \beta_{13} x_{1i3} + \beta_{21} x_{2i} + \beta_{22} x_{2i2} + e_i \quad [2.64]$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i \quad [2.65]$$

Model z visokimi potencami je neprimeren:

1. razlaga (strokovna)
2. statistično (zakon skromnosti)
3. matematično

$$X^n \cong \alpha + \left(\frac{\partial x^n}{\partial x}\right)_{x=\alpha}(x-\alpha) + \frac{1}{2}\left(\frac{\partial^2 x^n}{(\partial x)^2}\right)_{x=\alpha}(x-\alpha)^2 + \cdots =$$

$$\cong \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \ldots + \gamma_n x^k$$

Torej x^n se lahko izrazi kot linearna funkcija nižjih polinomov. Težja je aproksimacija nižje potence (npr. x^4), vendar ni nobenega problema izraziti x^{15}.

Modeli z nepolnim rangom

Sistemi z nepolnim rangom imajo več parametrov kot linearno neodvisnih enačb. Zato nimajo nobene rešitve ali pa jih imajo neskončno število. Ocenljivi parametri so dejansko le funkcije odvečnih parametrov. Šele, ko se omejimo na določene vrednosti za neznane parametre, lahko enolično določimo ostale ("ocenljive"). To odločitev imenujemo **restrikcija**.

PRIMER: Sistem dveh enačb s tremi neznankami (parametri x, y in z). Sistem lahko delno rešimo. Tako lahko dva parametra (spremenljivki) (npr. x in y) izrazimo kot funkcijo tretjega (z v našem primeru).
$(x, y) = f(z)$

Tretja spremenljivka lahko zavzame katerokoli vrednost znotraj prostora parametrov (slanina 0-30). Ko poznamo njeno vrednost, lahko izračunamo ostale. Sistem enačb ima neskončno število rešitev. Ne moremo torej pričakovati ene same rešitve. Rezultate bomo preverjali tako, da bomo ocenili povprečje razreda.

PRIMER: Če npr. določimo z-ju vrednost nič, dobita parametra x in y vrednosti 22 in -6. Omejili smo se na eno rešitev.

$x + 2y + z = 10$ $\quad x = 10 - 2y - z = 22 - 5z \qquad x = f_x(z)$

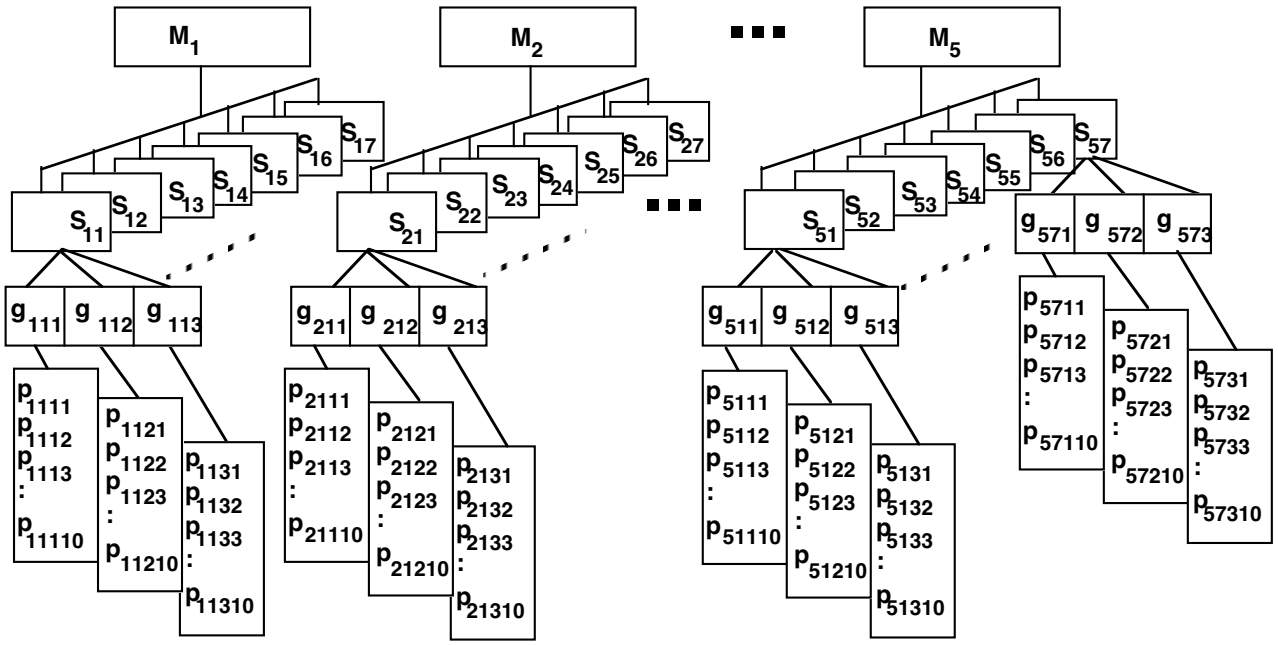
$x + 3y - z = 4$ $\quad y = -6 + 2z \qquad y = f_y(z)$

Isti rezultat bi dobili, če bi že na začetku za parameter z uporabili "izbrano vrednost". Sistem bi se poenostavil na sistem dveh enačb (dve polni celici) z dvema neznankama. Dobili bi torej reducirani sistem s polnim rangom in tako z eno samo rešitvijo, če le-ta obstaja.

Rang sistema in linearno odvisne enačbe ni vedno enostavno določiti. Odvisen je tako od modela, kar je enostavno določiti, kot strukture podatkov, pri čemer pa so odvisne enačbe lahko zelo zakrite. Linearno odvisne enačbe, ki izvirajo iz modela, lahko vedno odpravimo z **reparametrizacijo**. Težje delo pa imamo z enačbami, ki so linearno odvisne zaradi strukture podatkov. Pomaga le temeljita analiza strukture podatkov. Vedno lahko določimo število ocenljivih parametrov, saj lahko ocenimo vedno samo srednje vrednosti posameznih najnižje vgnezdenih zasedenih celic.

Dve šoli

- reparametrizacija-Henderson, Gianola
- splošna inverza-Rao, SEARLE (Harvey)



Slika 2.7: Hierarhična zgradba poskusa in vgnezdeni vplivi

2.3.3 Z ozirom na strukturo podatkov

2.3.3.1 Hierarhični modeli

V hierarhičnem modelu imamo samo vgnezdene vplive.

Vzemimo pujske (a) iz enega gnezda! Vsi pujski pripadajo gnezdu (g), gnezdo pa samo eni svinji, svoji materi (S). Ne morejo imeti dveh mater hkrati. Svinja je praviloma za vsako gnezdo, zlasti pa v poskusnih razmerah, parjena samo z enim merjascem, medtem ko pri osemenjevanju lahko z enim merjascem pripustimo več svinj hkrati. Svinje so torej vgnezdene znotraj merjasca (M). Poskusimo določiti. Začnemo z vplivom, ki je najbolj na vrhu, merjascem torej. Ker je to v modelu prvi vpliv, mo dobil indeks i. Oznaka vpliva merjasca je torej M_i. Naslednji vpliv je svinja, ki je vgnezdena znotraj merjasca, zato bo poleg indeksa merjasca, s katerim je bila pripuščena, nosila tudi svoj indeks. Označena bo torej S_{ij}. Z indeksom j označujemo vse svinje, ki smo jih pripustili oziroma osemenili z enim merjascem. Prav gotovo sta bili najmanj dve svinji pripuščeni z vsaj enim merjascem. Naslednji vpliv je gnezdo, ki pripada svinji. Tako bo dobil indekse od svinje, dodatni indeks k pa bo štel gnezda znotraj svinje. Oznaka za naključni vpliv skupnega okolja v gnezdu je g_{ijk}. Gnezdo je vgnezdeno tudi znotraj merjasca, ker je naveden tudi indeks i za merjasca. Vpliv pujska a je vgnezdeni znotraj gnezda, ker pujsek pripada enemu gnezdu in tako dobi vse indekse od gnezda. V poskusu je iz enega gnezda 10 pujskov, zato dobi vsak pujsek še dodaten indeks l. Pujska bomo pravilno opisali kot a_{ijkl}. Kot vidimo pujsek pripada merjascu i in svinji ij.

PRIMER 1: Vzemimo, da imamo pet merjascev, daje vsak merjasec parjen s sedmimi svinjami. Vsaka svinja ima po tri gnezda, iz vsakega gnezda je vzetih 10 pujskov. Vsakega pujska smo stehtali 8 krat. Narišimo strukturo podatkov in označimo merjasca in svinjo kot sistematska vpliva, gnezdo in pujska pa kot naključna vpliva. Rezultat je nakazan na sliki 2.7. Napišimo tudi osnovni in možni model! Preštejte število opazovanj! Določite število parametrov in stopinje prostosti!
PRIMER 2: Privzemimo informacije iz primera 1 in dodajmo, da je pol pujskov v vsakem gnezdu svinjk in druga polovica kastratov. Ponovite vajo!
PRIMER 3: Vzemimo drugi primer. V njem bomo spremenili le to, da so svinje praviloma parjenje z različnimi merjasci za posamezna gnezda. Ponovite vajo!

Tabela 2.19: Struktura podatkov pri hierarhičnem modelu

Vpliv	Pasma 1	Pasma 2	Pasma 3
Krma 1	n_{11}		
Krma 2	n_{21}		
Krma 3		n_{32}	
Krma 4		n_{42}	
Krma 5			n_{53}
Krma 6			n_{63}

Tabela 2.20: Struktura podatkov pri križno klasificiranem modelu

Vpliv	Pasma 1	Pasma 2	Pasma 3
Krma 1	n_{11}	n_{12}	n_{13}
Krma 2	n_{21}	n_{22}	n_{23}
Krma 3	n_{31}	n_{32}	n_{33}
Krma 4	n_{41}	n_{42}	n_{43}
Krma 5	n_{51}	n_{52}	n_{53}
Krma 6	n_{61}	n_{62}	n_{63}

Križno-klasificirani modeli

Pri križno klasificiranem modelu so vsi vplivi križno klasificirani. Pri dobrem poskusu imamo meritve pri vseh možnih kombinacijah glavnih vplivov, ki jo iz praktičnih razlogov poimenujemo celice. Število opazovanj je lahko različno, v vsaki celici pa mora biti najmanj eno opazovanje. To eno opazovanje ni dobro za kakovostno izveden poskus, a pomaha premostiti računske probleme. Pri poskusu moramo dobro vedeti, da so zaključki vredni le toliko kot najmanj točna informacija.

V preglednici 2.20 je prikazan enostaven križno klasificirani model z dvema vplivoma: pasmo in krmo. Ker ima pasma 3 nivoje, krma pa 6, je možnih 18 kombinacij. Pri vseh kombinacijah (celicah) imamo opazovanja, zato lahko rečemo, da je poskus dobro zasnovan. Vprašanje je le, zakaj smo hkrati preizkušali toliko krm.

V drugem poskusu (pregl. 2.21) sta vpliva tudi križno klasificirana, a načrt je nepopoln. Nobena od pasem ni dobila vseh krm. Vzroki so lahko različnji, med njimi bi lahko bili celo opravičljivi. Produktivne pasme ne moremo krmiti s skromnimi obroki, ki so dobri za stare, avtohtone pasme. Pri slednjih pasmah pa je obrok, naravnan na potrebe moderne pasme, preobilen in bi bil ekonomsko neupravičen, lahko pa bi povzročil tudi prehitro rast in zamastitev, kar je neugodno zlasti pri plemenski vzreji.

Kombinirani modeli

Modeli pogosto vsebujejo tako križno klasificirane vplive kot vgnezdene vplive, kar velja predvsem za podatke, ki jih beležimo v proizvodnih razmerah. Strukturo podatkov moramo poznati ne glede na to, da

Tabela 2.21: Struktura podatkov pri križno-klasificiranem modelu

Vpliv	Pasma 1	Pasma 2	Pasma 3
Krma 1	n_{11}	n_{12}	n_{13}
Krma 2	n_{21}	0	n_{23}
Krma 3	0	n_{32}	n_{33}
Krma 4	n_{41}	n_{42}	n_{43}
Krma 5	n_{51}	n_{52}	0
Krma 6	n_{61}	n_{62}	0

bi za samo obdelavo s statističnimi paketi to niti ne bilo potrebno. Dobro poznavanje podatkov pripomore k dobri interpretaciji.

2.3.4 Z ozirom na naravo parametrov (prisotnost vplivov)

Modeli s sistematskimi vplivi (fixed model)

Model ima samo eno slučajno spremenljivko - ostanek. Ostali vplivi so sistematski. Modele pogosto srečamo pri analizi podatkov načrtovanih poskusov, pri katerih smo se izognili koreliranim slučajnim vplivom.

$$y_{ijk} = \mu + P_i + b\left(x_{ijk} - \bar{x}\right) + e_{ijk}$$

Modeli z naključnimi vplivi (random model)

Model ima samo en sistematski vpliv - srednjo vrednost (mean). Vsi ostali vplivi so naključni. Model se je pogosto uporabljal pri nekaterih metodah analize variance (Henderson I). Podatki so se predhodno očistili sistematskih vplivov - korigirali. Sedaj pa modelov praktično ne srečamo.

$$y_{ijk} = \mu + u_i + v_j + (uv)_{ij} + e_{ijk}$$

Mešani modeli (mixed model)

Pri mešanih modelih imamo prisotne tako sistematske kot naključne vplive. Na modele naletimo v živinoreji, ko obdelujemo podatke iz proizvodnje. Pri analizi fenotipske variance so praktično neizogibni, zlasti če obdelujemo podatke v populaciji podvrženi selekciji. Neizogibni so pri napovedovanju plemenskih vrednosti. Poleg vpliva živali (aditivni direktni genetski vpliv) bodo v mešanih modelih lahko prisotni še naslednji naključni vplivi: aditivni maternalni genetski vpliv, dominanca, skupno okolje v gnezdu ali v čredi, permanentno okolje. Med sistematskimi vplivi pogosto srečamo v teh modelih sezono, starost ali maso, pasmo oz. genotip, genetske skupine...

$$y_{ijk} = \mu + P_i + b\left(x_{ijk} - \bar{x}\right) + a_{ij} + e_{ijk}$$

2.3.5 Z ozirom na število vplivov

Enorazsežni model

PRIMER:

$y_{ij} = \mu + \alpha_i + e_{ij}$ $\quad i = 1, ...p.$ $\quad j = 1, ...n_i$

Število parametrov:	$\mu,\ \alpha_1,\ \alpha_2,\ \ldots,\ \alpha_p$	$= 1 + p$
Število zasedenih celic:		$s \leq p$

Dvorazsežni model brez interakcije

PRIMER:

$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$ $\quad i = 1, ...p;$ $\quad j = 1, ...q;$ $\quad k = 1, ...n_{ij}$

Število parametrov:	$\mu,\ \alpha_1,\ \alpha_2,\ \ldots \alpha_p, \beta_1, \beta_2,\ \ldots,\ \beta_q,\ \ldots,$	$= 1 + p + q$
Število zasedenih celic:		$s \leq p + q$

Dvorazsežni model z interakcijo

PRIMER:

$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk} \qquad i = 1, \ldots, p; \qquad j = 1, \ldots, q; \qquad k = 1, \ldots, n_{ij}$

Število parametrov: $\mu, \alpha_1, \alpha_2, \ldots, \alpha_p, \beta_1, \beta_2, \ldots, \beta_q, \alpha\beta_{11}, \alpha\beta_{1q}, \alpha\beta_{21}, \alpha\beta_{2q}, \alpha\beta_{p1}, \alpha\beta_{pq}$

$= 1 + p + q + pq$

Število zasedenih celic: $s \leq pq$

Ocenimo lahko največ toliko parametrov, kot je zasedenih celic.

PRIMER:

farma	Pasma B_1	Pasma B_2	Pasma B_3
A_1	y_{111}, y_{112}	y_{121}, y_{122}	y_{131}
A_2	y_{211}, y_{212}	y_{221}, y_{222}	

- model: $y_{ijk} = \mu + A_i + B_j + AB_{ij} + e_{ijk}$

μ

$A_1 A_2$

$B_1 B_2 B_3$

$AB_{11} AB_{12} AB_{13} AB_{21} AB_{22}$ $\mathbf{AB_{23}}$

V tem primeru imamo samo 9 opazovanj, radi pa bi ocenili kar 12 parametrov in pet neodvisnih enačb-stopinj prostosti za model (za μ 1, za A 1, za B 2, za interakcije AB pa 1, skupaj 5 stopinj prostosti). Nemogoče je oceniti vse parametre zaradi manjkajočega razreda (AB_{23}). Število neodvisnih funkcij je enako številu polnih celic.

Če so prisotne manjkajoče celice, je potrebno pazljivo uporabljati statistične pakete (SAS). "Rezultati" so lahko odvisni od vrstnega reda vplivov v modelu, celo vrstnega reda nivojev. To pa je seveda povsem nezaželeno, saj bi lahko eden lahko dobil značilne razlike, drugi pa bi model ali podatke malo pomešal in dobil povsem drugačne zaključke. Odgovoriti si moramo na vprašanje, kaj je ocenljivo. Pri ocenljivih parametrih pa bodo rezultati vedno enaki.

Vedno je ocenljivo poprečje celic, saj so modeli z najmanjšimi celicami polnega ranga. Tako lahko model napišemo:

$$y_{ij} = \mu_i + e_{ij}$$

$$y_{ijk} = AB_{ij} + e_{ijk}$$

Kompleksni modeli

Modeli so redkokrat tako preprosti - samo v skrbno načrtovanih poskusih. Praviloma pa imamo v živinoreji opravka z več vplivi in nam enostavni modeli služijo kot učni primeri. Poleg tega pa smo se zgoraj naučili, da sta lahko enorazsežni in dvorazsežni model z interakcijo ekvivalentna - omogočata povsem iste zaključke. Delitev je bila v navadi pri statistikih, ki so uporabljali skalarno algebro in ročno računanje. Pri uporabi matrik in statističnih paketov v statistiki pa ta delitev izgubi na pomenu.

2.4 Pogojno linearni modeli

Pri pogojno linearnih modelih lahko osnovni model transformiramo tako, da prvi odvodi odvisne spremenljivke z ozirom na parametre ne vsebujejo parametrov (neznank).

Inverzni model

$$y_i = (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i)^{-1} \quad [2.66]$$

Je model linearen?

$$\frac{\partial y_i}{\partial \beta_0} = \frac{\partial \left[(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i)^{-1}\right]}{\partial \beta_0} \quad [2.67]$$

NE, še vedno vsebuje parametre.

Poskusimo transformirati model:

$$y_i^* = \frac{1}{y_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad [2.68]$$

$$\frac{\partial y_i^*}{\partial \beta_0} = 1; \qquad \frac{\partial y_i^*}{\partial \beta_1} = x_{1i}; \qquad \frac{\partial y_i^*}{\partial \beta_2} = x_{2i} \quad [2.69]$$

V transformiranemu modelu smo pri vseh prvih parcialnih odvodih izgubili parametre. Torej, model je pogojno linearen. Transformirane podatke bomo lahko obdelali po običajnih metodah. Pri interpretaciji pa moramo paziti: rezultate ne dobimo na normalni skali.

PRIMER: Logit - model

$$y_i = \frac{1}{1 + exp\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i\}} \quad [2.70]$$

Model je nelinearen (e_i predstavlja napako v eksponentu). Poskusimo poiskati transformacijo.

$$\frac{1}{y_i} - 1 = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i}$$

$$\underbrace{ln\left(\frac{1}{y_i} - 1\right)}_{logit} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad [2.71]$$

Zgornji model je linearen po transformaciji, torej je pogojno linearen.

Če je napaka izven eksponenta, pri transformaciji dobimo njen logaritem.

$$y_i = \frac{1}{1 + exp\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\} + e_i} \quad [2.72]$$

$$\underbrace{ln\left(\frac{1 - y_i}{y_i}\right)}_{logit} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ln\, e_i \quad [2.73]$$

$$\underbrace{ln\left(\frac{1 - y_i}{y_i}\right)}_{logit} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \underbrace{ln\, e_i}_{novi\ ostanek} \quad [2.74]$$

Pri modelih je zelo dobrodošla lastnost, da je ostanek normalno porazdeljen. Slednji model bi bil dobrodošel v primerih, ko se variabilnost povečuje s povprečjem.

PRIMER: Količina mleka in dnevni prirast:

$\bar{x}$	300	500	900
σ	30	50	100

Če obravnavamo vse tri priraste kot isto lastnost (prirast), jih bomo transformirali.

Log-transformirani modeli

$$ln(y_i) = \alpha + \gamma ln(x_i) + e_i \qquad i = 1, \ldots, n \tag{2.75}$$

$$\frac{\partial(y_i)}{\partial\alpha} = 1 \qquad i = 1, \ldots, N \tag{2.76}$$

$$\frac{\partial(y_i)}{\partial\gamma} = ln(x_i) \qquad i = 1, \ldots, N \tag{2.77}$$

Logistični (logit) model

2.5 (Pogojno) nelinearni modeli

Prvi odvodi vsebujejo parametre in hkrati ni transformacije, ki bi model linearizirala.

2.5.1 Rastna krivulja

Model:

$$y_i = \underbrace{A\left(1 - Be^{-kt_i}\right)^{-1}}_{\eta(\boldsymbol{\beta})} + e_i \tag{2.78}$$

kjer pomeni:

y_i - opazovanje
t_i - čas, starost $\qquad i = 1, 2, ..., n$
B - masa ob rojstvu
A - odrasla velikost
k - parameter, ki je povezan z ukrivljenostjo
e_i - ostanek
e - ... konstanta

Parametri v modelu: $(A, B, k) = \boldsymbol{\beta}$

$$\frac{\partial y_i}{\partial A} = \left(1 - Be^{-kt_i}\right)^{-1} \tag{2.79}$$

$$\frac{\partial y_i}{\partial B} = A\frac{e^{-kt_i}}{\left(1 - Be^{-kt_i}\right)^2} \tag{2.80}$$

$$\frac{\partial y_i}{\partial k} = AB\frac{e^{-kt_i}}{\left(1 - Be^{-kt_i}\right)^2} \tag{2.81}$$

Model je nelinearen, ker v prvih odvodih ostajajo parametri. Model ima nekatere parametre, ki so bolj linearni od drugih. Parameter A se pojavlja v odvodih redkeje in v bolj enostavnih izrazih kot drugi parametri. Tako je parameter A, ki ga je lažje oceniti. V odvodih pa sta ostala tudi druga dva parametra B in k.

Nelinearne modele rešujemo iterativno tudi v primeru, če uporabimo metode najmanjših kvadratov ali metodo največje zanesljivosti.

Pseudo-linearni model (linearna aproksimacija)

$$y_i = \eta(\beta_1, \beta_2, \beta_3, ...) + e_i$$

Pri teh modelih nelinearno enačbo nadomestimo z linearnim modelom. Model ni čisto pravi, vendar na opazovanem intervalu ne bomo dobili pomembnih odstopanj. Nelinearni krivulji se bomo približali s polinomom ali kakšno drugo funkcijo. Morda bomo pred tem preoblikovali, transformirali neodvisno spremenljivko. Pri tem pa je pomembno, da je model, s katerim poskušamo opisati pravo funkcijo, linearen.

Laktacijske krivulje

Obstaja tudi možnost sestavljanja različnih funkcij na različnih intervalih. Pri rastni krivulji bi na začetku rasti uporabili polinom druge stopnje s pozitivnim regresijskim koeficientom pri kvadratnem členu. V času največje rasti zadostuje linearna regresija. Ko pa se živali približujejo odrasli velikosti, pa hitrost rasti pojenja. Na tem intervalu je ponovno primernejši polinom druge stopnje, regresijski koeficient pri kvadratnem členu pa bo negativen.

2.6 Ekvivalentni modeli

Ekvivalentni modeli so različice modela, ki pa popolnoma enakovredno opisujejo podatke. Iščemo jih lahko samo med modeli z istim številom ocenljivih parametrov. Modele preoblikujemo zaradi interpretacije ali numerične stabilnosti sistema enačb. Zaradi računalniške nenatančnosti predstavitve realnih števil, pa tudi numeričnih napak pri računskih operacijah, ki so še nekoliko bolj verjetne pri neuravnoteženih poskusih, računalnik ne more zanesljivo proglasiti, katere enačbe so linearno odvisne od drugih.

Preoblikovanje modela imenujemo reparametrizacija. Kot ime samo pove poskusimo najti parametre, ki bodo parametre v starem modelu združili tako, da bomo odstranili čimveč "neocenljivih" parametrov, ali razdružili tako da bomo model prilagodili interpretaciji.

Poglejmo spodnje modele in jih primerjajmo. Vzemimo, da je prvi model (2.82) pravilen. Modelom določimo najprej število zahtevanih (željenih) parametrov, število ocenljivih parametrov (stopinj prostosti), potem pa poiščimo linearne kombinacije parametrov v manjšem modelu za izpuščene parametre iz prvega modela.

$$y_{ijk} = \mu + A_i + B_j + AB_{ij} + e_{ijk} \quad [2.82]$$

$$y_{ijk} = \mu + A_i + AB_{ij} + e_{ijk} \quad [2.83]$$

$$y_{ijk} = \mu + A_i + B_{ij} + e_{ijk} \quad [2.84]$$

$$y_{ijk} = \mu + AB_{ij} + e_{ijk} \quad [2.85]$$

$$y_{ijk} = \mu + AB_{ij} + e_{ijk} \quad [2.86]$$

$$y_{ijk} = \mu + A_i + B_j + e_{ijk} \quad [2.87]$$

Rezultat: samo zadnji model (2.87) ni ekvivalenten prvemu. Drugi model (2.83) je nekoliko "nerodno" napisan, lahko pa bomo izpeljali povsem iste ocenljive funkcije. Povsem jasno je, da so interakcije vgnezdene znotraj vpliva A, so pa prav tako vgnezdene znotraj vpliva B. Zakaj smo ga torej izpustili? Bolj sprejemljiv je tretji model (2.84). Statistično gledano (numerično) sta modela 2.82 in 2.84 ekvivalentna, različna pa je interpretacija (vsebina). Četrti (2.85) in peti (2.86) pa se ujemata s prvim tudi v interpretaciji, znebimo pa se linearno odvisnih vrstic.