

Poglavje 1

OCENJEVANJE IN INTERPRETACIJA PARAMETROV

V tem poglavju se bomo seznanili s statističnimi metodami za oceno lokacijskih parametrov in kriteriji oziroma pogoji, ključnimi za izbiro metode. Izbor metode je v največji meri odvisen od strukture podatkov in porazdelitve lastnosti.

Za oceno parametrov so nam na voljo številne metode. Izberemo lahko enostavna povprečja, metodo najmanjših kvadratov, tehtanih najmanjših kvadratov, splošnih najmanjših kvadratov, metodo največje zanesljivosti itd.

1.1 Kriteriji za izbor metode

1.1.1 Struktura podatkov

Uravnoteženi in neuravnoteženi poskusi

Pri uravnoteženih poskusih, kamor lahko štejemo tudi skrbno načrtovane poskuse po posebnih shemah (split-plot, latinski kvadrat...) lahko obdelamo z enostavno metodo, ki je znana pod imenom analiza variance z oznako ANOVA. Uravnoteženi poskusi so v živinoreji redki. V poskusih, ki jih izvajamo na kmetijah ali celo večjih obratih, smo vezani na velikost obratov in strukturo črede, ki jo imajo. Le redko imamo priložnost, da je hlev prazen in naselimo tiste živali, ki jih želimo

Vsekakor morajo biti ti poskusi izvedeni korektno. Ne smemo si zatiskati oči pred pomembnimi vplivi, ki jih nismo mogli izničiti ali kontrolirati. Sorodne živali, če jih imamo, morajo biti uravnoteženo porazdeljene po skupinah. Če imamo v poskusu s prašiči štiri skupine, bomo v poskus vključili nesorodne živali ali pa vzeli štiri prašiče iz istega gnezda in v vsako skupino dali po enega. Imamo sicer še nekaj možnosti, vseh niti ne moremo naštetih. Paziti moramo, da so "kršene" predpostavke uravnoteženo porazdeljene med skupinami. To pravilo je na videz v nasprotju z zahtevo po naključni porazdelitvi živali oziroma enot v poskus. Tisto, kar poskus moti, to pa so "nezaželeni" vplivi in "kršene" predpostavke, moramo načrtno izbrati in porazdeliti po skupinah. Le tako lahko zagotovimo, da ne okužimo tistih vplivov, ki jih moramo proučiti. To velja pravzaprav za vse poskuse, tudi za tiste, za katere že vnaprej vemo, da bomo uporabili zahtevnejšo metodo za obdelavo podatkov. Nobena metoda ne more nadoknaditi slabo zastavljenega poskusa!

Naključni in selekcionirani vzorci

Pri običajnih poskusih bomo živali ali kakšne druge poskusne enote praviloma naključno uvrstili v skupine. To pa vedno ni mogoče. Eden od razlogov je etične narave (primer 4.1). V živinoreji se pogosto srečamo z obilico proizvodnih podatkov zbranih v dokaj urejenih informacijskih sistemih. Ti podatki niso naključno izbrani: to so proizvodni rezultati živali na kmetijah, praviloma vseh.

Neumnost bi bilo iz teh podatkov izbrati naključni vzorec, da bi zadovoljili pogoje za obdelavo. Prav gotovo vsi podatki več povedo, kot manjši naključni vzorec. Imajo pa kmetje različno število živali, različnih genotipov, različno kakovost krme, različni intervali med molžama itd. Še najbolj problematično pa je, da potomce odbirajo od najboljših krav in odličnih bikov. Potomci torej niso naključni: ker smo v teorijo dedovanja prepričani, morajo biti tisti, ki so namenjeni prirerji in razmnoževanju, boljši.

PRIMER : Ko želimo proučiti neko lastnost pri zdravih in bolnih živalih, bi naključni izbor pomenil, da izberemo za poskus zdrave živali in jih okužimo. To bomo naredili le izjemoma, z dobro utemeljenim razlogom. Rezultati poskusa morajo biti dovolj tehtni, da za to žrtvujemo zdravje živali. Praviloma bomo postopali raje drugače. Poiskali bomo obolele živali in iz njih nastavili eno skupino, v drugo pa dali zdrave. Ali je tu kršeno pravilo o naključnih vzorcih? Pravzaprav ne! Zdravstveno stanje je vpliv in tega izberemo načrtno. V ozadju bolezni se lahko sicer skriva marsikaj, da so ene živali obolele, druge pa ne. Lahko je to izpostavljenost povzročiteljem bolezni, kar pravzaprav želimo proučevati, in to samo po sebi ne bi dalo izkrivljenih rezultatov. Lahko so bile živali izpostavljene dodatnim stresom, med njimi pa so lahko take, ki bi motile poskus. Vzemimo, da so bile obolele živali preslabo krmljene, zdrave pa ravno prav. V tem primeru bodo proizvodni rezultati morda bolj posledica slabe prehrane kot pa bolezni in poskus ne bi dal pravih rezultatov. Težav, torej tudi poskusa, se moramo lotiti na povsem drugem koncu. Bolezen je torej posledica, prav bi torej bilo, da bi bila objekt proučevanja, naša opazovana lastnost. Odpraviti bo potrebno vzroke, da je do bolezni sploh prišlo. Bolj bi bilo torej primerno, da proučujemo, kako prehrana vpliva na pojav bolezni...

1.1.2 Porazdelitev

Porazdelitev preverjamo za vse naključne vplive, ostanek in opazovanje. Najprej preverimo porazdelitev za opazovanja, ker pač ostankov še nimamo. Vsekakor pa pri opazovanjih, pri katerih imamo dvome, opravimo preizkuse na ostanku.

Porazdelitvena funkcija

Zaželjeni so normalno porazdeljeni podatki. Pri teh podatkih lahko pač uporabimo enostavne metode, kot so ANOVA, metoda najmanjših kvadratov, metoda tehtanih najmanjših kvadratov in metodo splošnih najmanjših kvadratov. Opišemo jih lahko s pričakovanimi vrednostmi ter merami razpršenosti (parametri disperzije). Pri drugih porazdelitvah pred povprečjem izberemo parametre, ki to porazdelitev najbolj opišejo. V živinoreji se boste dostikrat srečali tudi z lastnostmi, ki imajo specifično porazdelitev, biološko utemeljeno, a jih ne moremo opisati z znanimi porazdelitvami. Pri teh porazdelitvah pogosto odpadejo običajne statistike, povprečje in razpršenost prav nič ne povesta o podatkih.

Pri izboru metode upoštevamo porazdelitev opazovanj, ki je posledica porazdelitve naključnega dela modela. Pri sistematskih modelih, ki so zelo pogosti pri obdelavi poskusov, je porazdelitev opazovanj odvisna od ostanka.

Ko se izkaže, da lastnost ni porazdeljena po znani porazdelitvi, najprej poskusimo z različnimi transformacijami. Na transformiranih opazovanjih in ostankih preizkus ponovimo. Druga možnost je tudi približek prave porazdelitve z eno od znanih porazdelitev, kadar so odstopanja zanemarljiva. Pri tem statističnih parametrov za odločitev pravzaprav nimamo in so potrebne predvsem izkušnje in zadostno število opazovanj. Zanemarjanje porazdelitve je pogosto posledica nepoznavanja drugih metod, rezultati takih analiz pa so lahko zavajajoči.

S porazdelitvami, ki se ne dajo transformirati ali aproksimirati, se boste srečali že med študijem in pri diplomskih nalogah. Nimamo prostora, da bi imenovali vse primere in se bomo omejili le na nekatere. Vam so že poznani nekatere lastnosti obnašanja, dnevni ritmi in podobno. Pri rejah pa se boste spoznavali z nekaterimi lastnostmi plodnosti, ki se ne obnašajo, kot bi v statistiki radi. Omenimo lahko lastnosti povezane z vitalnostjo, preživitvijo ali izgubami mladičev, z dobami od poroda (odstavitve) do pripusta in z življenjsko prirejo. Za začetek lahko lastnosti opišemo enostavno s porazdelitvijo in porazdelitve primerjamo med seboj. Morda se lahko poslužimo mediane ali modusov, kombiniranih transformacij, a je zelo težko dati splošno veljaven postopek obdelave podatkov. Razgovor z nekom, ki ima bogate izkušnje pri obdelavi podatkov, bo gotovo primeren začetek.

Identična ali heterogena porazdelitev

V manjših poskusih velja poskusiti zagotoviti homogenost variance. To dosežemo tako, da imamo iste instrumente, stalno in usklajeno ekipo. Ko pa živali rastejo in jim spremljamo maso od rojstva pa jih ne moramo tehtati z isto tehtnico: tista za odrasla goveda, razlik med rojenimi teleti skoraj ne zazna. Tista, ki pa je primerna za tehtanje telet, pa bi se potrla pod maso odraslega goveda.

Neodvisna ali odvisna porazdelitev

V praksi dostikrat privzamemo, da so nivoji znotraj posameznega vpliva neodvisni. Do sedaj smo omejnili odvisnost med sorodniki pri aditivnem genetskem vplivu. Povezanost smo ugotavljali iz porekla živali in jo vgradili v matriko sorodstva. To velja tako za direktne, maternalne in paternalne aditivne genske vplive. Povezanost imamo tudi pri neaditivnih genetskih vplivih: dominanci in epistazi, vendar pa se v okviru tega predmeta z njimi ne bomo ukvarjali.

Samo omenili bomo tudi podobnost med primerjalnimi skupinami. Pri prašičih na testni postaji preizkušamo prašiče od 30 do 100 kg. Živali naseljujemo vsak teden in vsak teden zapuščajo testno postajo. Primerjalno skupino tvorijo živali, ki jih istočasno preizkušamo. Če smo natančni, to pomeni, da tvorijo primerjalno skupino živali, ki so v istem tednu končale preizkus. Teh pa je pogosto premalo, da bi dale zadostno oceno vpliva skupnega okolja, ki ga praviloma imenujemo kar sezona. Pomagamo si lahko z živalmi, ki so končale test en teden prej. Morda moramo zaradi števila živali dodati še kakšen teden...

Transformacije

S transformacijami lahko spremenimo porazdelitev spremenljivke tako, da je porazdeljena normalno ali po drugem znanem porazdelitvenem zakonu. Porazdelitev je pomembna samo pri odvisnih spremenljivkah. Pri neodvisnih spremenljivkah jo naredimo takrat, kadar z njeno transformacijo dosežemo enostavnejšo povezavo med odvisno in neodvisno spremenljivko, npr. iz eksponentne funkcije z logaritmiranjem dobimo linearno povezavo. Logaritmiranje pa ni edina transformacija. Poslužujemo se lahko tudi različnih korenov pri desno asimetričnih porazdelitvah ali potenciranja pri levo asimetričnih porazdelitvah. Pri transformacijah moramo paziti, da lahko pri transformaciji vse vrednosti spremenimo. Če je zaloga vrednosti pri spremenljivki večja ali enaka nič, transformacija z logaritmom ni mogoča, ker ne poznamo vrednosti $\log(0)$. Aproksimacija z eno od zelo majhnih vrednosti lahko rezultat močno preoblikuje. Ni vseeno ali se vrednosti 0 približamo z vrednostjo 10^{-2} ali 10^{-10} . Pri log-transformaciji dobimo v prvem primeru vrednost -2 , v drugem pa -10 .

Aproksimacije

Pri manjših odstopanjih od normalne porazdelitve lahko pravo porazdelitev pravzaprav zanemarimo in kot približek vzamemo normalno porazdelitev. Kot primer lahko navedemo velikost gnezda pri prašičih, ki je porazdeljena po Poissonovi porazdelitveni funkciji s povprečjem pri 10, minimumom 0 in maksimumom okrog 20. Tukaj lahko privzamemo normalno porazdelitev. To pa ne velja za velikost gnezda pri drobnici, kjer je porazdelitev tudi Poissonova, povprečje pa je praviloma med 1 in 2, v gnezdu je najmanj 0 živorojenih mladičev in praktično nikoli ne presega pa 5 mladičev.

1.1.3 Število opazovanj

V statistiki poznamo tudi zakon velikih števil. V preprostem jeziku pove, da ne glede na porazdelitev opazovanj se porazdelitev pričakovanih vrednosti približuje normalni porazdelitvi, če je le število opazovanj dovolj veliko. Ta zakon pravzaprav dovoljuje nadomestitev prave porazdelitve z normalno porazdelitvijo. Kaj je zadostno število, pa je povezano s porazdelitvijo in številom parametrov v sistematskem delu modela. Bolj kot porazdelitev odstopa od normalne, več opazovanj potrebujemo, da lahko predpostavimo kar normalno porazdelitev.

1.1.4 Načelo skromnosti, praktičnost, izvedljivost in interpretacija

V statistiki velja načelo skromnosti tudi pri izbiri metode. Uporabili bomo najpreprostejšo metodo, ki nam še vedno daje zadovoljivo oceno rezultatov poskusa. Če je to navadno povprečje, bomo izbrali pač povprečje.

1.1.5 Funkcija tveganja in funkcija izgube, loss function

Funkcija tveganja (risk function) ali kar tveganje in funkcija izgub ali izguba (loss function)

- vsota kvadratov za ostanek najmanjša
- vsota tehtanih kvadratov za ostanek najmanjša
- vsota splošnih kvadratov za ostanek najmanjša
- največja zanesljivost

1.2 Enostavna analiza variance (ANOVA)

V poskusih, kjer je struktura podatkov uravnotežena, lahko uporabljamo enostavno metodo, imenovano ANOVA. Uporabna je v primerih, ko smo opravili uravnotežen preizkus in nimamo na voljo statističnih paketov. Izračunamo in primerjamo lahko povprečja po skupinah. Za preizkuse postavljenih hipotez pa uporabimo enostavne izračune vsot kvadratov za posamezne vplive, kar bomo opisali v poglavju o preizkušanju hipotez. Analizo lahko opravimo kar s kalkulatorjem.

Toda v živinoreji prepogosto kršimo in to metodo uporabimo na neuravnoteženih podatkih in bolj sestavljenih modelih. Prav nič ne izgubimo, kadar uravnotežene podatke obdelamo z metodo najmanjših kvadratov. Kadar bi lahko uporabili metodo ANOVA, bomo pri metodi najmanjših kvadratov dobili popolnoma enake zaključke. Če pa se rezultati razlikujejo, pa je metoda najmanjših kvadratov boljša.

1.3 Metoda najmanjših kvadratov

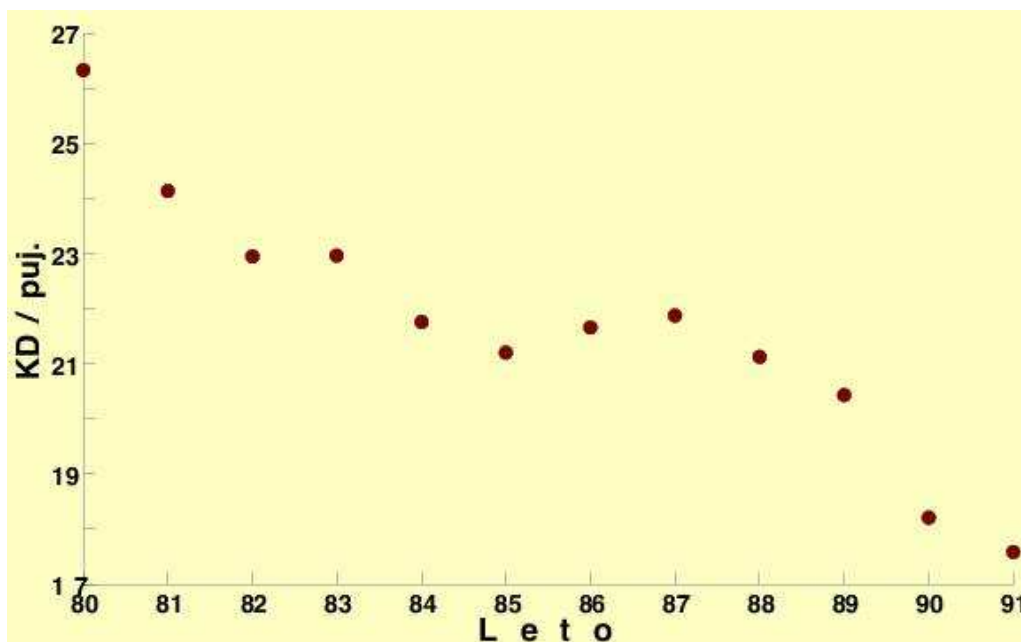
Pri metodi najmanjših kvadratov (ang. Ordinary Least Square, OLS) predpostavimo, da je odgovarjajoči model sistematski, ostanki pa so identično in neodvisno porazdeljeni. Število meritve po posameznih skupinah se lahko razlikuje. Torej so podatki lahko neuravnoteženi. Pri uporabi metode ni nujno zahtevati normalno porazdelitev, dokler se zadovoljimo samo z rešitvami. Ne smemo pa pozabiti, da se pri "divji" porazdelitvi ocenjena vrednost lahko pojavi celo izven intervala zalog vrednosti. Torej dobimo lahko vrednost, ki je nemogoča in neuporabna za interpretacijo.

1.3.1 Ilustracija metode najmanjših kvadratov

Najprej postopek ilustrirajmo s primerom. Za ilustracijo vzemimo podatke o spremembah krmnih dni na živorojenega pujska v Sloveniji po letih iz tabele 1.1 in narišimo sliko 1.1! Krmni dnevi predstavljajo lastno ceno živorojenega pujska v trdni "prašičerejski" valuti in so tako dober pokazatelj uspešnega rejskega dela. Tako iz tabele kot iz grafa dobro vidimo, da so slovenski rejci izboljšali rejo prašičev. Dosežen rezultat v posameznem letu je na grafu označen z rdečo kroglico.

Tabela 1.1: Število krmnih dni na živorojenega pujska po letih

80	81	82	83	84	85	86	87	88	89	90	91
26.33	24.14	22.95	22.97	21.76	21.21	21.66	21.88	21.13	20.44	18.21	17.59



Slika 1.1: KD na živorojenega pujska po letih

Za obrazložitev uporabimo kar preprost model, ki vsebuje samo linearno regresijo z neodvisno spremenljivko x_i kot sistematski vpliv. Verjetno bi polinom tretje stopnje bil primernejši, a bi si samo otežili izračune in prikaze. Lahko pa vajo s polinomom tretje stopnje naredite doma za vajo. V skalarni obliki bo torej model imel preprosto obliko, kot nakazuje enačba 1.1, v enačbi 1.2 pa imamo opisano pričakovano vrednost in strukturo variance. Za metodo najmanjših kvadratov opišeta tudi običajne predpostavke.

$$y_i = \mu + bx_i + e_i \quad [1.1]$$

$$y_i \sim IID(\mu + bx_i, \sigma^2) \quad [1.2]$$

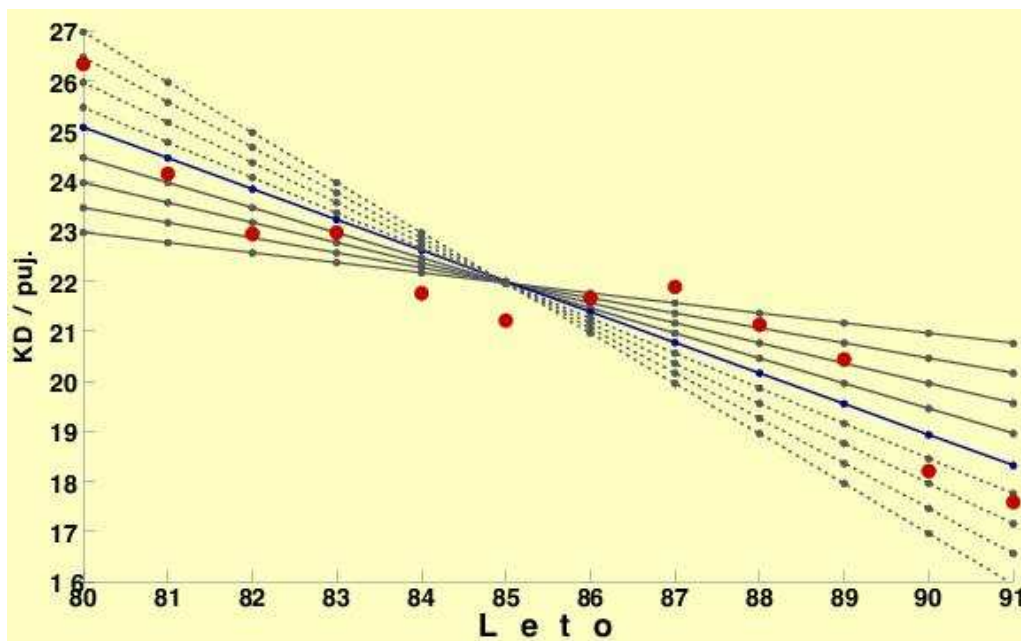
Tako je iz enačbe 1.2 razvidno, da so meritve y_i identično in neodvisno porazdeljene, na kar nas opozori oznaka *IID* (*ang.* Identically and Independently Distributed) in varianca σ^2 , s pričakovano vrednostjo $\mu + bx_i$.

Porazdelitev naključnih spremenljivk v modelu moramo nujno preveriti, ko preizkušamo značilnost postavljenih hipotez. Za preveritev hipotez bi lahko preprosto rekli, da preverjamo verodostojnost dobljenih rezultatov. Ta nadaljni korak je praviloma pričakovan in povsem logičen, saj brez statističnega preizkusa ne moremo rezultatov interpretirati. Analiza brez statističnega preizkusa bi bila torej brez pravega učinka. Za presojo rezultatov je porazdelitev meritev pomembna. Dokazali pa smo že, da je porazdelitev odvisna od naključnega dela modela - od naključnih spremenljivk. Kadar razmišljamo o varianci, predpostavimo, da je spremenljivka normalno porazdeljena ali pa je porazdeljena tako, da lahko pri obdelavi predpostavimo normalno porazdelitev. Takšna predpostavka nam pride prav, ker lahko uporabimo metode iz skupine najmanjših kvadratov. V živinorejskih poskusih tega ne smemo vedno narediti. Da bi ne naredili napake, moramo vedno preveriti porazdelitev!

Prilagodimo točkam na grafu 1.1 najprej 9 premic iz tabele 1.3. Tako dobimo graf 1.2 s šopom premic. Zanima nas, katera od njih podatke najbolj opiše ali pojasni. Z drugimi besedami: zanima nas, katera premica se podatkom najbolj prilega.

Tabela 1.2: Nekaj izbranih enačb za opis zgornjih podatkov

Oznaka	Nekatere izbrane enačbe	Oznaka	Nekatere izbrane enačbe
1	$y_i = 39.033 - 0.20084 x_i + e_i$	6	$y_i = 81.533 - 0.70084 x_i + e_i$
2	$y_i = 47.533 - 0.30084 x_i + e_i$	7	$y_i = 90.033 - 0.80084 x_i + e_i$
3	$y_i = 56.033 - 0.40084 x_i + e_i$	8	$y_i = 98.533 - 0.90084 x_i + e_i$
4	$y_i = 64.533 - 0.50084 x_i + e_i$	9	$y_i = 107.033 - 1.00084 x_i + e_i$
5	$y_i = 74.146 - 0.6135 x_i + e_i$		



Slika 1.2: KD na živorojenega pujska in prileganje različnih premic

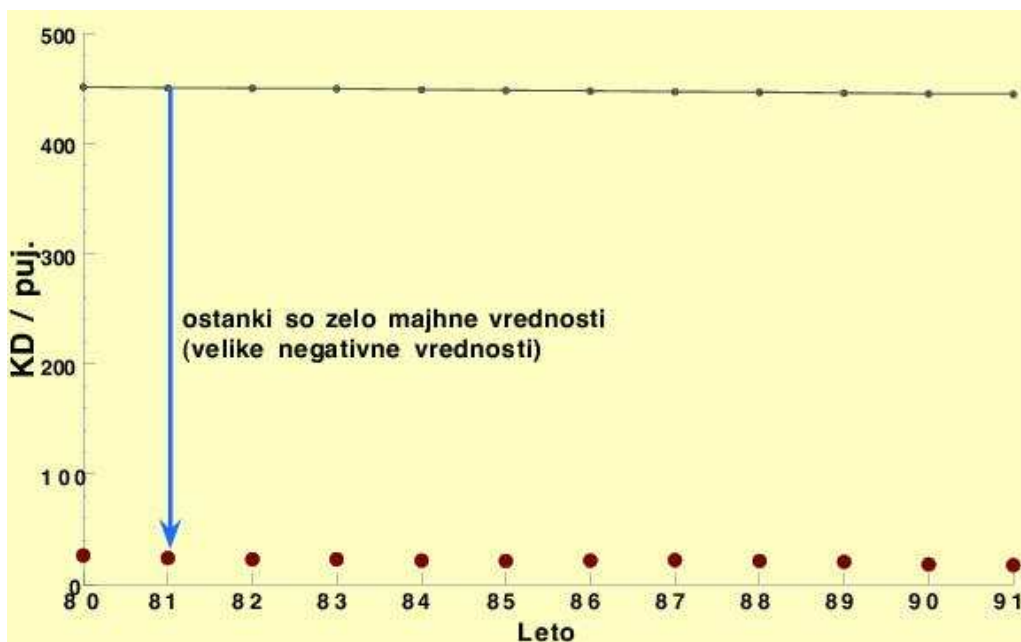
Odgovor je enostaven: najbolje se prilega srednja - modra premica. Izbrali smo jo zato, ker se točke modri premici najbolj prilegajo. "Prilegajo" pomeni, da so od nje najmanj oddaljene. Kakšna od točk lahko leži kar na najboljši premici, tu in tam pa imamo tudi precej oddaljene točke. Ker je več točk, moramo tako najti neko statistiko, ki bo merila skupno oziroma povprečno oddaljenost. Oddaljenost točk (meritev) od premice (pričakovane vrednosti) pa imenujemo ostanek. Ker pravih ostankov ne poznamo, ocenimo pa jih lahko kot razliko med izmerjeno vrednostjo y_i in pripadajočo pričakovano vrednostjo $E(y_i)$, kot to prikazuje enačba 1.3.

$$\hat{e}_i = y_i - E(y_i) \quad [1.3]$$

Pri razmišljanju nas dostikrat zapelje želja, da bi izbrali rešitev tako, da bi bila vsota ostankov enaka 0 (1.4) ali morda celo najmanjša (1.5). V zgornjem primeru (graf 1.2) je vsota ostankov pri vseh premicah enaka 0. Pravzaprav ta kriterij izpolnjuje cel šop premic, ki gredo skozi presečišče premic na grafu. Premic je celo neskončno mnogo. Torej je med njimi po tem kriteriju ni nobene najboljše - po tem kriteriju ne bomo našli dobre rešitve.

$$\sum_i (y_i - E(y_i)) = 0 \quad [1.4]$$

$$\sum_i (y_i - E(y_i)) = \min. \quad [1.5]$$



Slika 1.3: Vsota ostankov je minimalna

Pri drugem kriteriju (1.5), da bi bila vsota ostankov najmanjša, tudi ne bi bili uspešni. Najmanjše niso vrednosti blizu 0, ampak zelo “velike” negativne vrednosti. Po tem kriteriju bi poiskati tisto premico (graf 1.3), ki bi dala najbolj negativno vsoto. Premice prav gotovo ne bi imeli na grafu! Od točk - opazovanj - bi bila odmaknjena neskončno daleč. Do nje bi potovali več svetlobnih let, če vam je tako všeč. Rešitev, ki je daleč od opazovanj, bi o opazovanjih zelo malo povedala. Potem pa to sploh ni rešitev!

Poiskali bi lahko še kak kriterij in morda bi bila obrazložitev, zakaj z njim nismo najbolj zadovoljni, celo zahtevna. Poskusimo zato kar s kriterijem, na katerem je osnovana metoda najmanjših kvadratov. Uporabiti moramo odklone, da se bo premica najboljše prilegala. Ker je v nazivu metode beseda “kvadratov”, moramo odklone kvadrirati. Tako dobimo za vsako meritev eno vrednost - kvadrirani odklon. Če pa kvadrirane odklone seštejemo, pa imamo statistiko, ki pa nam da rešitev. Statistiko bomo imenovali vsota kvadratov ali vsota kvadratnih odklonov. Poglejmo, če je razmišljanje dobro!

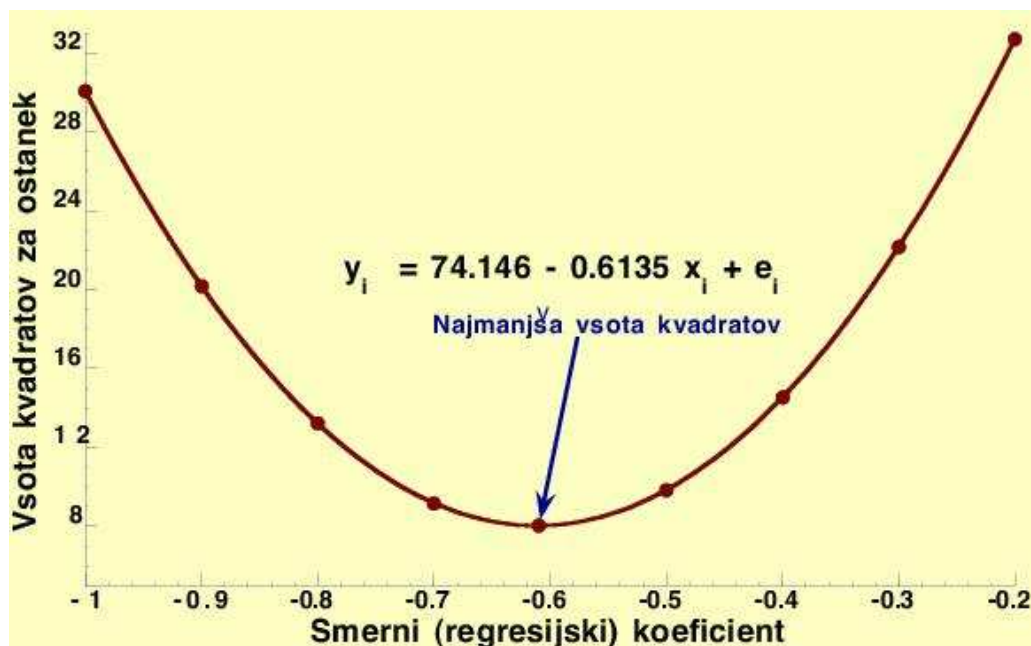
Za vsa leta in vse enačbe imamo kvadratne odklone shranjene v tabeli 1.3 in jih v zadnji vrsti še seštejemo. Tako smo dobili vsoto kvadratnih odklonov (1.6), uporabljali bomo pa kar oznako RSS , kar je povzeto po angleškem izrazu "Residual Sum of Squares". Katera premica je torej najboljša? Zaradi izbora metode prav gotova tista, pri kateri je vsota kvadratnih odklonov RSS najmanjša. To je v našem primeru enačba 5, ki tudi predstavlja rešitev sistema. To pravzaprav ni naključno: pred tem postopkom smo na skrivaj opravili izračun.

$$RSS = \sum_i (y_i - E(y_i))^2 \quad [1.6]$$

Tabela 1.3: Kvadrati za ostanke pri različnih prirejenih premicah (tabela, graf)

En.	1	2	3	4	5	6	7	8	9
b	-0.20	-0.30	-0.40	-0.50	-0.61	-0.70	-0.80	-0.90	-1.00
80	11.316	8.2025	5.5885			0.74650	0.13264	0.01844	0.40424
81	1.8907	0.95070	0.33067			0.39057	1.0505	2.0305	3.3305
82	0.14890	0.00737	0.04585	0.26432	0.66279	1.2413	1.9997	2.9382	4.0567
83	0.36811	0.16542	0.04273	0.00004	0.03736	0.15467	0.35198	0.62929	0.98661
84	0.16196	0.25245	0.36293	0.49342	0.64391	0.81439	1.0049	1.2154	1.4459
85	0.56491	0.56491	0.56490	0.56490	0.56490	0.56490	0.56490	0.56490	0.56491
86	0.01015	0.00000	0.00985	0.03970	0.08954	0.15939	0.24925	0.35909	0.48893
87	0.10245	0.27048	0.51851	0.84655	1.2544	1.7426	2.3107	2.9587	3.6867
88	0.05248	0.00501	0.13758	0.45014	0.94269	1.6152	2.4678	3.5003	4.7129
89	0.51587	0.10128	0.00668	0.23210	0.77750	1.6429	2.8283	4.3337	6.1591
90	7.5482	5.0508	3.0534	1.5560	0.55801	0.06121	0.06381	0.56641	1.5690
91	10.027	6.5872	3.8674	1.8675	0.58761	0.02774	0.18787	1.0680	2.6681
RSS	32.7067	22.1582	14.5290	9.81981	8.02982	9.16137	13.212	20.183	30.074

Vseeno pa narišemo še graf (1.4): na absciso nanesimo regresijske koeficiente, na ordinato pa vsote kvadratnih odklonov. Točke leže na paraboli, peta točka pa predstavlja minimum te parabole. Vsota kvadratov je pozitivna vrednost. Vsaka druga premica, razen optimalne, pa daje večjo vsoto kvadratnih odklonov.



Slika 1.4: Spreminjanje vsote kvadratov v odvisnosti od regresijskega koeficienta

Pri odločitvi smo torej poiskali minimum za vsoto kvadratnih odklonov. Kljub temu, da je metoda najmanjših kvadratov za nas še neznanca, pa smo pravkar spoznali, da bo še najbolj sprejemljiva. Sedaj pa metodo najmanjših kvadratov še izpeljimo!

1.3.2 Izpeljava metode v skalarni obliki

V tem poglavju se želimo osredotočiti na izpeljavo metode najmanjših kvadratov (OLS - Ordinary Least Square), ki je v živinoreji je pogosto uporabljena. Metodo bomo izbrali takrat, ko število opazovanj ni

uravnoteženo (balansirano) med posameznimi razredi. Istočasno pa mora biti izpolnjen še en osnovni pogoj: ostanki morajo biti neodvisno in identično porazdeljeni. Pri zgornjem modelu, opisanem v enačbah 1.1, ?? in 1.2, je slednji pogoj izpolnjen, kar dokazujeta oznaka *IID* pri opisu porazdelitve v enačbi 1.2 in tudi identična matrika **I** pri opisu strukture variance. Zgornji model bomo za izpeljavo metode najmanjših kvadratov razširili še za sistematski vpliv A_i z nivoji 1.7. Novi model se ne nanaša več na zgornje podatke. Primeren pa bi bil za analizo sprememb po letih za več farm. Takrat bi sicer vpliv z nivoji preimenovali, da bi nas oznaka spominjala na farmo.

$$y_{ij} = \mu + A_i + bx_{ij} + e_{ij} \quad [1.7]$$

$$y_{ij} \sim IID(\mu + A_i + bx_{ij}, \sigma^2) \quad [1.8]$$

Rešitve sistema poiščemo tako, da je vsota kvadratov za ostanek minimalna. Ker iščemo minimum vsote kvadratov, bomo funkcijo 1.7 odvajali po vseh neznankah in vse prve delne odvode izenačili z nič. Neznanke so lokacijski parametri μ , A_1 , A_2 , ... in A_{p_A} in b . Enačbo 1.9 za RSS bomo najprej malo preoblikovali.

$$RSS = \mathbf{e}'\mathbf{e} = \sum_{i=1}^p \sum_{j=1}^{n_i} (e_{ij})^2 \quad [1.9]$$

V enačbo vstavimo izraz za ostanek in poenostavimo zapis pri vsoti (1.10).

$$RSS = \sum_{ij} (y_{ij} - E(y_{ij}))^2 \quad [1.10]$$

Potrebujemo torej pričakovano vrednost, ki jo predstavlja praviloma le sistematski del modela (1.11).

$$E(y_{ij}) = \mu + A_i + bx_{ij} \quad [1.11]$$

Tako dobimo za vsoto kvadratov za ostanek zapisano v naslednji obliki:

$$RSS = \sum_{ij} (y_{ij} - \mu - A_i - bx_{ij})^2 \quad [1.12]$$

Od tu naprej sta dve možnosti. Po enem postopku lahko enačbo 1.12 razčlenimo. To lahko naredite sami za vajo. Tu pa bomo izhajali kar iz te enačbe. Torej vrnimo se k naši nalogi. Iščemo torej minimum funkcije za izračun RSS (1.13)!

$$\sum_{ij} (y_{ij} - \mu - A_i - bx_{ij})^2 = \min. \quad [1.13]$$

Funkcija ima obliko parabole. V prvem koraku moramo poiskati vse parcialne odvode. Najprej odvajamo na parameter μ (1.14).

$$\frac{\partial \sum_{ij} (y_{ij} - \mu - A_i - bx_{ij})^2}{\partial \mu} \quad [1.14]$$

Pri odvajanju bomo uporabili naslednje pravilo 1.15.

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g)}{\partial g(x)} \cdot \frac{\partial g(x)}{\partial x} \quad [1.15]$$

Rezultat v enačbi 1.16 izenačimo z 0. Hkrati parametre nadomestijo ocene, rešitve, ki bodo izpolnjevale pogoje zahtevane pri metodi najmanjših kvadratov. Parametre in njihove ocene ločimo tako, da ocene nosijo strešico (enačba 1.17).

$$2 * \sum_{ij} (y_{ij} - \mu - A_i - bx_{ij}) * (-1) \quad [1.16]$$

$$2 * \sum_{ij} (y_{ij} - \hat{\mu} - \hat{A}_i - \hat{b}x_{ij}) * (-1) = 0 \quad | \quad \div (-2) \quad [1.17]$$

V enačbi 1.17 lahko desno in levo stran delimo z 2 ter dobimo enačbo 1.18.

$$\sum_{ij} (-y_{ij} + \hat{\mu} + \hat{A}_i + \hat{b}x_{ij}) = 0 \quad [1.18]$$

Sedaj pa poskusimo enačbo še malo poenostaviti. Prav vseeno je, ali vsako meritev najprej očistimo sistematskih vplivov in tako dobljene ostanke seštejemo, kakor kaže enačba 1.18. Lahko pa levo stran enačbe 1.18 lahko najprej razčlenimo. Tako lahko posebej seštejemo vsa opazovanja ($\sum_{ij} y_{ij}$), nato še vse srednje vrednosti ($\sum_{ij} \hat{\mu}$), vplive A_i ($\sum_{ij} \hat{A}_i$) in prispevke regresije ($\sum_{ij} \hat{b}x_{ij}$) in na koncu opravimo odštevanje, kot je to nakazano v enačbi 1.19.

$$-\sum_{ij} y_{ij} + \sum_{ij} \hat{\mu} + \sum_{ij} \hat{A}_i + \sum_{ij} \hat{b}x_{ij} = 0 \quad [1.19]$$

Člene z neznanimi parametri zadržimo na levi strani enačbe, člen brez njih pa prenesimo na desno stran. Poleg tega pa opravimo še kratek premislek. Pri izračunu ostanka smo vsakemu opazovanju odšteli srednjo vrednost, torej smo to naredili natanko n -krat. Torej lahko vsoto nadomestimo z zmnožkom (enačba 1.20).

$$\sum_{ij} \hat{\mu} = n * \hat{\mu} \quad [1.20]$$

Podobno ravnamo tudi pri vplivih A_i . Nivojev za vpliv A_i je več in sicer smo predpostavili, da jih je p_A , kjer p_A predstavlja število nivojev in s tem tudi neznanih parametrov pri vplivu A . Vsota za vpliv A torej vključuje n_1 opazovanj pri nivoju A_1 , n_2 opazovanj pri nivoju A_2 , ... in n_{p_A} opazovanj pri nivoju A_{p_A} . Tretji člen iz enačbe 1.19 zapišemo v obliki 1.21.

$$\sum_{ij} \hat{A}_i = n_1 * \hat{A}_1 + n_2 * \hat{A}_2 + \dots + n_{p_A} * \hat{A}_{p_A} \quad [1.21]$$

Tudi četrti člen iz enačbe 1.19 lahko preuredimo, kot kaže enačba 1.22, ker je vsaki člen pri vsoti pomnožen s parametrom \hat{b} .

$$\sum_{ij} \hat{b}x_{ij} = \hat{b} \sum_{ij} x_{ij} \quad [1.22]$$

Sedaj smo si pogledali vsak člen posebej in lahko tako preurejene vstavimo v enačbo 1.19. Naredimo samo še nekaj: člene s parametri obdržimo na levi strani enačbe, člen brez parametrov pa prenesimo na desno stran. Tako dobimo enačbo 1.23.

$$n * \hat{\mu} + n_1 * \hat{A}_1 + n_2 * \hat{A}_2 + \dots + n_{p_A} * \hat{A}_{p_A} + \hat{b} \sum_{ij} x_{ij} = \sum_{ij} y_{ij} \quad [1.23]$$

Dobili smo enačbo, s katero lahko iz vrednotimo parameter μ . V enačbi pa so še drugi neznani parametri, zato moramo poiskati še druge enačbe. Da bi problem rešili, potrebujemo za vsak neznan parameter po eno enačbo. Na razpolago imamo še druge parcialne odvode.

Poskusimo odvajati tudi na parametre A_1, A_2, \dots, A_{p_A} . Postopek je identičen za vse parametre vpliva A , zato bomo izpeljali postopek za katerikoli parameter in ga označili z $A_{i'}$. Na koncu pa bomo razvili enačbo za vsak parameter posebej, saj mora imeti po eno enačbo za vsak neznan parameter. Odvajati moramo izraz v 1.24.

$$\frac{\partial \sum_{ij} (y_{ij} - \mu - A_i - bx_{ij})^2}{\partial A_{i'}} = \quad [1.24]$$

Dobimo dve možnosti. Kadar je parameter A , na katerega odvajamo, isti kot tisti iz tretjega člena v števcu enačbe 1.24, velja, da sta indeksa i in i' enaka. Tako moramo najti odvod iz 1.25.

$$\frac{\partial (-A_i)}{\partial A_i} = -1 \quad [1.25]$$

V primeru, da sta indeksa i in i' različna, pa moramo rešiti odvod v 1.26.

$$\frac{\partial (-A_i)}{\partial A_{i'}} = 0 \quad [1.26]$$

Končno lahko napišemo odvod, ki smo ga zastavili v enačbi 1.24. Omenili smo dve možnosti, ki jih lahko nakažemo, kot prikazuje enačba 1.27.

$$= 2 * \sum_{ij} (y_{ij} - \mu - A_i - bx_{ij}) * \begin{cases} (-1); & \text{kjer } i = i' \\ (0); & \text{kjer } i \neq i' \end{cases} \quad [1.27]$$

Druga možnost v enačbi 1.27 je 0, iz prve možnosti pa dobimo izraz 1.28.

$$= 2 * \sum_{ij} (y_{ij} - \mu - A_i - bx_{ij}) (-1); \quad \text{kjer } i = i' \quad [1.28]$$

Delimo z -2 in izenačimo z 0. Parametre ob tem zamenjamo z njihovimi ocenami (enačba 1.29). Spodnji izraz predstavlja več enačb in sicer za $i = 1, 2, \dots, p_A$.

$$\sum_{ij} (y_{ij} - \hat{\mu} - \hat{A}_i - \hat{b}x_{ij}) = 0; \quad \text{kjer } i = i' \quad [1.29]$$

Najprej prikažimo enačbo, kjer je $i = 1$. Vstavimo vrednost 1 namesto oznake i in tako dobimo 1.30.

$$\sum_{1j} y_{1j} - \sum_{1j} \hat{\mu} - \sum_{1j} \hat{A}_1 - \sum_{1j} \hat{b}x_{1j} = 0 \quad [1.30]$$

Prvi člen, ki ne vsebuje parametrov, prenesimo na desno stran. Drugi člen predstavlja vsoto srednjih vrednosti $\hat{\mu}$ za vse meritve opravljene pri nivoju A_1 . Meritev je natanko n_1 , srednje vrednosti imajo iste vrednosti, zato lahko drugi člen zapišemo tudi v obliki iz 1.31.

$$\sum_{1j} \hat{\mu} = n_1 \hat{\mu} \quad [1.31]$$

Tudi tretji člen lahko poenostavimo, kot prikazujemo v 1.32.

$$\sum_{1j} \widehat{A}_1 = n_1 * \widehat{A}_1 \quad [1.32]$$

Zadnji člen preuredimo po zgledu 1.22. Po preureditvi dobimo 1.33.

$$n_1 * \widehat{\mu} + n_1 * \widehat{A}_1 + \widehat{b} \sum_{1j} x_{1j} = \sum_{1j} y_{1j} \quad [1.33]$$

V enačbi ni členov z ostalimi parametri vpliva A , vendar jih lahko brez škode dodamo, če jih pomnožimo s konstanto 0.

$$n_1 * \widehat{\mu} + n_1 * \widehat{A}_1 + 0 * \widehat{A}_2 + \dots + 0 * \widehat{A}_{p_A} + \widehat{b} \sum_{1j} x_{1j} = \sum_{1j} y_{1j} \quad [1.34]$$

Postopek od koraka 1.30 do 1.34 lahko ponovimo še za vse i od 2 do p_A . Tako dobimo še preostale enačbe za neznane parametre iz vpliva A . Prikazujemo enačbi za $i = 2$ (1.35) in $i = p_A$ (1.36).

$$n_2 * \widehat{\mu} + 0 * \widehat{A}_1 + n_2 * \widehat{A}_2 + \dots + 0 * \widehat{A}_{p_A} + \widehat{b} \sum_{2j} x_{2j} = \sum_{2j} y_{2j} \quad [1.35]$$

$$n_{p_A} * \widehat{\mu} + 0 * \widehat{A}_1 + \dots + 0 * \widehat{A}_{p_A-1} + n_{p_A} * \widehat{A}_{p_A} + \widehat{b} \sum_{p_Aj} x_{p_Aj} = \sum_{p_Aj} y_{p_Aj} \quad [1.36]$$

Odvajati moramo še po enem parametru. To je regresijski koeficient b , kot smo zastavili v 1.37.

$$\frac{\partial \sum_{ij} (y_{ij} - \mu - A_i - bx_{ij})^2}{\partial b} \quad [1.37]$$

Pri odvodu moramo paziti, da odvajamo po parametru b in ne po neodvisni spremenljivki x_{ij} . To nas lahko hitro zavede, ker smo bili pri matematiki vajeni, da smo neznanko poimenovali z x . V našem primeru pa so neznani parametri v modelu. Torej je neznanka regresijski koeficient b , neodvisne spremenljivke x_{ij} pa so bile izmerjene, ko smo poskus opravili. Tako so za oznako skrite konstante. Odvod smo dobili v enačbi 1.38.

$$2 * \sum_{ij} (y_{ij} - \mu - A_i - bx_{ij}) * (-x_{ij}) \quad [1.38]$$

Izenačimo dobljeni parcialni odvod z nič in spremenimo parametre v ocene parametrov (parameter s strešico)! V enačbi 1.39 še nakažemo deljenje desne in leve strani s konstanto -2 .

$$2 * \sum_{ij} (y_{ij} - \widehat{\mu} - \widehat{A}_i - \widehat{b}x_{ij}) * (-x_{ij}) = 0 \quad | \div (-2) \quad [1.39]$$

Člene z ocenami parametrov zadržimo na desni strani enačbe, na levo pa prestavimo člene brez njih. Prestavljen člen je vsota produktov med opazovanji y_{ij} in pripadajočimi neodvisnimi spremenljivkami x_{ij} (enačba 1.40).

$$\sum_{ij} \widehat{\mu}x_{ij} + \sum_{ij} \widehat{A}_i x_{ij} + \widehat{b} \sum_{ij} x_{ij}^2 = \sum_{ij} y_{ij}x_{ij} \quad [1.40]$$

Iz vsot izpostavimo parametre (enačba 1.41). Pri tem dobimo nekaj več členov: pri vplivu A smo izpostavili parametre za posamezne nivoje.

$$\hat{\mu} \sum_{ij} x_{ij} + \hat{A}_1 \sum_{1j} x_{1j} + \hat{A}_2 \sum_{2j} x_{2j} + \dots + \hat{A}_{p_A} \sum_{p_A j} x_{p_A j} + \hat{b} \sum_{ij} x_{ij}^2 = \sum_{ij} y_{ij} x_{ij} \quad [1.41]$$

Tako, sedaj imamo vse prve parcialne odvode pripravljene ??, 1.33, 1.34, 1.35 in 1.41. Enačbe lahko uredimo v matrično obliko. Ocene parametrov $\hat{\mu}$, \hat{A}_1 , \hat{A}_2 , ..., \hat{A}_{p_A} in \hat{b} zberemo v vektor z ocenami parametrov $\hat{\beta}' = [\hat{\mu} \ \hat{A}_1 \ \hat{A}_2 \ \dots \ \hat{A}_{p_A} \ \hat{b}]$. Vse meritve nanizamo v vektor opazovanj \mathbf{y} . Matrika \mathbf{X} je matrika dogodkov za sistematske vplive in poskrbi, da potrebna opazovanja seštejemo. Začnimo na desni strani enačbe (*ang.* Right Hand Side, RHS): vektor je zmnožek matrike dogodkov in vektorja opazovanj. Torej ga lahko predstavimo z $\mathbf{X}'\mathbf{y}$. Na levi strani vektor ocen $\hat{\beta}$ že poznamo. Matrika na levi se imenuje tudi matrika koeficientov ali matrika varianc in kovarianc. Dobimo pa jo z množenjem transponirane matrike \mathbf{X}' z matriko dogodkov \mathbf{X} . Povsem sprejemljiva je torej oznaka $\mathbf{X}'\mathbf{X}$. Ker se z matrikami nismo ukvarjali jo bomo označili kar s črko \mathbf{C} .

$$\begin{bmatrix} n & n_1 & n_2 & \dots & n_{p_A} & \sum_{ij} x_{ij} \\ n_1 & n_1 & 0 & \dots & 0 & \sum_{1j} x_{1j} \\ n_2 & 0 & n_2 & \dots & 0 & \sum_{2j} x_{2j} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ n_{p_A} & 0 & 0 & \vdots & n_{p_A} & \sum_{p_A j} x_{p_A j} \\ \sum_{ij} x_{ij} & \sum_{1j} x_{1j} & \sum_{2j} x_{2j} & \dots & \sum_{p_A j} x_{p_A j} & \sum_{ij} x_{ij}^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{A}_1 \\ \hat{A}_2 \\ \vdots \\ \hat{A}_{p_A} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \sum_{ij} y_{ij} \\ \sum_{1j} y_{1j} \\ \sum_{2j} y_{2j} \\ \vdots \\ \sum_{p_A} y_{p_A j} \\ \sum_{ij} y_{ij} x_{ij} \end{bmatrix} \quad [1.42]$$

Sistem enačb 1.42 lahko zapišemo v matrični obliki 1.43. Dokaz si lahko preberete v poglavju, ki prikazuje izpeljavo metode najmanjših kvadratov v matrični obliki (??).

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \quad [1.43]$$

$$\mathbf{C}\hat{\beta} = \mathbf{X}'\mathbf{y} \quad [1.44]$$

Sistem enačb pri metodi najmanjših kvadratov si bomo dobro zapomnili in ga znali tudi nastaviti.

1.3.3 Vaje

Izpeljite metodo najmanjših kvadratov in nastavite sistem enačb v matrični obliki za naslednje modele!

$$y_i = \mu + b_{11}x_{1i} + b_{12}x_{1i}^2 + b_{13}\sin(x_{1i}) + b_{21}x_{2i} + b_{22}x_{2i}^2 + e_i \quad [1.45]$$

$$y_{ijk} = \mu + P_i + S_j + e_{ijk} \quad [1.46]$$

$$y_{ijk} = \mu + P_i + S_j + PS_{ij} + e_{ijk} \quad [1.47]$$

$$y_{ijk} = \mu + P_i + S_j + b_x x_{ijk} + e_{ijk} \quad [1.48]$$

$$y_{ijk} = \mu + P_i + S_j + b_x x_{ijk} + b_z z_{ijk} + e_{ijk} \quad [1.49]$$

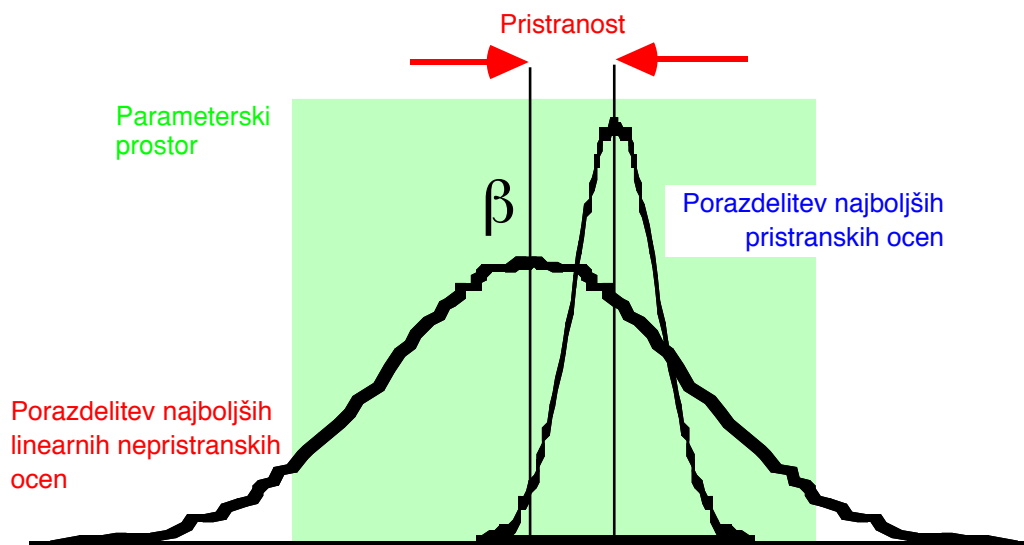
$$y_{ijk} = \mu + P_i + S_j + b_{11}x_{1i} + b_{12}x_{1i}^2 + b_{21}x_{2j} + b_{22}x_{2j}^2 + e_{ijk} \quad [1.50]$$

Pri modelih 1.46, 1.47, 1.48, 1.49 in 1.50 vzemite, da je $i = 1, 2, 3$ in $j = 1, 2, 3, 4$.

1.3.4 Pristranost in nepristranost linearnih ocen

Pa le omenimo primer iz živinoreje, da bi se prehitro ne razveselili. Prav verjetno bo kdo izmed vas našel službo v selekcijski službi. Tam se na veliko ukvarjamo z ocenjevanjem komponent (ko)variance, ki je osnovana na vsotah kvadratov. Vsote kvadratov so "kvadratne kombinacije opazovanj", imenovali jih bomo strokovno tudi kvadratne oblike. Problem rešimo z več metodami. Nekatere med njimi so celo nepristranske, vendar pa je rezultat včasih prav nemogoč. Tako dobimo lahko negativne variance. Delež posameznih komponent variance preseže vrednost 1, kar pomeni 100 %. To bi v praksi pomenilo, da je ena komponenta variance večja kot vsota vseh. Če iz komponent varianc in kovarianc izračunamo korelacije, so ocene izven parametrskega prostora - izven intervala možnih vrednosti. Tako so se bolj obnesle metode, ki dajejo pristranske rezultate. Za te metode, ki jih običajno tudi izberemo, je značilno, da so asimptotično nepristranske. Kadar imamo dovolj podatkov, so ocene torej tako malo pristranske, da lahko to zanemarimo. Ne smemo pa na to pozabiti in komponent variance ocenjevati z nekaj 100 ali še manj meritvami! tudi 10000 meritev ni prav veliko, raje jih imamo nekaj 100000. Nekaj pa je tudi izjem, a o njih ne bomo razpravljali prezgodaj. Tako se kaj rado zgodi, da postanejo običajna praksa in celo dokaz za "pravilnost" napačnih pristopov.

Poskusimo primerjati nepristransko in pristransko metodo (1.5). Sliko smo malo pretiravali, da bi bila razlika bolj jasna. Če imamo kolikor toliko dobre podatke, bodo razlike bistveno manjše. Pristranske ocene bodo manj precenjene, nepristranske pa bolj zanesljivo ocenjene (manj razpršene). Zeleno oziroma pikčasto območje pa predstavlja parametrski prostor ali zalogo vrednosti. Kar je večje ali manjše od parametrskega prostora, so vrednosti, ki jih naš proučevani parameter pač ne mora imeti v nobenem primeru.



Slika 1.5: Primejava pristranske in nepristranske metode

Nepristranske meritve bodo, če bomo poskuse ponavljali, lahko precej različne. Tako kot vedno bodo natančnost izvedbe preizkusa, natančnost meritev in število meritev odločali o zanesljivosti ocene. Pri vsakem od ponovljenem poskusu pa obstaja enaka verjetnost, da bo rezultat večji ali manjši od parametra, zato ne govorimo o tem, da so rezultati v posameznem poskusu pristranski. Ocene so porazdeljene okrog parametra. Ker dejanskih vrednosti parametrov ne poznamo, tudi ne moremo oceniti za koliko odstopajo. To bi lahko dobili le, če bi poskus mnogokrat ponovili. To pa bi ne bilo nič drugega kot en sam večji poskus. Ker pa poskusi vedno stanejo, delamo malo večje poskuse šele, ko smo zbrali zadosti zanesljivih dokazov, da je uspeh zagotovljen.

Pristranske ocene pa niso porazdeljene okrog parametra. Vrh porazdelitve je pomaknjen v desno. Ocene, ki nam jih metoda ponuja, so bolj pogosto večje od parametra. V tem primeru so ocene precenjene, na naši sliki skoraj vedno. Če znamo izračunati pričakovano vrednost, bomo lahko ocenili tudi pristranost.

Pristranost ne bo ocenjevala odstopanje našega rezultata, ampak odstopanja pričakovane vrednosti ocene od parametra. Običajno je to funkcija parametra.

1.4 Metode preverjanja ocenljivosti

Metodo preverjanja ocenljivosti bomo ponazorili v matrični obliki.

Linearna kombinacija \mathbf{k}' je ocenljiva, če velja:

$$\mathbf{k}'\mathbf{C}^{-1}\mathbf{C} = \mathbf{k}' \quad [1.51]$$

Matrika

Linearne kombinacije iščemo pri interpretaciji rezultatov in testiranju hipotez. Obravnavamo lahko samo tiste linearne kombinacije \mathbf{k}' , ki so ocenljive. Tako smo že ugotavljali, da vpliv pasme ni ocenljiv, ocenljive pa so razlike med pasmami - pa še to ne vedno! V primeru, da ni ocenljiva razlika med pasmama, je gotovo slabo načrtovan poskus: struktura podatkov je slaba.

Le redko se lotimo testiranja posameznih hipotez. Radi imamo najprej namig, ali se z določeno skupino sploh splača ukvarjati. Npr., zanima nas, ali se pasme značilno razlikujejo. Vse linearno neodvisne kombinacije nanizamo v matriko \mathbf{K} in preizkusimo ocenljivost.

$$\mathbf{K} = \begin{bmatrix} \mathbf{k}'_1 \\ \mathbf{k}'_2 \\ \vdots \\ \mathbf{k}'_p \end{bmatrix} \quad [1.52]$$

Tako velja:

$$\mathbf{K}\mathbf{C}^{-1}\mathbf{C} = \mathbf{K} \quad [1.53]$$

Hipoteze iz matrike \mathbf{K} so ocenljive.

$$\mathbf{H}\mathbf{C}^{-1}\mathbf{C} = \mathbf{H} \quad [1.54]$$

Hipoteze iz matrike \mathbf{H} niso ocenljive, vsaj ena med njimi ni, medtem ko so hipoteze iz matrike \mathbf{K} iz enačbe 1.54 ocenljive.

Primer

$$\mathbf{C} = \begin{bmatrix} 22 & 6 & 8 & 8 \\ 6 & 6 & & \\ 8 & & 8 & \\ 8 & & & 8 \end{bmatrix} \quad [1.55]$$

Če črtamo prvo vrstico in prvi stolpec (enačba), smo v preostanku matrike dobili diagonalno matriko s tremi vrsticami in stolpci.

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 6 & & \\ 0 & & 8 & \\ 0 & & & 8 \end{bmatrix} \quad [1.56]$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1/6 & & \\ 0 & & 1/8 & \\ 0 & & & 1/8 \end{bmatrix} = \mathbf{A}^- = \mathbf{C}^- \quad [1.57]$$

- Ali lahko napišete model za zgornji primer?
- Dobili smo eno od neskončno mnogo splošnih inverz:
... črtali prvo vrstico in prvi stolpec ...

$$\mathbf{C}^- \mathbf{C} = \quad [1.58]$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1/6 & & \\ 0 & & 1/8 & \\ 0 & & & 1/8 \end{bmatrix} \begin{bmatrix} 22 & 6 & 8 & 8 \\ 6 & 6 & & \\ 8 & & 8 & \\ 8 & & & 8 \end{bmatrix} \quad [1.59]$$

- Pazimo na vrstni red matrik!
- Matrika služi kot filter za preverjanje hipotez

$$\mathbf{K} \mathbf{C}^- \mathbf{C} = \quad [1.60]$$

$$\begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0.5 & 0 & -0.5 \\ 0 & 0 & 2 & -2 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad [1.61]$$

- Pogoju ocenljivosti smo zadostili
- Lahko bi uporabili katerokoli drugo splošno inverzo

$$\mathbf{C} = \begin{bmatrix} 22 & 6 & 8 & 0 \\ 6 & 6 & & 0 \\ 8 & & 8 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad [1.62]$$

$$\mathbf{C}^- = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1/6 & & \\ 0 & & 1/8 & \\ 0 & & & 1/8 \end{bmatrix} \quad [1.63]$$