UNIVERZITET U NOVOM SADU
PRIRODNO – MATEMATIČKI FAKULTET
DEPARTMAN ZA MATEMATIKU I INFORMATIKU

Maja Zobenica

Analiza longitudinalnih podataka

- Master rad -

Novi Sad, 2013.

SADRŽAJ

Uvod

Promene su prisutne u svakodnevnom životu: bebe uče da hodaju i govore, deca uče da čitaju i pišu, starije osobe postaju zaboravne. Pored ovih prirodnih promena, promene mogu prouzrokovati i određene intervencije, npr. rezultati ispita se mogu popraviti posle nastavnog kursa, krvni pritisak se može regulisati korišćenjem lekova. Merenjem i beleženjem prirodnih i eksperimentalnih promena zaključuje se da je vreme bitan faktor promena. Trenutni podaci koji su laki za prikupljanje i široko dostupni, nisu dovoljni za ispitivanje tih promena.

Istraživanje promena se radi već generacijama. Metodolozi su tek 1980-ih godina razvili klasu odgovarajućih statističkih modela preko kojih su istraživači mogli da proučavaju ovakve promene. 1960-ih i 1970-ih godina veći broj metodologa je insistirao da istraživači ne treba ni da pokušaju da mere promene. Danas podatke prikupljene posmatranjem ovakvih promena nazivamo longitudinalni podaci. Longitudinalne studije predstavljaju studije longitudinalnih podataka, odnosno one u kojima je određena karakteristika izmerena na istom subjektu nekoliko puta. Ovakve studije mogu biti eksperimentalne ili posmatračke. Većina longitudinalnih studija su posmatračke (posmatraju stanje bez manipulacije, uplitanja posmatrača). Neki istraživači smatraju da takva istraživanja imaju manju snagu da otkriju uzročne veze nego eksperimentalne studije. Ipak, zbog ponovljenog posmatranja na individualnom nivou imaju veću moć od studija sa trenutnim posmatranjima. Vreme može biti izraženo različitim jedinicama: godinama, semestrima, mesecima, danima i tako dalje. Raspored prikupljanja podataka može biti fiksni (svaka osoba ima istu periodičnost) ili fleksibilni (svaka osoba ima jedinstven raspored). Longitudinalne studije se bave rešavanjem pitanja kako se rezultat posmatranja menja tokom vremena i kako možemo predvideti razlike tih promena. Prvo pitanje je deskriptivno i obuhvata određene osobine svake jedinke tokom vremena, kao i oblik promena rezultata posmatranja. Drugo pitanje je odrediti relaciju koja opisuje vezu između nezavisnih promenljivi (ponovljenih merenja) i načina promena.

Zbog svojih prednosti, analiza longitudinalnih podataka je našla primenu u mnogim oblastima. U medicini se koristi da otkrije pokazatelje određenih bolesti, u psihologiji i ekonomiji da pokaže vezu između određenih pojava, u marketingu se koristi da identifikuje promene koje se su se desile u stavovima i ponašanju ciljne publike usled određene reklamne kampanje.

Kada se kao objekat istraživanja uzme promena neke pojave, karakteristike, jedini način ispitivanja je da se prikupe podaci koji su rezultat višestrukog merenja. Posmatranja jednog subjekta nisu međusobno nezavisna, pa je stoga neophodno primeniti određene statističke tehnike koje uzimaju u obzir činjenicu da su ponovljena posmatranja svakog subjekta povezana (u korelaciji). Osnova metoda analize koje ćemo predstaviti u ovom radu je analiziranje kovarijansne ili korelacione matrice.

1. Osnovni pojmovi

Longitudinalni podaci predstavljaju podatke u vidu ponovljenih merenja koji su prikupljeni na jedinkama posmatrane populacije (ili češće na uzorku iz populacije). Termin longitudinalno znači uzdužno i sugeriše da su podaci prikupljeni tokom vremena. Iako se ponovljena merenja najčešće dešavaju tokom vremena, to nije jedini način na koji se merenja mogu uzeti više puta na istoj jedinki. Na primer:

- Jedinke mogu biti ljudi. Za svakog čoveka, kao subjekat, smanjenje dijastolnog (donjeg) krvnog pritiska je izmereno nekoliko puta i svako od njih uključuje davanje različitih doza leka. Prema tome, subjekat je meren više puta u zavisnosti od doziranja.
- Jedinke može biti drveće u šumi. Za svako drvo se meri prečnik stabla u nekoliko različitih tačaka duž stabla. Dakle, drvo je izmereno više puta u zavisnosti od pozicije duž stabla.
- Jedinke mogu biti učenici. Svaki učenik daje ocenu kao meru lepote azijatskog, belačkog i mešanog lica. Ovde se beleži ocena više puta u odnosu na boju kože.

Treći primer je malo različit od prva dva jer ne postoji prirodan red kako bi se merenja ponovila.

Za nas termin longitudinalni podaci predstavlja podatke u obliku ponovljenih merenja koji mogu biti u zavisnosti od vremena, što je i najčešće, ali takođe mogu biti i u zavisnosti od drugih uslova.

Kod longitudinalnog istraživanja prvo određujemo pitanje koje nas zanima (pitanje od interesa), a potom ga rešavamo odgovarajućom analizom. Koristićemo termin „rezultat“ za označavanje rezultata posmatranja koja nas interesuju. Pošto su jedinke uglavnom ljudski ili životinjski subjekti, koristiće se termini jedinka, individua i subjekat naizmenično.

1.1. Razlika između trenutnih i longitudinalnih podataka

Koncept istraživanja u mnogim studijama zavisi od prirode pitanja koja se istražuju. Prvi korak u sprovođenju analize je da saznamo koju vrstu podataka će istraživanje prikupiti što je poznato kao metodologija.

Recimo da želimo ispitati odnos između svakodnevnih treninga i telesne težine. Jedna od prvih stvari koje moramo odrediti je tip istraživanja koji će nam najviše reći o tom odnosu. Da li želimo da uporedimo telesnu težinu između različitih populacija sportista i nesportista u istom trenutku? Ili želimo da merimo telesnu težinu u jednoj populaciji sportista tokom određenog vremenskog perioda? Prvi pristup je tipičan za trenutna istraživanja, a drugi za longitudinalna istraživanja.

Analiza trenutnih podataka je posmatračka. To znači da istraživači beleže informacije o određenim subjektima bez menjanja njihovog okruženja. U našem slučaju, mi bismo jednostavno merili telesnu težinu sportista i nesportista zajedno sa drugim karakteristikama koje mogu biti korisne. Takođe, ne treba podsticati nesportiste da vežbaju, niti sportiste da manje ili više vežbaju. Ukratko, ne treba da se uplićemo.

Bitna kartakteristika trenutnog istraživanja je ta što ono može da poredi različite grupe populacije u istom trenutku. Možemo razmišljati o tome kao o trenutku fotografisanja tj. snimku.

U predstavljenom primeru možemo izabrati da merimo telesnu težinu sportista kroz dve starosne grupe, preko 30 i ispod 30 godina i uporediti ih sa telesnom težinom nesportista u istim starosnim grupama. Možemo, na primer, napraviti podgrupe i gledati pol, režim ishrane, prihod i stepen obrazovanja u odnosu na vežbanje i težine, sa malo ili bez dodatnih troškova. U svakom slučaju, ne bismo razmatrali težinu iz prošlosti ili budućnosti. Gledali bi samo težine u jednom trenutku u vremenu.

Trenutna istraživanja ne mogu dati informacije o uzročno-posledičnim odnosima. To je zato što ova istraživanja nude „snimak“ jednog trenutka u vremenu, ne razmatraju šta se desilo pre ili posle tog snimka. Dakle, ne možemo za sigurno znati da li su naši sportisti imali manju težinu pre režima vežbanja ili da li su treninzi sa sigurnošću pomogli snižavanju težine koji je prethodno bila velika. Analiza longitudinalnih podataka daje mnogo više informacija o uzorcima. Ključ je u tome što longitudinalna analiza postoji i van određenog momenta u vremenu.

Longitudinalno istraživanje je takođe posmatračko. Longitudinalni podaci su podaci dobijeni posmatranjem subjekata (ljudi, životinja, biljaka, laboratorijskih uzoraka itd.) koji su izmereni više puta na istom subjektu tokom određenog vremenskog perioda. Taj vremenski period ponekad traje i godinama. Svrha longitudinalnog istraživanja je da analiziramo promene tretmana tokom određenog perioda. Kada je promena uzeta kao objekat istraživanja, jedini način ispitivanja te promene je da se prikupe podaci koji su rezultat višestrukog merenja. Pored medicine, podaci prikupljeni na ovaj način se široko primenjuju u poljoprivredi, biologiji, fizici i tehničkim naukama.

Prikupljanje longitudinalnih podataka može biti izazovno. Uzorak je potrebno izmeriti bar dva puta, što košta više nego trenutni podaci. Takođe, u svakom momentu moramo da osiguramo saradnju onih koji su u startu učestvovali, jer ne možemo zameniti subjekte koji odbiju saradnju, sa drugima koji nisu učestvovali na prethodnom merenju. Čak i u dobro definisanom i kontrolisanom istraživanju, takve situacije se neminovno javljaju u longitudinalnim istraživanjima. Ovi problemi se javljaju pod nazivom nedostatak podataka. Prikupljanjem longitudinalnih podataka, pošto merimo isti subjekat više puta, dobijamo zavisna posmatranja.

Primer longitudinalnih podataka: Posmatra se šest učenika tokom pet nedelja koliko imaju opravdanih časova. Podaci su dati u sledećoj tabeli:

Tabela 1. Longitudinalni podaci.

i	**1**	**2**	**3**	**4**	**5**
1	5	3	6	21	0
2	3	5	0	3	0
3	2	4	0	21	6
4	4	6	1	4	3
5	7	18	11	34	8
6	0	12	22	6	2

1.2. Terminologija i notacija

Polazna osnova analize prirode nekog fenomena su podaci koji se odnose na jedan ili više skupova objekata (ljudi, biljke, predmeti, pojave itd.). Često nismo u prilici da kompleksnu prirodu objekata sagledamo u potpunosti, ali na raspolaganju nam stoji mogućnost da obuhvatimo njihove različite karakteristike. Te karakteristike, odnosno obeležja predstavljaju predmet našeg merenja i modeliramo ih slučajnim promenljivama. Model je statistički kada promenljive nisu deterministički, već stohastički povezane. Statistički model opisuje kako se jedna ili više slučajnih promenljiva odnose na jednu ili više drugih promenljivih. Statistički model koristi raspodele verovatnoća za opisivanje mehanizma koji generiše podatke. To jest, rezultati su predstavljeni slučajnim promenljivama čije se raspodele verovatnoća koriste za opisivanje promena koje se dešavaju kada rezultat uzima različite vrednosti. Varijaciju rezultata uzrokuje više faktora. Prema tome, da bi se izgradio statistički model i odlučilo koja raspodela verovatnoća je pogodna, treba pažljivo razmotriti sve karakteristike.

Sa Y ćemo označavati obeležje koje posmatramo. $E(Y)$ označava očekivanu vrednost slučajne promenljive Y (matematičko očekivanje ili prosek populacije). Često se koristi i simbol μ.

Varijansa populacije (disperzija) $var(Y)$ se može posmatrati kao mera rasipanja svih mogućih vrednosti koje se posmatraju od očekivane vrednosti,

$$\sigma^2 = var(Y) = E\left[\left(Y - E(Y)\right)^2\right].$$

Upotreba kvadratnog odstupanja uzima u obzir udaljenost od "centra", ali ne i pravac, tako da se meri širenje uopšte (u bilo kom smeru).

Standardna devijacija populacije σ je apsolutna mera disperzije populacije. Ona nam govori koliko u proseku elementi skupa odstupaju od aritmetičke sredine skupa. Označava se grčkim slovom sigma (σ). Formula za njeno izračunavanje je

$$\sigma = \sqrt{var(Y)}.$$

Često se u istraživanjima koristi uzorak. Ako se posmatra n subjekata, imaćemo slučajne promenjive $Y_1, \dots, Y_n$. Očekivanje uzorka $\bar{Y}$ je ocena proseka populacije:

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

Varijansa uzorka se koristi kao ocena varijanse populacije. Deljenje sa $(n-1)$, pre nego sa n, koristi se da bi ocena bila nepristrasna (dobro procenjuje varijansu populacije čak i kada je veličina uzorka n mala):

$$S^2 = (n-1)^{-1}\sum_{j=1}^{n}\left(Y_j - \bar{Y}\right)^2.$$

Standardna devijacija uzorka je pozitivni kvadratni koren varijanse uzorka i predstavljena je simbolom *S*.

Kovarijansa je merilo koliko dve slučajne promenljive, Y_j i Y_k variraju zajedno. Kovarijansa između Y_j i Y_k je:

$$cov(Y_j, Y_k) = E\left[\left(Y_j - \mu_j\right)(Y_k - \mu_k)\right],$$

gde je μ_j očekivanje za Y_j, a μ_k očekivanje za Y_k. Za $j,k = 1,\dots,n$ je $var(Y_i) = \sigma_j^2$, a kovarijansa je $cov(Y_j, Y_k) = \sigma_{jk}$.

Različiti su uzroci varijacija obeležje koja posmatramo:

1. ***Biološka varijacija.*** Poznato je da su biološki subjekti različiti, iako bića istog tipa imaju tendenciju da imaju slične osobine (karakteristike), oni nisu isti (osim možda u slučaju genetski identičnih klonova). Dakle, čak i ako posmatramo grupu osoba istih karakteristika (godine, pol i slično) i merimo njihovu težinu, očekujemo varijacije u mogućim rezultatima posmatranih subjekata, a to možemo očekivati zbog prirodne biološke varijacije.

Neka Y predstavlja težinu slučajno odabrane osobe iz posmatrane grupe, koja ima funkciju raspodele sa očekivanjem μ. Kada bi sve osobe bile biološke identične varijansa populacije Y bi bila jednaka 0 i očekivali bismo da sve osobe imaju tačno težinu μ. Kako osobe nisu biološki identične varijansa će biti veća od 0. Težina slučajno odabrane osobe nije jednaka μ već odstupa od μ za neku pozitivnu ili negativnu veličinu. Možemo predstaviti jednostavan statistički model za Y

$$Y = \mu + b \tag{1.1}$$

gde je b slučajna promenljiva sa očekivanjem $E(b) = 0$ i varijansom $var(b) = \sigma_b^2$. Primetimo da se Y sastoji iz dva sabirka: očekivane vrednosti μ i slučajnog odstupanja b koje predstavlja za koliko težina osobe može odstupati od očekivane vrednosti težine zbog specifičnih bioloških faktora.

2. ***Greške merenja.*** Razmatrali smo težinu kao karakteristiku jedne populacije i kada pred sobom imamo osobu čiju težinu merimo, možemo registrovati njenu vrednost. U idealnom slučaju, izmerena veličina bi predstavljala pravu težinu osobe svaki put kad se meri, ali uglavnom su uređaji kojim se takva merenja vrše nesavršeni, pa merenja iste jedinke mogu varirati svaki put. Veličina za koju se merenje razlikuje od prave vrednosti smatra se greškom merenja - odnosno odstupanje na više ili na niže od prave vrednosti koja bi bila izmerena sa savršenim uređajem. Sa „fer“ ili nepristrasnim uređajem greške mogu biti u bilo kom smeru bez pravila. Dakle, ako imamo samo jednu objektivnu skalu na kojoj se meri težina ljudi, izmerena težina osobe koju posmatramo ne odražava samo pravu težinu (koja varira među osobama koje posmatramo), već i grešku merenja. Slučajnu promenljivu koja sadrži sve moguće greške merenja težine ljudi posmatrane populacije, označićemo sa e.

S obzirom da smatramo da težina ljudi varira i zbog bioloških faktora i zbog grešaka merenja, proširićemo statistički model za Y (izmerena težina slučajno izabranog subjekta) tako da ga čine sve vrednosti sastavljene od stvarne težine osobe iz populacije i svih navedenih mogućih grešaka. Dobijamo,

$$Y = \mu + b + e = \mu + \varepsilon, \tag{1.2}$$

gde je b već predstavljeno u (1) tj. b je odstupanje uzrokovano specifičnim biološkim faktorima, sa $E(b) = 0$ i varijansom $var(b) = \sigma_b^2$, a e je greška pri merenju.

U (1.2) $\varepsilon = b + e$ predstavlja odstupanje uzrokovano kako biološkom varijacijom, tako i greškom merenja. Pri tome važi: $E(\varepsilon) = 0$ i $var(\varepsilon) = \sigma^2 = \sigma_b^2 + \sigma_e^2$, a iz toga dalje sledi $E(Y) = \mu$ i $var(Y) = \sigma^2$. σ^2 odražava odstupanja stvarnih težina i odstupanja zbog mogućih grešaka.

Postoje i drugi izvori varijacije koje bismo mogli uzeti u obzir. Za sada je važno naglasiti da postoje različiti izvori variacije koje uzrokuju da rezultati merenja variraju.

Rezultate ponovljenih posmatranja iste jedinke potrebno je sagledati zajedno, pa zato uvodimo slučajne vektore. Slučajni vektor je vektor čiji su elementi slučajne promenljive:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}.$$

Svaki element od $\boldsymbol{Y}$, $Y_j, j = 1, \dots, n,$ je slučajna promenljiva sa svojim očekivanjem, varijansom i raspodelom:

$$E(Y_j) = \mu_j, \;\; var(Y_j) = E\left[(Y_j - \mu_j)^2\right] = \sigma_j^2.$$

Dalje pretpostavljamo da Y_j ima normalnu raspodelu:

$$Y_j \sim \mathcal{N}(\mu_j, \; \sigma_j^2).$$

Raspodelu jedne slučajne promenljive često zovemo jednovarijantna (univarijantna) raspodela. Raspodela više slučejnih promenljivih se naziva multivarijantna.

Ako posmatramo elemente od $\boldsymbol{Y}$ zajedno, moramo uzeti u obzir činjenicu da oni dolaze zajedno u grupi, pa će verovatno postojati veza između njih. Specijalno, ako posmatramo $\boldsymbol{Y}$ tako da sadrži moguća posmatranja iste jedinke, za očekivati je da će realizovana vrednost jednog posmatranja u jednom vremenskom momentu biti u vezi sa vrednošću u drugom vremenskom momenu. Na primer, ako merimo težinu jedne osobe tokom n dana, možemo očekivati da ukoliko izmerena vrednost težine „velika“ jednog dana, da će biti „velika“ i narednog dana. Dakle, ako je prirodno misliti da su posmatranja jedne jedinke u vezi onda želimo da vidimo kako posmatranja posmatrana zajedno uzimaju svoje vrednosti.

Za n - dimenzionalni slučajni vektor $\boldsymbol{Y}$, uzimamo očekivanja svakog elementa i predstavljamo ih u obliku vektora

$$\boldsymbol{\mu} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}.$$

Definišemo očekivanu vrednost od $\boldsymbol{Y}$ kao:

$$E(\boldsymbol{Y}) = \boldsymbol{\mu}.$$

Analogno slučajnom vektoru, slučajna matrica se definiše kao matrica čiji su elementi slučajne promenljive, pri čemu svaki element ima svoje očekivanje. Možemo razmatrati kovarijanse između elemenata slučajnog vektora:

$$(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})' =$$

$$= \begin{bmatrix} (Y_1 - \mu_1)^2 & (Y_1 - \mu_1)(Y_2 - \mu_2) & \dots & (Y_1 - \mu_1)(Y_n - \mu_n) \\ (Y_2 - \mu_2)(Y_1 - \mu_1) & (Y_2 - \mu_2)^2 & \dots & (Y_2 - \mu_2)(Y_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (Y_n - \mu_n)(Y_1 - \mu_1) & (Y_n - \mu_n)(Y_2 - \mu_2) & \dots & (Y_n - \mu_n)^2 \end{bmatrix}$$

Dalje je,

$$E[(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})'] =$$

$$= \begin{bmatrix} E[(Y_1 - \mu_1)^2] & E[(Y_1 - \mu_1)(Y_2 - \mu_2)] & \dots & E[(Y_1 - \mu_1)(Y_n - \mu_n)] \\ E[(Y_2 - \mu_2)(Y_1 - \mu_1)] & E[(Y_2 - \mu_2)^2] & \dots & E[(Y_2 - \mu_2)(Y_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(Y_n - \mu_n)(Y_1 - \mu_1)] & E[(Y_n - \mu_n)(Y_2 - \mu_2)] & \dots & E[(Y_n - \mu_n)^2] \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix} = \boldsymbol{\Sigma}, \;\; j,k = 1,\dots,n, \;\; var(Y_j) = \sigma_j^2, cov(Y_j, Y_k) = \sigma_{jk}.$$

Matrica $\boldsymbol{\Sigma}$ se zove kovarijansna matrica za $\boldsymbol{Y}$. Kako je $\sigma_{jk} = \sigma_{kj}$, matrica $\boldsymbol{\Sigma}$ je simetrična kvadratna matrica formata $(n \times n)$.

Dalje sledi da ako razmatramo zajedničku raspodelu za sve elemente od $\boldsymbol{Y}$, $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ su karkteristike ove raspodele (centar i širenje). Nazivamo ih očekivanje populacije i kovarijansna matrica populacije.

Koeficijent korelacije predstavlja meru zajedničkog variranja dve ili više promenljivih i stepena njihove linearne povezanosti. On pokazuje da li između variranja dve promenljive postoji kvantitativno slaganje. Koeficijent korelacije populacije između Y_j i Y_k se definiše kao

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_j^2}\sqrt{\sigma_k^2}},$$

gde $\sigma_j = \sqrt{\sigma_j^2}$ predstavlja standardnu devijaciju populacije za Y_j, na istoj skali merenja kao Y_j i analogno je σ_k standardna devijacija za Y_k. Vidimo da koeficijent korelacije predstavlja normalizovanu kovarijansu između Y_j i Y_k.

Koeficijent korelacije zadovoljava sledeću jednačinu

$$-1 \le \rho_{jk} \le 1,$$

gde znak koeficijenta govori o smeru povezanosti dve slučajne promenljive. Ukoliko koeficijent ρ_{jk} uzme donju ili gornju graničnu vrednost, tada kažemo da postoji perfektna linearna veza između Y_j i Y_k i to sa negativnim ($\rho_{jk} = -1$), odnosno pozitivnim predznakom ($\rho_{jk} = 1$). Kada je $\rho_{jk} = 0$ ($\sigma_{jk} = 0$) znači da su promenljive Y_j i Y_k nekorelirane. U slučaju da imaju zajedničku normalnu raspodelu, one su i nezavisne.

Koeficijenti korelacije se prikazuju u obliku matrice

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix},$$

koju nazivamo korelaciona matrica za slučajni vektor.

1.3. Modeli linearne regresije

Jedan od ciljeva u velikom broju istraživanja je da se opiše veza među pojavama koje nas okružuju. To se može postići pronalaženjem formule ili jednačine koja povezuje veličine koje posmatramo. Na primer, može nas interesovati veza između temperature i pritiska u hemijskom procesu; ili veza između broja jabuka na drveću u voćnjaku i količine đubriva koje je korišćeno u voćnjaku; ili kako nova vakcina utiče na bolest. Veza između količine padavina, temperature i vlažnosti ili veza između prinosa i sorte žita.

U statistici pronalaženjem statističkih veza između pojava bavi se regresiona analiza. Regresiona analiza je od velikog značaja, kako u ekonomiji i privredi, tako i u drugim prirodnim naukama, kao što su: hemija, fizika, biologija, farmakologija, toksikologija, biohemija i sudska medicina...

Problem opisivanja ovakvih veza svodi se na pronalaženje modela koji povezuje jednu ili više *zavisnih* promenljivh $Y_1, Y_2, ..., Y_p$ sa jednom ili više *nezavisnihih, objašnjavajućih,* promenljivih $x_1, x_2, ..., x_k$ pomoću neke funkcionalne zavisnosti. Oblik ove funkcionalne zavisnosti je najčešće nepoznat, pa ostaje na istraživaču da izabere onu koja je po nekom kriterijumu najbolja. Veoma često se koriste polinomne funkcije, ali isto tako i eksponencijalne ili neke druge funkcije. Opšti problem nalaženja funkcije koja dobro aproksimira dobijeni skup podataka, često se naziva "*fitovanje*" krive ili određivanje *regresione linije.*

Veza između promenljivih može biti različitog oblika, a regresioni model kojim se opisuje linearna međuzavisnost između dve promenljive naziva se prosti linearni regresioni model.

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

gde su Y i ϵ slučajne promenljive, x je neslučajna promenljiva. Nezavisna promenljiva x je kontrolisana, vrednosti zavisno promenljive Y se mogu meriti, dok se vrednosti promenljive ϵ, koja se naziva i greška, ne mogu meriti. Promenljiva ϵ sadrži u sebi sve ostale promenljive koje utiču na vrednosti promenljive Y .

Uzorački model linearne regresije se dobija kada za fiksirane vrednosti $x_1, \dots, x_n$, posmatramo odgovarajuće slučajne promenljive $Y_j, j = 1, \dots, n$. Uzorački model proste linearne regresije predstavlja se na sledeći način:

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j, \ j = 1, \dots, n \tag{1.5}$$

gde je ϵ_j slučajna promenljiva takva da važi $E(\epsilon_j) = 0, \ var(\epsilon_j) = \sigma^2$. Sledi da je $E(Y_j) = \beta_0 + \beta_1 x_j$ i $var(Y_j) = \sigma^2, \ j = 1, \dots, n.$

Razlika između stvarne (empirijske) Y_j i očekivane vrednosti $E(Y_j)$ u osnovnom skupu predstavlja slučajnu grešku ϵ_j koja se dešava zbog biološke varijacije i grešaka merenja.

Ukoliko je Y_j neprekidna slučajna promenljiva, često se pretpostavlja da je njena raspodela normalna

$$\epsilon_j \sim \mathcal{N}(0, \sigma^2).$$

Iz ovoga dobijamo da za određeno x_j vrednosti Y_j takođe imaju normalnu raspodelu

$$Y_j \sim \mathcal{N}(\beta_0 + \beta_1 x_j, \sigma^2).$$

Dakle, model govori da važi simetrija: jednake su šanse da će vrednost Y_j biti iznad ili ispod očekivanja $\beta_0 + \beta_1 x_j$.

Važna pretpostavka linearne regresije je da su slučajne promenljive Y_j ili ekvivalentno promenljive ϵ_j, nezavisne. To znači da je način na koji Y_j u x_j uzima svoju vrednost nepovezan sa načinom na koji posmatranje Y_k u x_k uzima svoju vrednost. Ponašanje svakog posmatranja jedinke je nepovezano sa posmatranjima drugih.

Neka važi da je $Y_j \sim \mathcal{N}(\mu_j, \sigma^2)$, odnosno da su očekivanja različita, a varijanse iste i da su Y_j međusobno nezavisne promenljive. Zadatak regresione analize je da odredi liniju koja se najbolje prilagođava stvarnim vrednostima promenljive Y_j, i dobija se primenom metode najmanjih kvadrata. Metoda najmanjih kvadrata se sastoji u tome da se oceni očekivana vrednost za Y_j, tako da se minimizira suma kvadrata odstupanja $\sum_{j=1}^{n}(Y_j - \mu_j)^2$ ili ekvivalentno da minimizira

$$\sum_{j=1}^{n} \frac{(Y_j - \mu_j)^2}{\sigma^2}, \tag{1.6}$$

pri čemu je σ^2 konstantno. Odstupanja koja odgovaraju svakom posmatranju su sumirana, tako da svako odstupanje doprinosi (1.6) na sopstveni način i nije u vezi sa doprinosima drugih.

Ukoliko važi $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$, to jest da su očekivanja različita, kao i da su varijanse različite i da su Y_j međusobno nezavisne, za svako x_j, minimiziraćemo sledeću sumu kvadrata odstupanja

$$\sum_{j=1}^{n} \frac{(Y_j - \mu_j)^2}{\sigma_j^2}. \tag{1.7}$$

Kada je broj ponovljenih merenja mali, regresiona linija se zasniva na samo nekoliko vremenskih tačaka, a to procenu čini nepouzdanom.

U slučaju analize uticaja dve ili više objašnjavajućih promenljivih na zavisnu promenljivu govorimo o višestrukoj regresiji. Nezavisnu promenljivu često nazivamo faktor, kovarijata, dok njene vrednosti nazivamo tretmani ili nivoi faktora. Dakle, ispituje se uticaj faktora na zavisnu promenljivu. Napomenimo da bez obzira na to da li je reč o modelu proste ili višestruke regresije, model uvek sadrži samo jednu zavisnu promenljivu.

Opšti model je

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon,$$

gde su Y i ϵ slučajne promenljive, a promenljive $x_1, \dots, x_k$ su neslučajne promenljive, a β_i su nepoznati parametri. β_0 predstavlja odsečak ili konstantu, a $\beta_1, \beta_2, \dots, \beta_k$ su koeficijenti nagiba (nazivaju se i parcijalni regresioni koeficijenti). Ukoliko se vrednost jedne od nezavisnih promenljivih, na primer x_k, poveća za jednu jedinicu, tada se vrednost očekivanja poveća za β_k. Pozitivna vrednost koeficijenta β_i u regresionom modelu ukazuje na pozitivnu zavisnost promenljive Y_j od x_{ji}, dok negativna vrednost koeficijenta β_i ukazuje na negativnu zavisnost.

Da bi se ocenili parametri, potrebno je posmatrati uzorak. Za n različitih vrednosti $x_{j1}, \dots, x_{jk}$, $j = 1, \dots, n$ nezavisnih promenljivih mere se vrednosti zavisne promenljive: $Y_1, \dots, Y_n$. Na primer, merimo težinu n osoba nakon određenog režima vežbanja, a takođe merimo i x_1 – početnu težinu, x_2 – godine, x_3 – visinu, x_4 – tip ishrane i x_5 - nivo stresa. Cilj nam je napraviti statistički model koji će predstaviti težinu kao funkciju od nezavisnih promenljivih. Cilj ovog istraživanja bi mogao biti da se napravi model koji će opisati kako određen režim vežbanja utiče na težinu osobe sa određenim karakteristikama (početna težina,

godine, visina, tip iskrane i nivo stresa) i shodno tome da se razvije odgovarajući program vežbi.

Dobija se uzorački model

$$Y_j = \beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk} + \epsilon_j, \qquad \mu_j = \beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk},$$

$j = 1, \dots, n$ Ovde je ϵ_j slučajno odstupanje tako da važi $E(\epsilon_j) = 0$ i $var(\epsilon_j) = \sigma^2$, x_{jk} su uslovi u kojima su rađena posmatranja (na primer težina, godine), Deo $\beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk}$ je deterministički deo, a ϵ_j je stohastički deo.

Uobičajene pretpostavke su da važi

$$\epsilon_j \sim \mathcal{N}(0, \sigma^2) \text{ ili ekvivalentno } Y_j \sim \mathcal{N}(\mu_j, \sigma^2)$$

i da su Y_j međusobno nezavisne.

Često se model piše u matričnom obliku

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.8}$$

gde je

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}_{n\times p}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}.$$

$\boldsymbol{\epsilon}$ je vektor ukupnih grešaka i važi $p = k + 1$. Na osnovu reprezentacije kao što je (1.8) možemo govoriti o opštijem modelu za longitudinalne podatke, što ćemo i videti kasnije.

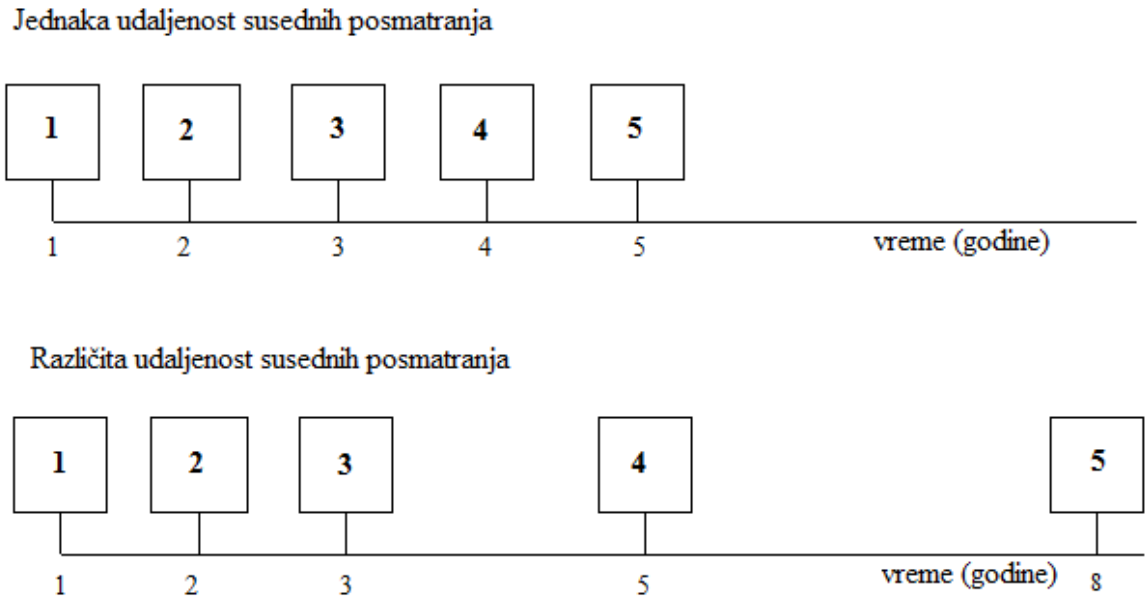
2. Uvod u modelovanje longitudinalnim podacima

Uvodimo osnovni statistički model za longitudinalne podatke. Drugi modeli se najčešće dobijaju modifikacijama ovog modela, tako što se uključuju određene pretpostavke o izvorima varijacija i obliku vektora očekivanih vrednosti.

U longitudinalnoj analizi razlikujemo balansirane i nebalansirane podatke. Podaci su balansirani kada su merenja svih posmatranih jedinki ponovljena u *n* istih vremenskih trenutaka. U suprotnom su podaci nebalansirani. Kod balansiranih podataka razlikujemo slučaj kada je razmak između posmatranja ravnomeran (jednaka udaljenost susednih posmatranja) i kada se razmaci između posmatranja razlikuju (različita udaljenost susednih posmatranja).

Prvo ćemo razmotriti modele kada su podaci balansirani, a kasnije ćemo kroz primere videti šta se dešava kada podaci nisu balansirani. Na Slici 1. vidimo različite mogućnosti međusobne udaljenosti balansiranih podataka u odnosu na vreme.

Slika 1. Posmatranja balansiranog skupa podataka.



2.1. Osnovni statistički model

Ponovimo da je u osnovi analize longitudinalnih podataka posmatranje iste reakcije više puta tokom vremena (ili u odnosu na neki drugi uslov) svake individualne jedinke. U najjednostavnijem slučaju, jedinke mogu biti slučajni uzorci iz jedne populacije. Međutim, jedinke mogu poticati i iz različitih populacija; mogu im biti dodeljeni različiti tretmani ili mogu biti različitih tipova (na primer: muškarci i žene). U nekim slučajevima mogu se evidentirati i dodatne informacije o karakteristikama pojedinačne jedinke, poput dobi, visine i drugo.

Pretpostavimo da obeležje od interesa svakog subjekta beležimo n puta $t_1 < t_2 < \cdots < t_n$, pri čemu su jedinke iz iste populacije. Svakom $t_j,\ j = 1, \ldots, n$, odgovara slučajna promenljiva $Y_j,\ j = 1, \ldots, n$, sa funkcijom raspodele koja objašnjava način na koji rezultati svih jedinki populacije uzimaju svoje vrednosti u vremenu t_j. Vrednosti koje te jedinke uzimaju u bilo kom vremenu t_j mogu da variraju zbog različitih izvora varijacije, koji su ranije obrazloženi.

Elementi slučajanog vektora

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \tag{2.1}$$

su promenljive raspoređene prema tekućem vremenskom poretku.

$\boldsymbol{Y}$ (2.1) ima multivarijantnu raspodelu verovatnoća koja predstavlja način na koji svi rezultati jedinki populacije uzimaju svoje vrednosti u vremenima $t_1, t_2, \ldots, t_n$. Ova raspodela ima očekivanu vrednost $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ sa elemetima $\mu_j = E(Y_j),\ j = 1, \ldots, n$ i kovarijansnom matricom $\boldsymbol{\Sigma}$. Koristićemo indeks $j = 1, \ldots, n$, za nabrajanje vremenskih jedinica.

Za slučajan uzorak od m jedinki datog obeležja od interesa, imamo m slučajnih vektora dimenzije $(n \times 1)$.

$$\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m,$$

gde koristimo indeks $i = 1, \ldots, m$, za nabrajanje jedinki. Svaka od njih je rezultat n merenja tokom vremena na jedinkama, tako da svaka ima karakteristike (npr. multivarijantne raspodele verovatnoća) identične sa $\boldsymbol{Y}$. Vektore $\boldsymbol{Y}_i, i = 1, \ldots, m,$ možemo predstaviti na sledeći način

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{bmatrix},$$

pri čemu važi $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}$ i $var(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}$, s obzirom da promenljive $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$, imaju istu raspodelu.

Dakle, imamo matricu dimenzije $(n \times m)$

$$\begin{bmatrix} Y_{11} & Y_{21} & & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & & \vdots \\ Y_{n1} & Y_{n2} & & Y_{nm} \end{bmatrix}.$$

Prirodno je da su komponente $Y_{ij}, j = 1, \ldots, n,$ korelisane. Rezultati posmatranja na jednom subjektu teže da budu sličniji u odnosu na rezultate drugih jedinki (na primer, ako je izmerena težina jedne jedinke znatno manja od težina drugih jedinki u nekom vremenskom trenutku, za očekivati je da će i u većini drugih posmatranja težina te jedinke biti manja u odnosu na druge jedinke). Zato je realno očekivati da je

$$cov(Y_{ij}, Y_{ik}) \neq 0 \text{ za svako } j \neq k = 1, \ldots, n.$$

pa $\boldsymbol{\Sigma}$ neće biti dijagonalna matrica.

Ukoliko svako $\boldsymbol{Y}_i$ odgovara različitim jedinkama i jedinke nisu ni u kakvoj vezi (na primer, različite biljke pod drugačijim tretmanima), tada je realno pretpostaviti da rezultati bilo kog posmatranja jedinke i nisu u vezi sa mogućim rezultatima posmatranja jedinke

$l \neq i$. Sledi da su vektori $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_m$ su međusobno nezavisni. Dakle, ukoliko je Y_{ij} rezultat jedinke i i Y_{lk} je rezultat jedinke l, $cov(Y_{ij}, Y_{lk}) = 0$ za svako $j = k$ (isti vremenski trenutak, ali različite jedinke).

- Osnovni statistički model

Spajajući sve ovo zajedno, dobijamo m međusobno nezavisnih vektora $\boldsymbol{Y}_i, i = 1, \dots, m$, sa očekivanjem $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}$, $var(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}$.

Analogno kao u univarijantnom slučaju u (1.1), dobijamo sledeći statistički model:

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \;\; E(\boldsymbol{\epsilon}_i) = \boldsymbol{0}, \;\; var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}, \tag{2.2}$$

gde je $\boldsymbol{\epsilon}_i$ slučajni vektor odstupanja, tako da su $\boldsymbol{\epsilon}_i$, $i = 1, \dots, m$, međusobno nezavisne promenljive i važi da je $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{in})'$, gde svako $\epsilon_{ij}, j = 1, \dots, n$, predstavlja odstupanje Y_{ij} od svog očekivanja μ_j zbog svih izvora varijacije. Važi i to da su $\epsilon_{ij}, j = 1, \dots, n$ u korelaciji, ali $\boldsymbol{\epsilon}_i$ su međusobno nezavisni za sve i.

Možemo primetiti da se $\boldsymbol{Y}_i$ sastoji iz dve komponente:

- $\boldsymbol{\mu}$ je *deterministički* deo modela koji opisuje očekivane rezultate tokom vremena. Ova reprezentacija zahteva da je dužina svakog vektora podataka $\boldsymbol{Y}_i$ ista i da je n.
- $\boldsymbol{\epsilon}_i$ je *slučajni* deo modela koji opisuje koliko $\boldsymbol{Y}_i$ odstupa od svog očekivanja.

Vrednosti promenljive $\boldsymbol{Y}_i$ se mogu meriti, dok se vrednosti promenljive $\boldsymbol{\epsilon}_i$, koja se naziva *greška* ili *rezidual*, ne mogu meriti. Promenljiva $\boldsymbol{\epsilon}_i$ u sebi sadrži sve ostale promenljive koje utiču na vrednost promenljive $\boldsymbol{Y}_i$, ali nisu obuhvaćene posmatranjem. Sve ove promenljive dovodiće do slučajnih odstupanja od predviđenog modela. Iako se vrednosti promenljive $\boldsymbol{\epsilon}_i$ ne mogu meriti, često se kao deo modela daju pretpostavke o njihovoj raspodeli.

Ukoliko su rezultati promenljive koja se posmatra neprekidnog tipa, često se koristi pretpostavka da Y_{ij} i ϵ_{ij} imaju normalnu raspodelu. Dakle uz (2.2) stavlja se uslov,

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \;\text{ i }\; \boldsymbol{Y}_i \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad i = 1, \dots, m,$$

gde su $\boldsymbol{Y}_i$ međusobno nezavisne.

- Merenja na više gupa

Pretpostavimo da subjekti mogu biti slučajno izabrani iz q različitih populacija (na primer: $q = 2$ ukoliko posmatramo dečake i devojčice). Dakle, imamo m nezavisnih slučajnih vektora $\boldsymbol{Y}_i$, gde ukoliko $\boldsymbol{Y}_i$ odgovara jedinki iz grupe $l, l = 1, \dots q$, tada $\boldsymbol{Y}_i$ ima multivarijantnu raspodelu verovatnoća sa

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_{\boldsymbol{l}}, \;\; var(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_l, \;\; i = 1, \dots, l$$

Vidimo da svaka grupa može imati različite vektore očekivanih vrednosti i kovarijansne matrice. Ekvivalentno, dobijamo

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_l + \boldsymbol{\epsilon}_i,$$

$E(\boldsymbol{\epsilon}_i) = 0, \;\; var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_l,$ za i iz grupe $l = 1, \dots, q$. Možemo pretpostaviti $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_l)$ za jedinku iz grupe l, pa je

$$\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \;\; i = 1, \dots, l$$

za i iz grupe l.

Možemo pretpostaviti i da se svi izvori varijacije ponašaju slično u svakoj populaciji, pa sledi da je $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}$, zajednička kovarijansna matrica za sve populacije.

Model tada postaje

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_l + \boldsymbol{\epsilon}_i, \tag{2.3}$$

gde važi da je $E(\boldsymbol{\epsilon}_i) = 0,\ var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma},$ za *i* iz grupe *l = 1, ..., q* i $\boldsymbol{\Sigma}$ je kovarijansna matrica zajednička za sve grupe.

Primetimo da, u ovom slučaju, iako je $\boldsymbol{\Sigma}$ zajednička za sve populacije, dijagonalni elementi matrice $\boldsymbol{\Sigma}$ mogu biti različiti za različite $j = 1, \dots, n$, pa varijanse mogu biti različite za različita vremena, ali u svakom pojedinačnom trenutku, varijansa je ista za sve grupe. Slično, kovarijanse između j - tog i k - tog elementa $\boldsymbol{Y}_i$ mogu biti različite za različite izbore j i k, ali za određen par (j, k), kovarijanse su iste za sve grupe.

2.1.1. Merenja ponovljena dva puta

Najjednostavniji oblik longitudinalne analize je kada se obeležje od interesa meri dva puta u vremenu. Tada se pitamo da li postoji razlika u rezultatima promenljive $\boldsymbol{Y}$ u vremenima t_1 i t_2. Da bismo rešili postavljeno pitanje od interesa, možemo koristiti t – test parova. Testira se hipoteza da li je razlika u očekivanju Y_{t1} i Y_{t2} jednaka nuli. Kako statistički test koristi pojedinačne razlike, on uzima u obzir činjenicu da su zapažanja u okviru jednog subjekta zavisna jedna od drugog. t - test statistika parova je

$$t = \frac{\bar{\boldsymbol{d}}}{\left(\frac{\sigma_d}{\sqrt{\boldsymbol{m}}}\right)}$$

gde je t test statistika, $\bar{d}$ prosek razlika, σ_d standardna devijacija razlika i m broj subjekata. Ova test statistika ima t – raspodelu sa $m - 1$ stepenom slobode.

Pretpostavke za korišćenje t – testa su da su zapažanja različitih subjekata nezavisna i da razlike između dva merenja imaju približno normalnu raspodelu. U istraživanjima kada je broj subjekata velik (preko 25), t – test parova se koristi bez ikakvih problema, dok sa manjim brojem subjekata, pretpostavka normalne raspodele je vrlo važna.

Ukoliko pretpostavke t – testa nisu zadovoljene, moguće je koristiti neparametarski ekvivalent – Wilcoxonov test sume rangova. Ovaj test je zasnovan na dodeljivanju rangova individualnim razlikama i ne zahteva nikakve pretpostavke o distribuciji rezultata.

Kada se ponavljanja vrše više od dva puta situacija longitudinalne analize postaje kompleksnija i ne može se koristiti t – test. Pitanje od interesa da li se zavisna promenljiva **Y** menja tokom vremena možemo rešiti analizom varijanse ponovljenih merenja. Osnovna ideja statističkih tehnika je slična kao kod t – testa.

Kada su merenja ponovljena dva puta analiza ne daje toliko sigurne odgovore za tok situacije tokom vremena. Najčešće se vrši analiza longitudinalnih podataka kada su merenja ponovljena više od dva puta.

2.2. Izvori varijacije i korelacije longitudinalnih podataka

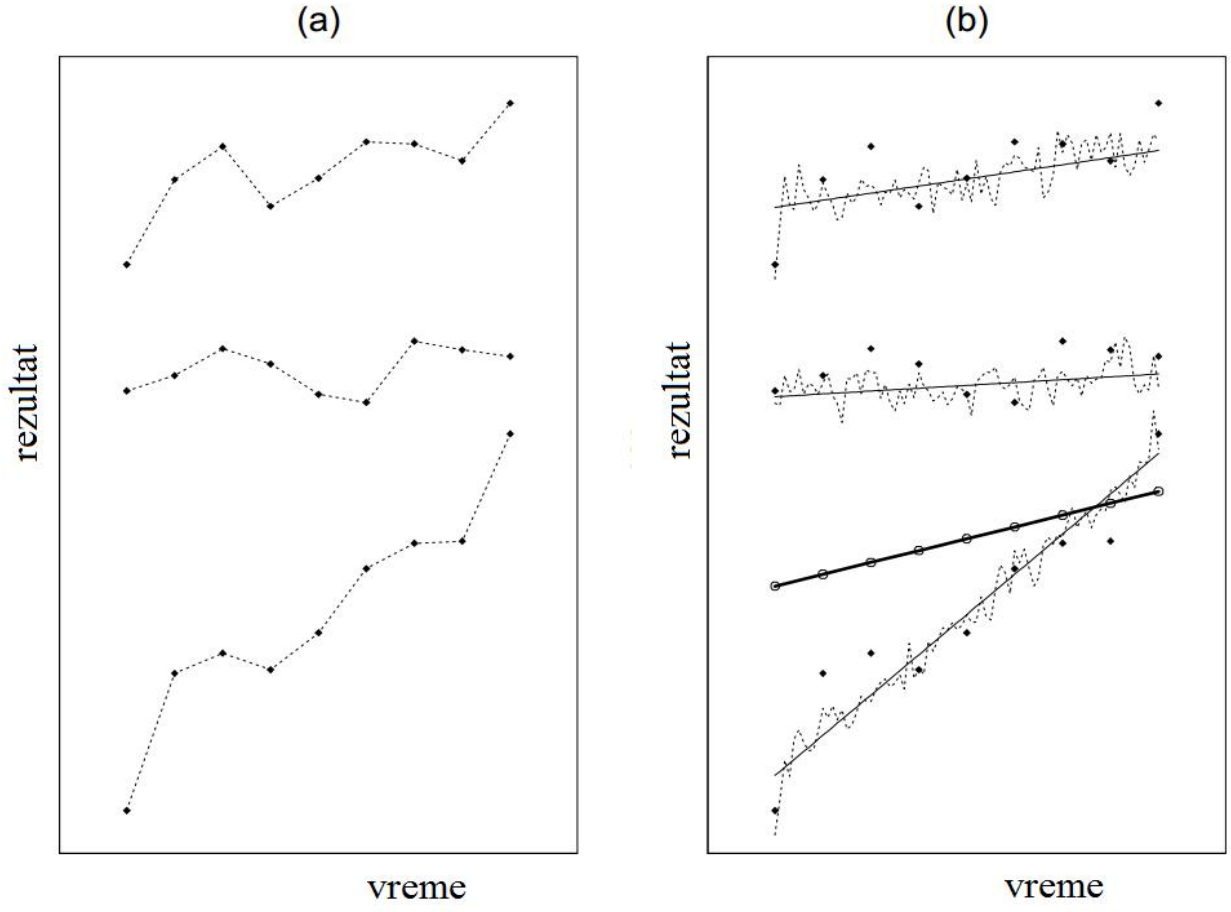
Mogući izvori varijacija za longitudinalne podatke se najčešće dele na dva tipa:

1. Slučajna varijacija u populaciji zbog biološke varijacije: različiti rezultati iz razloga što jedinke biološki variraju (nisu sve identične). Ovo se naziva ***slučajnom varijacijom među jedinkama.*** (among units)
2. Slučajna varijacija zbog fluktuacije pri posmatranju jedne jedinke i grešaka merenja, što se naziva ***slučajna varijacija unutar jedinke***. (within unit)

Ovi izvori će biti značajni za određivanje prirode kovarijansne matrice vektora podataka. Korisno je razmotriti način na koji se longitudinalni podaci mogu pojavljivati. Razmotrimo slučaj jednog obeležja i modela

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i\,, \quad i = 1, \dots, m.$$

Slika 2. a) Longitudinalni podaci za $m = 3$ jedinke u $n = 9$ perioda označeni su tačkama.
b) Mogući izvori varijacije. Otvoreni kružići spojeni debelom crnom linijom predstavljaju očekivanja μ_j, $j = 1, \dots, 9$ za populacije svih posmatranja u svih 9 perioda. Tanke prave linije predstavljaju "trend" za svaku jedinku. Tačkaste linije predstavljaju kretanje obrasca grešaka svake jedinke tokom vremena koje fluktuiraju u odnosu na trend. Crne tačke predstavljaju rezultate ovih merenja koji su podložni greškama.



Možemo primetiti sledeće: Slika 2 a) predstavlja stvarne rezultate tri jedinke, gde su u merenja uključeni efekti svih izvora varijacije. Slika 2 b) je zamišljena reprezentacija mogućih karakteristika u ovom slučaju. Otvoreni kružići na debeloj liniji predstavljaju elemente u vektoru $\boldsymbol{\mu}$ za svaki od devet perioda. Na primer, prvi kružić skroz levo predstavlja očekivanje μ_1 zavisne promenljive u trenutku t_1, nastalo usrednjavanjem svih odstupanja ϵ_{i1} nastalih zbog varijacija unutar i među subjektima i. Očekivana vrednost tokom vremena, u ovom primeru, leži na pravoj liniji, ali to generalno ne mora da važi. Crne (podebljane) tačke predstavljaju rezultate za pojedinačne jedinke. Ako posmatramo samo prvu vremensku tačku, primetimo da rezultati jedinke i variraju oko μ_1.

- Za svaku individuu, možemo uočiti "trend", označenim tankim pravim linijama (trend ne mora biti prava linija). "Trend" predstavlja ponašanje jedinki u populaciji i razlikuje se između jedinki**.** Položaj trenda neke jedinke u svakom momentu određuje da li se jedinka nalazi "iznad" ili "ispod" u odnosu na zajedničku sredinu $\boldsymbol{\mu}$**.**
- Tačkaste linije su fluktuacije oko trenda jedinke, predstavljajući varijaciju kako rezultati jedinke mogu evoluirati. Biološka fluktuacija oko trenda je rezultat procesa unutar posmatrane jedinke.
- Dalje, rezultati jedinke (crne podebljane tačke) ne leže tačno na tačkastim linijama, već variraju oko njih. Ovo se dešava zbog grešaka merenja, a takve greške se dešavaju unutar same jedinke.

Raščlanimo dalje elemente modela (2.2), tako da *j*- ti element $\boldsymbol{Y}_i$, Y_{ij}, predstavlja sumu nekoliko komponenti:

$$Y_{ij} = \mu_j + \epsilon_{ij} = \mu_j + b_{ij} + e_{ij} = \mu_j + b_{ij} + e_{1ij} + e_{2ij}, \qquad (2.4)$$

gde je $E(b_{ij}) = 0$, $E(e_{1ij}) = 0$ i $E(e_{2ij}) = 0$.

- b_{ij} je odstupanje koja predstavlja varijaciju među jedinkama u vremenu t_j (zbog zbog biološke varijacije).
- e_{1ij} predstavlja odstupanje zbog fluktuacija jedinki oko trenda (varijacija unutar jedinki).
- e_{2ij} je odstupanje zbog grešaka merenja (unutar jedinki).
- Zbir $e_{ij} = e_{1ij} + e_{2ij}$ označava ukupno odstupanje unutar jedinki.
- Zbir $\epsilon_{ij} = b_{ij} + e_{1ij} + e_{2ij}$ predstavlja ukupno odstupanje od μ_j zbog svih uzroka odstupanja.

  Možemo pisati

$$\boldsymbol{\epsilon}_i = \boldsymbol{b}_i + \boldsymbol{e}_i = \boldsymbol{b}_i + \boldsymbol{e}_{1i} + \boldsymbol{e}_{2i},$$

gde se naglašava da $\boldsymbol{\epsilon}_i$ uključuje komponente varijacije i unutar jedinki i među jedinkama.

Navedimo neke pretpostavke o varijaciji među i unutar jedinki kao i to kako proističe korelacija među Y_{ij} (ekvivalentno, među ϵ_{ij}).

- b_{ij} određuje trend svojstven jednoj jedinki u smislu da $\mu_j + b_{ij}$ predstavljaju poziciju trenda za jedinku i u trenutku j . Y_{ij} prema tome teže da budu blizu ovog trenda tokom vremena (j) za jedinku i.

  Prema tome, očekujemo da su elementi od $\boldsymbol{\epsilon}_i$ (pa i za $\boldsymbol{Y}_i$) u korelaciji zbog činjenice da dele ovaj zajednički trend. Korelacija nastala na ovaj način je korelacija između jedinki.
- Kako su e_{1ij} odstupanja zbog procesa fluktuacije, prirodno je da mislimo da postoji korelacija e_{1ij} u odnosu na j. Ukoliko je proces "visok" u odnosu na svojstven trend

u trenutku t_j (e_{1ij} je pozitivno), može se očekivati da je "visok" i u t_j, koje se nalazi blizu t_j. Očekujemo da su elementi $\boldsymbol{\epsilon}_i$, pa i $\boldsymbol{Y}_i$, u korelaciji kao posledica takve fluktuacije (jer su elementi $\boldsymbol{e}_{1i}$ korelisani). Primetimo da ako se fluktuacija dešava u kratkom vremenskom periodu u odnosu na momenat t_j, to što je proces "visok" u momentu t_j može imati malu ili ne imati nikakvu vezu sa tim da li je visok u susednim vremenima. U ovom slučaju, možemo verovati da je korelacija među subjektima zanemarljiva. Kao što se vidi, ovo je opravdana pretpostavka, često opravdano ukazujući da su t_j udaljeni u vremenu.

- Opšti obrazac korelacije za $\boldsymbol{\epsilon}_i$ (pa i $\boldsymbol{Y}_i$) je rezultat kombinovanih efekata ova dva izvora (unutar subjekta i među subjektima).
- Svaki put kad se koriste uređaji za merenje, oni teže da naprave "slučajne" greške, te je stoga razumno pretpostaviti da su e_{2ij} nezavisni za sve j. Dakle, ne očekujemo doprinos opštem obrascu korelacije.

Zaključak je da moramo razmotriti varijanse $\boldsymbol{b}_{ij}, \boldsymbol{e}_{1ij}$ i $\boldsymbol{e}_{2ij}$.

2.3. Ispitivanje srednje vrednosti i kovarijanse

Skup efekata svih uzroka varijacije određuje oblik kovarijansne matrice za $\boldsymbol{\epsilon}_i$, pa zbog toga i za $\boldsymbol{Y}_i$. Prilikom ocenjivanja parametara koji nas interesuju, važno je uzeti u obzir da su posmatranja korelisana i možda imaju različite varijanse. Zbog toga je tačna reprezentacija *var*($\boldsymbol{\epsilon}_i$) od velikog značaja. Prvi korak u analizi najčešće je ispitavanje podataka radi sagledavanja najverovatnijeg oblika kovarijansne matrice.

Razmotrimo uzorački model kada se posmatrano obeležje $\boldsymbol{Y}$, koje se posmatra u *n* vremenskih momenata na *m* jedinki

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{bmatrix} = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \tag{2.5}$$

$$E(\boldsymbol{\epsilon}_i) = \mathbf{0},\ var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma},\ i = 1, \dots, m.$$

Na osnovu ovog uzorka, želeli bismo da steknemo uvid o najverovatnijem obliku $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$. Dijagonalni elementi matrice $\boldsymbol{\Sigma}$ su varijanse rezultata ponovljenih merenja, a vandijagonalni elementi predstavljaju kovarijansu između tih rezultata u odnosu na različita ponovljena merenja.

Prirodna ocena za očekivanje $\boldsymbol{\mu}_j$ u j – tom periodu je aritmetička sredina uzorka (srednja vrednost uzorka):

$$\bar{Y}_{.j} = m^{-1} \sum_{i=1}^{m} Y_{ij},$$

gde tačka označava prosek u odnosu na prvi indeks *i* (po subjektima). Srednja vrednost uzorka može da se izračuna za svaku vremensku tačku, $j = 1, \dots, n$, pa je ocena za vektor očekivanja $\boldsymbol{\mu}$, vektor srednjih vrednosti uzorka čiji su elementi $\bar{Y}_{.j}$, dat sa

$$\bar{\boldsymbol{Y}} = m^{-1} \sum_{i=1}^{m} \boldsymbol{Y}_i = \begin{bmatrix} \bar{Y}_{.1} \\ \vdots \\ \bar{Y}_{.n} \end{bmatrix}.$$

Slučajan vektor $\overline{\boldsymbol{Y}}$ je nepristrasna ocena za $\boldsymbol{\mu}$:

$$E(\overline{\boldsymbol{Y}}) = \boldsymbol{\mu}.$$

Sticanje uvida o obliku matrice $\boldsymbol{\Sigma} = E[(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})']$ može se dobiti pomoću nepristrasne ocene za $\boldsymbol{\Sigma}$ i ocene za korelacionu matricu. Dijagonalni elementi $\boldsymbol{\Sigma}$ su varijanse σ_j^2 promenljivih $\boldsymbol{Y}_j$ u svakom od perioda $j = 1, \dots, n.$. Za m jedinki, prirodna ocena za σ_j^2 je varijansa uzorka u periodu j,

$$S_j^2 = m^{-1} \sum_{i=1}^{m} (Y_{ij} - \bar{Y}_{.j})^2,$$

i može se pokazati da je to nepristrasna ocena za σ_j^2.

Vandijagonalni elementi matrice $\boldsymbol{\Sigma}$ su kovarijanse

$$\sigma_{jk} = E\big[\big(Y_j - \mu_j\big)(Y_k - \mu_k)\big].$$

pa je njihova ocena

$$S_{jk} = (m-1)^{-1} \sum_{i=1}^{m} \big(Y_{ij} - \bar{Y}_{.j}\big)(Y_{ik} - \bar{Y}_{.k}).$$

koja je takođe nepristrasna.

Ocena za $\boldsymbol{\Sigma}$ je uzoračka korelaciona matrica:

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} S_1^2 & S_{12} & \dots & S_{1n} \\ S_{21} & S_2^2 & \dots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n1} & S_{n2} & \dots & S_n^2 \end{bmatrix},$$

ili u matričnom zapisu

$$\widehat{\boldsymbol{\Sigma}} = (m-1)^{-1} \sum_{i=1}^{m} (\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})'.$$

Suma $\sum_{i=1}^{m} (\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})'$ se zove matrica sume kvadrata i uzajamnih proizvoda[1], jer su njeni elementi sume kvadrata odstupanja i uzajamni proizvodi odstupanja od uzoračke sredine.

Kovarijansna matrica uzorka se može koristiti kao ocena kovarijansne matrice. Iako dijagonalni elementi mogu obezbediti informaciju o stvarnoj vrednosti varijanse u svakom vremenskom trenutku, vandijagonalni elementi su teški za interpretaciju.

Ako je $\widehat{\boldsymbol{\Sigma}}$ ocena kovarijansne matrice $\boldsymbol{\Sigma}$ sa elementima $\widehat{\Sigma}_{jk}, k = 1, \dots, n$, tada prirodna ocena za korelacionu matricu $\boldsymbol{\Gamma}$ je $(n \times n)$ matrica $\widehat{\boldsymbol{\Gamma}}$ sa jedinicama na dijagonali i (j, k) vandijagonalnih elemenata datih izrazom:

$$\frac{\widehat{\Sigma}_{jk}}{\sqrt{\widehat{\Sigma}_{jj}\widehat{\Sigma}_{kk}}}.$$

U slučaju jedne populacije, gde je $\widehat{\boldsymbol{\Sigma}}$ kovarijansna matrica uzorka, vandijagonalni elementi su

$$\frac{S_{jk}}{S_j S_k},$$

a koji su ocene korelacija

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}.$$

[1] Često se koristi skraćenica SS&CP; originalno Sums of Squares and Cross Product matrix

U ovom slučaju, ocena $\hat{\mathbf{\Gamma}}$ se zove korelaciona matrica uzorka. Izraz $\frac{s_{jk}}{s_j s_k}$ se zove uzorački koeficijent korelacije između posmatranja u vremenima j i k, jer je ocena za odgovarajući koeficijent korelacioje ρ_{jk}.

2.4. Kovarijansni i korelacioni modeli

Postoje različiti pristupi za predstavljanje longitudinalnih podataka pomoću statističkog modela. Jedan od načina da se naprave razlike između tih modela i metoda jesu pretpostavke o strukturi kovarijansne matrice vektora podataka. Predstavićemo često korišćene modele za različite strukture kovarijansne matrice longitudinalnih podataka i to za balansirane i za nebalansirane podatke. Svaki kovarijansni model ima odgovarajući korelacioni model.

2.4.1. Modeli za balansirane podatke

Posmatramo longitudinalne podatke i pretpostavimo da je kovarijansna matrica $var(\boldsymbol{\epsilon}_i) = \mathbf{\Sigma}$ ista za sve i, gde $\mathbf{\Sigma}$ može da ima različite oblike. Predstavimo neke od njih:

1. Nestruktuiran kovarijansni model

U nekim slučajevima se ne uočavaju sistematska pravila u ponašanju varijansi i kovarijansi. Tada kažemo da imamo nestruktuiran kovarijansni model.

U nestruktuiranoj kovarijansnoj matrici moguće je da postoji n različitih varijansi, jedna za svaki trenutak u vremenu i $n(n-1)/2$ različitih elemenata van dijaginale koji predstavljaju različite kovarijanse za svaki par vremenskih trenutaka. Matrica ukupno ima $n + n(n-1)/2 = n(n+1)/2$ varijansi i kovarijansi. (Kako je kovarijansna matrica simetrična, za vandijagonalne elemente važi $(j,k) = (k,j)$, tako da je potrebno računati svaku kovarijansu jednom).

Dakle, u nestruktuiranom kovarijansnom modelu postoji mnogo parametara koji opisuju varijaciju i koje treba proceniti, naročito ako je n veliko. Na primer, ako je $n = 5$, što i nije tako veliko, imamo $(5 \times 6)/2 = 15$ parametara. Ako postoji q različitih grupa, svaka sa različitom kovarijansnom matricom, biće q puta više varijansi i kovarijansi. Ako kovarijanse imaju neko sistematsko ponašanje, zadržavanje pretpostavke o nestrukuiranosti bi rezultirao ocenjivanjem više parametara nego što je to potrebno.

U nekim slučajevima je lakše da se prvo posmatra korelacioni model, a zatim odgovarajući kovarijansni model.

2. Složeni simetrični kovarijansni modeli

U nekim slučajevima posmatra se kovarijansni model u kome je korelacija između posmatranja u bilo kojim vremenima t_j i t_k ista, a varijanse za različita vremena mogu biti različite. Pretpostavimo da je ρ parametar koji predstavlja zajedničku korelaciju za bilo koje dve tačke u vremenu. Na primer, za $n = 5$, korelaciona matrica je

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}.$$

Ista struktura je za bilo koje n. Ovde je $-1 < \rho < 1$.

Predstavimo dva popularna kovarijansna modela sa ovom korelacionom matricom .

- Ako su σ_j^2 i σ_k^2 varijanse za vremena t_j i t_k (mogu biti različite u različitim vremenima) i σ_{jk} odgovarajuća kovarijansa, tada je

  $$\rho = \frac{\sigma_{jk}}{\sigma_j \sigma_k}, \text{ pa je } \sigma_{jk} = \rho \sigma_j \sigma_k.$$

  Na primer, za $n = 3$, kovarijansna matrica je:

  $$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix},$$

  što važi za svako n.
- Ako su varijanse iste za svaki vremenski trenutak $\sigma_j^2 = \sigma^2$ za $j = 1, \dots, n$. Tada je

  $$\rho = \frac{\sigma_{jk}}{\sigma^2}, \text{ pa je } \sigma_{jk} = \rho\sigma^2.$$

  Za $n = 3$ imamo:

  $$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} = \sigma^2 \boldsymbol{\Gamma}.$$

  Naravno, postoje i alternativni modeli. $var(\boldsymbol{\epsilon}_i)$ može biti drugačija za različite grupe.

3. „Jedna zavisna“

Posmatranja vršena u bližim vremenskim momentima mogu biti više povezana od onih koja su vršena u udaljenijim vremenskim momentima. Ako su vremena posmatranja raspoređena tako da je vreme između dva nesusedna posmatranja prilično dugo, možemo pretpostaviti da će korelacija između dva uzastopna posmatranja (koja su susedna u vremenu) biti najveća. To jest, korelacije se javljaju u uzastopnim vremenskim trenucima. U odnosu na opseg ove korelacije, korelacija između posmatranja koja su razdvojena sa dva vremenska intervala može biti zanemarljiva. Na primer, za $n = 5$, korelaciona matrica koja to predstavlja je

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & 0 & 0 \\ 0 & \rho & 1 & \rho & 0 \\ 0 & 0 & \rho & 1 & \rho \\ 0 & 0 & 0 & \rho & 1 \end{bmatrix}.$$

Dakle, korelacija između susednih posmatranja je ista i jednaka ρ, $-1 \leq \rho \leq 1$. Ovaj model se često koristi kad su merenja podjednako raspoređena u vremenu (ponovljena merenja u uzrastima od 6, 8, 10 i 12 godina).

Dva popularna kovarijansna modela sa ovom korelacionom matricom su:

- Ako je $\sigma_j^2, j = 1, \dots, n$ varijansa u vremenu t_j, tada je

$$\rho = \frac{\sigma_{jk}}{\sigma_j \sigma_k}, \text{ pa je } \sigma_{jk} = \rho \sigma_j \sigma_k.$$

Na primer, za $n = 4$, kovarijansna matrica je:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 & 0 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & 0 \\ 0 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ 0 & 0 & \rho\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix},$$

što važi za svako n.

- Ako su varijanse iste za svaki vremenski trenutak $\sigma_j^2 = \sigma^2$ za $j = 1, \dots, n$. Tada je

$$\rho = \frac{\sigma_{jk}}{\sigma^2}, \text{ pa je } \sigma_{jk} = \rho\sigma^2.$$

$$\mathbf{\Sigma} = var(\boldsymbol{\epsilon}_i) = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & 0 & \dots & 0 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \rho\sigma^2 & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{\Gamma}.$$

Naravno, postoje i alternativni modeli. $var(\boldsymbol{\epsilon}_i)$ može biti drugačija za različite grupe. Ako radimo sa grupama, pretpostavka „jedna zavisna“ važi za svaku grupu, ali dozvoljava mogućnost da su varijansa σ^2 i korelacija ρ različite u svakoj grupi. Isto važi i za naredne modele koje ćemo razmatrati.

4. Autoregresivni model reda 1 (ravnomerno raspoređeni u vremenu)

Ovaj model kaže da korelacija opada kada se posmatranja udaljavaju jedna od drugih u vremenu. Ovaj model ima smisla samo ako su vremena posmatranja podjednako razmaknuta. Dakle, autoregresivni model reda 1 sa homogenom varijansom tokom vremena glasi:

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{bmatrix}.$$

- Ako je $\sigma_j^2, j = 1, \dots, n$ varijansa u vremenu t_j, tada je

$$\rho = \frac{\sigma_{jk}}{\sigma_j \sigma_k}, \text{ pa je } \sigma_{jk} = \rho \sigma_j \sigma_k.$$

Kovarijansna matrica je:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \dots & \rho^{n-1}\sigma_1\sigma_n \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \dots & \rho^{n-2}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{n-1}\sigma_1\sigma_n & \rho^{n-2}\sigma_2\sigma_n & \rho^{n-3}\sigma_3\sigma_n & \cdots & \sigma_n^2 \end{bmatrix},$$

što važi za svako n.

- Ako su varijanse iste za svaki vremenski trenutak σ^2, tada je

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \dots & \sigma^2\rho^{n-1} \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma^2\rho^{n-1} & \sigma^2\rho^{n-2} & \sigma^2\rho^{n-3} & \dots & 1\sigma^2 \end{bmatrix} = \sigma^2 \mathbf{\Gamma}.$$

5. Markovljev model (neravnomerno raspoređeni u vremenu)

Autoregresivni model reda 1 može biti generalizovan na vremena koja su neravnomerno raspoređena (na primer: 1, 2, 3, 4, 6, 9). Stepeni od ρ su uzeti za rastojanja u vremenu između posmatranja. Ako je

$$d_{jk} = |t_{ij} - t_{ik}|, \;\; j,k = 1, \dots, n,$$

onda je model za korelacionu matricu

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & \rho^{d_{12}} & \dots & \rho^{d_{1n}} \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{d_{n1}} & \rho^{d_{n2}} & \dots & 1 \end{bmatrix},$$

odnosno kovarijansna matirca je oblika:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2\rho^{d_{12}} & \dots & \sigma^2\rho^{d_{1n}} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma^2\rho^{d_{n1}} & \sigma^2\rho^{d_{n2}} & \dots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{\Gamma}.$$

2.4.2. Modeli za nebalansirane podatke

Za analizu nebalansiranih podatka potebno je postaviti druge pretpostavke o kovarijansnoj matrici. Pretpostavimo da smo u situaciji kad su sva vremena posmatranja ista za sve jedinke, ali za pojedine jedinke neka posmatranja nedostaju. Na primer, imamo vremena $(t_1, t_2, t_3, t_4) = (0, 1, 2, 3)$ i za jedinku i u vremenu t_3 rezultat posmatranja nije dostupan. Prema tome, vektor $\boldsymbol{Y}_i$ za ovu jedinku je dužine $n_i = 3$. Ovu situaciju možemo notaciono predstaviti na dva različita načina:

- Za i-tu jedinku ćemo pisati:

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{bmatrix} \text{ za posmatranja u vremenima } \begin{bmatrix} t_{i1} \\ t_{i2} \\ t_{i3} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}.$$

U ovoj notaciji za t_{ij}, j označava broj posmatranja u okviru jedinke, bez obzira na stvarne vrednosti za vremena. Postoje 3 različita vremenska momenta za ovu jedinku, pa je $j = 1,2,3$.

- Da bismo mogli modelirati kovarijansnu strukturu, razmotrimo drugačiji pristup. Neka nam sada j predstavlja vremena u kojima smo nameravali da posmatramo pojavu. Za jedinku i nedostaje vreme $j = 3$, pa ćemo ovo predstaviti na sledeći način:

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i4} \end{bmatrix} \text{ za vremena } \begin{bmatrix} t_1 \\ t_2 \\ t_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}.$$

Ovde ćemo razmatramo nekoliko kovarijansnih struktura i njihovo korišćenje kao moguće modele za $var(\boldsymbol{\epsilon}_i) = \mathbf{\Sigma}_i$. Modelovanjem $var(\boldsymbol{\epsilon}_i)$ nećemo zaista paraviti razliku između varijacija „među jedinkama“ i „u unutar jedinki“, već ćemo razmotriti modele za skup svih izvora zajedno.

Pretpostavimo da je $var(Y_{ij}) = \sigma^2$ za sve j. Na taj način želimo model za

$$\boldsymbol{\Sigma}_i = var(\boldsymbol{Y}_i) = \begin{bmatrix} \sigma^2 & cov(Y_{i1}, Y_{i2}) & cov(Y_{i1}, Y_{i4}) \\ cov(Y_{i2}, Y_{i1}) & \sigma^2 & cov(Y_{i2}, Y_{i4}) \\ cov(Y_{i4}, Y_{i1}) & cov(Y_{i4}, Y_{i2}) & \sigma^2 \end{bmatrix}.$$

1. Nestruktuiran kovarijanski model

Pod pretpostavkom nestruktuiranog modela, ova matrica postaje

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{24} \\ \sigma_{14} & \sigma_{24} & \sigma_4^2 \end{bmatrix}.$$

Svaka jedinka je posmatrana na sopstvenom skupu ponovljenih merenja, pa kovarijansni parametri nisu isti, a iz toga sledi da će matrica zavisiti od potpuno različitih vrednosti za svaku jedinku.

2. Složeni simetrični kovarijansni modeli

Pretpostavka složene simetrije bi bila predstavljena na isti način bez obzira na vrednost koja nedostaje. Posmatranja na bilo kojoj udaljenosti u vremenu imaju istu korelaciju. Pod ovom pretpostavkom, $\boldsymbol{\Sigma}_i$ bi bila (3×3) matrica:

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

3. „Jedna zavisna“

Kod modela „jedna zavisna“, gde su u korelaciji samo posmatranja susedna u vremenu, ova matrica postaje

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & 0 \\ \rho\sigma^2 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}.$$

4. Autoregresivni model reda 1 (ravnomerno raspoređeni u vremenu)

Ukoliko su vremenski momenti ponovljenih merenja jednako udaljeni možemo koristiti matricu ovog modela koja glasi:

$$\boldsymbol{\Sigma}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^3 \\ \rho & 1 & \rho^2 \\ \rho^3 & \rho^2 & 1 \end{bmatrix}.$$

Ovi primeri ilustruju najvažniju stvar – ako su sva merenja urađena u istim vremenskim momentima, pri čemu neka od njih nedostaju, kovarijansna matrica mora biti pažljivo konstruisana u skladu sa određenim vremenskim obrascima za svaku jedinku, koristeći odgovarajući kovarijansni model.

3. Univarijantna analiza varijanse sa ponovljenim merenjima

Jedna od osnovnih tehnika za analizu longitudinalnih podataka je analiza varijanse, (ANOVA) koju čine razne tehnike za identifikaciju i merenje raznih izvora varijabiliteta (rasipanja) unutar nekog skupa podataka. Razlikujemo sledeće podele ove tehnike:

- Podela prema broju nezavisnih promenljivih (faktora):
  - Jednofaktorski model analize varijanse (ili jednostruka analiza varijanse) – kada se posmatra samo jedan faktor, tretman.
  - Višefaktorski model analize varijanse (ili višestruka analiza varijanse) – kada se posmatra više od jednog faktora.
- Podela prema broju zavisnih promenljivih.
  - Univarijantna analiza varijanse – jedna zavisna promenljiva.
  - Multivarijantna analiza varijanse (MANOVA) – više od jedne zavisne promenljive.

U ovoj glavi razmatraćemo univarijantnu analizu varijanse, a u sledećoj glavi multivarijantnu analizu varijanse.

Pretpostavke analize varijanse su da podaci imaju normalnu raspodelu, da su varijanse uzoraka jednake (homoskedastičnost) i da su izabrani uzorci iz populacije međusobno nezavisni. Analiza varijanse ponovljenih merenja koristi isti uzorak jedinki za nekoliko merenja, pa je potrebno da vodimo računa o tome da su rezultati svakog subjekta zavisni tokom vremena.

Predstavićemo model koji se koristi u slučaju kada su podaci balansirani – merenja svake jedinke se vrše u istih n vremenskih momenata, bez izostavljenih vrednosti bilo koje jedinke. Dakle, svaka pojedinačna jedinka ima pridružen n - dimenzionalni slučajni vektor, čiji j – ti element predstavlja rezultat merenja u j – tom (zajedničkom) vremenskom trenutku.

Ovaj model zahteva specifičnu pretpostavku o kovarijansama vektora podataka, a ukoliko se metod koristi kada ta pretpostavka nije tačna, postoji mogućnost pogrešnih zaključaka u slučaju da ta pretpostavka nije relevantna za podatke koji se analiziraju. Model takođe obezbeđuje jednostavnu reprezentaciju vektora očekivanih vrednosti. Upravo zbog te jednostavnosti se model često koristi.

3.1. Osnovni statistički model

Osnovni faktori koji se posmatraju su:

- Jedinke su slučajno raspoređene u jednu od $q \geq 1$ grupa, tretmana. U literaturi se koristi izraz *faktor između jedinki ili grupa* (among-units factor, between-units factor). Na primer dve grupe – muškarci i žene ili tri različite starosne grupe.
- Obeležje koje se posmatra meri se n puta u n različitih uslova. U longitudinalnim studijama to je obično vreme, ali može biti i nešto drugo. Na primer merenja mogu biti maksimalan broj otkucaja srca u 10, 20, 30, 45 i 60 - om minutu posle hodanja na traci za vežbanje. Koristi se izraz *faktor unutar jedinki* (within-units factor).

Kao što je već objašnjeno, model razlikuje dva moguća izvora varijacije zbog kojih se rezultati merenja subjekata iz iste grupe (uzeti u isto vreme) razlikuju, a to su:

- slučajne varijacije među jedinkama,
- slučajne varijacije unutar jedinki.

Ovi izvori će biti značajni za određivanje prirode kovarijansne matrice vektora podataka.

Predstavljamo model sa notacijom malo različitom nego što smo do sada koristili. Definišimo slučajnu promenljivu

$$Y_{hlj} - \text{posmatranje jedinke } h \text{ iz } l - \text{te grupe u vremenu } j$$

gde

- $h = 1, \dots, r_l$, gde r_l označava broj jedinki u grupi l. Indeks h označava jedinku unutar posmatrane grupe.
- $l = 1, \dots, q$, je indeks grupe
- $j = 1, \dots, n$, je indeks vremena.
- Broj svih posmatranih jedinki je $m = \sum_{l=1}^{q} r_l$. Svaka jedinka je posmatrana u n vremenskih tačaka.

Model za Y_{hlj} je dat na sledeći način

$$Y_{hlj} = \mu + \tau_l + b_{hl} + \gamma_j + (\tau\gamma)_{lj} + e_{hlj} \tag{3.1}$$

gde je,

- μ je "ukupna" očekivana vrednost.
- τ_l je odstupanje od ukupne očekivane vrednosti povezano sa pripadanjem grupi l.
- b_{hl} je slučajan efekat sa $E(b_{hl}) = 0$ koji predstavlja odstupanje zbog činjenice da je Y_{hlj} izmereno na h – toj jedinki l – te grupe. Rezultati merenja se razlikuje od ukupne očekivane vrednosti zbog slučajne varijacije među jedinkama. Ukoliko sve jedinke populacije primaju tretman grupe l, svaka jedinka će imati svoju devijaciju jer se biološki razlikuje od drugih jedinki. Dakle, možemo smatrati da je populacija predstavljena raspodelom verovatnoća svih mogućih b_{hl} vrednosti, jedna za svaku jedinku populacije. Dakle, promenljiva b_{hl} karakteriše izvor slučajne varijacije nastao zbog razlike među jedinkama. Termin slučajni efekat je uobičajen da se opiše komponenta modela koja govori o varijaciji među subjetima.
- γ_j je odstupanje ukupne prosečne vrednosti povezano sa vremenom j.
- $(\tau\gamma)_{lj}$ je dodatno odstupanje povezano sa pripadnošću grupi l i vremenom j; to je interakcija efekta grupe l i vremena j.
- e_{hlj} je slučajno odstupanje sa $E(e_{hlj}) = 0$, koje predstavlja odstupanje uzrokovano ukupnim efektima fluktuacija unutar jedinki i greškama merenja. Ako posmatramo populaciju svih mogućih kombinacija fluktuacija i grešaka merenja koje mogu nastati, tu populaciju možemo predstaviti raspodelom verovatnoća svih mogućih e_{hlj} vrednosti. Najčešće se koristi izraz slučajna greška za ovu komponentu modela, ali se može koristiti i izraz slučajno odstupanje.

Treba uočiti da model (3.1) ne sadrži eksplicitne vrednosti vremena u kojima se merenja vrše (npr. godine: 3, 5, 7, 9). Umesto toga uvedeni su parametri γ_j i $(\tau\gamma)_{lj}$ povezani s vremenom. Dakle, u modelu nije uzeto u obzir da su merenja hronološki poređana, na primer, da li su jednake vremenske udaljenosti susednih merenja.

Model (3.1) pokazuje kako zamišljamo da sistematski faktori (kao što su vreme i tretmani - grupe) i slučajne varijacije mogu uticati na rezultat. Da bismo ovo bolje istražili, predstavimo model na drugačiji način,

$$Y_{hlj} = \mu_{lj} + \varepsilon_{hlj} \tag{3.2}$$

gde je

- $\mu_{lj} = \mu + \tau_l + \gamma_j + (\tau\gamma)_{lj}$
- $\varepsilon_{hlj} = b_{hl} + e_{hlj}$,

- $\varepsilon_{hlj} = b_{hl} + e_{hlj}$ je suma slučajnih odstupanja koja uzrokuju da se Y_{hlj} razlikuje od očekivanja u vremenu j za h – tu jedinku l – te grupe. Promenljiva ε_{hlj} sumira sve izvore slučajne varijacije.
- Primetimo da b_{hl} ne sadrži indeks j, što znači da je odstupanje h – te jedinke l – te grupe populacije u odnosu na očekivanje rezultata isto za sva vremena.
- Kako b_{hl} i e_{hlj} imaju očekivanje 0, dobijamo

$$E(Y_{hlj}) = \mu_{lj} = \mu + \tau_l + \gamma_j + (\tau\gamma)_{lj}.$$

Primetimo da μ_{lj} predstavlja očekivanje za jedinku iz l – te grupe u j – tom periodu (posmatranju). Vidimo da μ_{lj} obuhvata odstupanja od ukupnog očekivanja uzrokovana: fiksnim sistematskim efektom, dejstvom, grupe l na ukupno očekivanje u svim vremenskim tačkama ($\boldsymbol{\tau}_l$), fiksnim sistematskim efektom na ukupno očekivanje koje se dešava bez obzira na grupe u vremenu j (γ_j) i dodatnim fiksnim sistematskim efektom na ukupno očekivanje koji se javljaju za grupu l u vremenu j ($(\tau\gamma)_{lj}$).

➢ Pretpostavka o normalnoj raspodeli

Za podatke neprekidnog tipa često je realno pretpostaviti da se ponašaju po normalnoj raspodeli. Ako je $\boldsymbol{Y}_{hlj}$ neprekidno, za očekivati je i da su odstupanja nastala usled bioloških varijacija (između-jedinki) kao i iz izvora unutar-jedinki, takođe neprekidne promenljive. Zato, umesto pretpostavke da $\boldsymbol{Y}_{hlj}$ ima normalnu raspodelu, uobičajeno je da se pretpostavlja da svaka slučajna komponenta modela ima normalnu raspodelu.

Standardne pretpostavke koje obuhvataju i pretpostavke o varijansi su:

- $b_{hl} \sim \mathcal{N}(0, \sigma_b^2)$ i svi su nezavisni. To znači da je raspodela odstupanja u populaciji jedinki centrirana oko 0 (neka odstupanja su pozitivna, a neka negativna), sa varijacijom okarakterisanom komponentom varijanse σ_b^2. Činjenica da je normalna raspodela ista za sve $l = 1, \dots, q$ potiče od pretpostavke da jedinke slično variraju među sobom u svih q populacija. Pretpostavka o nezavisnosti posledica je realne činjenice da je dobijeni rezultat jednog subjekta populacije u bilo kom vremenu potpuno nepovezan sa dobijenim rezultatima drugog subjekta.
- Greške $e_{hlj} \sim \mathcal{N}(0, \sigma_e^2)$ i sve su nezavisne. To znači da je raspodela odstupanja nastalih iz uzroka unutar jedinki centrirana oko 0 (neka odstupanja su pozitivna, a neka negativna), sa varijacijom opisanom (zajedničkom) komponentom varijanse σ_e^2. Pretpostavlja se da je ova raspodela ista za sve $l = 1, \dots, q$ i $j = 1, \dots, n$. Varijansa σ_e^2 predstavlja varijaciju fluktuacije i grešaka merenja i pretpostavlja se da je konstantna tokom vremena i za sve grupe. Dakle, model pretpostavlja da su efekti varijacija unutar jedinki isti za sva vremena u svim grupama.
  Pretpostavka o nezavisnosti grešaka se mora pažljivo razmotriti. Uobičajeno je pretpostaviti da greške merenja u jednoj vremenskoj tački nisu povezane sa greškama merenja koje se dese u nekim drugim vemenskim momentima, odnosno da greške merenja nastaju slučajno. Tako, ako e_{hlj} predstavljaju greške merenja, pretpostavka o

nezavisnosti je razumna. Međutim, odstupanja unutar jedinke mogu biti povezana, korelisana, pa ako su merenja vršena u bliskim vremenskim trenucima, ne može se pretpostaviti nezavisnost. (Rezultati dobijeni u bliskim vremenskim momentima "teže" da se ponašaju na isti način, da budu zajedno "mali" ili "veliki").

- Za b_{hl} i e_{hlj} se pretpostavlja da su međusobno nezavisni. Ovo sledi iz toga što su odstupanja unutar jedinki slična po veličini bez obzira na veličinu odstupanja b_{hl} koja su povezana sa jedinkama na kojima se posmatranja vrše.

➢ Vektorska reprezentacija i kovarijansna matrica.

Razmatramo sada podatke pojedinačnog subjekta. Indeksi h i l označavaju $h-$tu jedinku u l -toj grupi. Za posmatranja u *n* perioda i dobijamo sledeće

$$\begin{bmatrix} Y_{hl1} \\ Y_{hl2} \\ \vdots \\ Y_{hln} \end{bmatrix} = \begin{bmatrix} \mu + \tau_l + \gamma_1 + (\tau\gamma)_{l1} \\ \mu + \tau_l + \gamma_2 + (\tau\gamma)_{l2} \\ \vdots \\ \mu + \tau_l + \gamma_n + (\tau\gamma)_{ln} \end{bmatrix} + \begin{bmatrix} b_{hl} \\ b_{hl} \\ \vdots \\ b_{hl} \end{bmatrix} + \begin{bmatrix} e_{hl1} \\ e_{hl2} \\ \vdots \\ e_{hln} \end{bmatrix} \tag{3.3}$$

$$\boldsymbol{Y}_{hl} = \boldsymbol{\mu_l} + \mathbf{1}b_{hl} + \boldsymbol{e}_{hl},$$

gde je $\mathbf{1}$ $(n \times 1)$ vektor jedinica. Model možemo zapisati na sledeći, kraći način:

$$\begin{bmatrix} Y_{hl1} \\ Y_{hl2} \\ \vdots \\ Y_{hln} \end{bmatrix} = \begin{bmatrix} \mu_{l1} \\ \mu_{l2} \\ \vdots \\ \mu_{ln} \end{bmatrix} + \begin{bmatrix} \varepsilon_{hl1} \\ \varepsilon_{hl2} \\ \vdots \\ \varepsilon_{hln} \end{bmatrix} \tag{3.4}$$

$$\boldsymbol{Y}_{hl} = \boldsymbol{\mu_l} + \boldsymbol{e}_{hl},$$

za vektor podataka *h* – te jedinke iz *l* – te grupe,

$$E(\boldsymbol{Y}_{hl}) = \boldsymbol{\mu_l}.$$

Primetimo da je $\boldsymbol{\mu_l}$ isto za sve jedinke iste grupe (*l* – te).

Videćemo da model implicira nešto posebno o tome kako rezultati među jedinkama i unutar jedinki kovariraju i o tome kakva je struktura očekivanja vektora podataka.

Kako su b_{hl} i e_{hlj} nezavisni, dobijamo sledeće

$$var(Y_{hlj}) = var(b_{hl}) + var(e_{hl}) + 2cov(b_{hl}, e_{hl}) = \sigma_b^2 + \sigma_e^2 + 0 = \sigma_b^2 + \sigma_e^2.$$

Dalje, kako slučajne komponente b_{hl} i e_{hl} imaju normalnu raspodelu, svako Y_{hlj} ima normalnu raspodelu, pa će i vektor $\boldsymbol{Y}_{hl}$ imati zajedničku normalnu raspodelu. Dakle, vektor podataka $\boldsymbol{Y}_{hl}$ pod pretpostavkama modela, ima multivarijantnu (*n* - dimenzionalnu) normalnu raspodelu sa očekivanjem $\boldsymbol{\mu_l}$.

Navedimo tri činjenice koje će nam koristiti u daljem radu, prilikom ispitivanja kovarijansi.

- Ukoliko su *b* i *e* dve slučajne promenljive sa očekivanjem μ_b i μ_e, tada $cov(b, e) = 0$, implicira da je $E(be) = E(b)E(e) = \mu_b\mu_e$.
- Ako *b* i *e* imaju zajedničku normalnu raspodelu i nezavisne su, tada je tada $cov(b, e) = 0$.

- Ako su b i e nezavisne i imaju zajedničku normalnu raspodelu, sledi da je $E(be) = \mu_b\mu_e$. Ako je još $E(b) = E(e) = 0$, tada je $E(be) = 0$.

Prvo, neka su Y_{hlj} i $Y_{h'l'j'}$ dva posmatranja različitih jedinki (h i h') iz različitih grupa (l i l') u različitim vremenima (j i j').

$$cov(Y_{hlj}, Y_{h'l'j'}) = E\left[(Y_{hlj} - \mu_{lj})(Y_{h'l'j'} - \mu_{l'j'})\right] = E\left[(b_{hl} + e_{hlj})(b_{h'l'} + e_{h'l'j'})\right]$$
$$= E(b_{hl}b_{h'l'}) + E(e_{hlj}b_{h'l'}) + E(b_{hl}e_{h'l'j'}) + E(b_{hl}e_{h'l'j'}) \qquad (3.5)$$

Kako je pretpostavljeno da su sve slučajne komponente modela međusobno nezavisne sa očekivanjem 0 sledi da je svaki sabirak u jednačini (3.5) jednak 0. To dalje implicira da dva rezultata različitih jedinki iz različitih grupa u različito vreme nisu korelisana.

Analogno, isto važi ako je $l = l'$, tj. ako su rezultati za dve jedinke iz iste grupe i/ili $j = j'$, odnosno posmatranja različitih jedinki u isto vreme.

$$cov(Y_{hlj}, Y_{h'lj'}) = 0, \quad cov(Y_{hlj}, Y_{h'l'j}) = 0, \quad cov(Y_{hlj}, Y_{h'lj}) = 0.$$

Može se zaključiti da model (3.1) implicira da bilo koja dva rezultata različitih jedinki imaju kovarijansu 0. Dalje, kako ova posmatranja imaju normalnu raspodelu, sledi da su bilo koja dva rezultata različitih jedinki nezavisna. Vektori $\boldsymbol{Y}_{hl}$ i $\boldsymbol{Y}_{h'l'}$ za različite jedinke, gde je $l \neq l'$ ili $l = l'$, u ovom modelu su nezavisni.

Sada razmotrimo dva posmatranja iste jedinke, recimo h – te jedinka iz l – te grupe Y_{hlj} i $Y_{hlj'}$ u dva vremenska momenta.

$$cov(Y_{hlj}, Y_{hlj'}) = E\left[(Y_{hlj} - \mu_{lj})(Y_{hlj'} - \mu_{lj'})\right] = E\left[(b_{hl} + e_{hlj})(b_{hl} + e_{hlj'})\right]$$
$$= E(b_{hl}b_{hl}) + E(e_{hlj}b_{hl}) + E(b_{hl}e_{hlj'}) + E(b_{hl}e_{hlj'})$$
$$= \sigma_b^2 + 0 + 0 + 0 = \sigma_b^2. \qquad (3.6)$$

Ovo dobijamo jer su sve slučajne promenljive u poslednja tri sabirka međusobno nezavisne, po pretpostavkama je i

$$E(b_{hl}b_{hl}) = E(b_{hl} - 0)^2 = var(b_{hl}) = \sigma_b^2.$$

Sumirajući gore rečeno dobijamo oblik kovarijansne matrice

$$var(\boldsymbol{Y}_{hl}) = \begin{bmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \dots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma_e^2 \end{bmatrix} \qquad (3.7)$$

Ovu matricu smo mogli dobiti korišćenjem matričnih operacija primenjenih na matrični zapis u (3.3). Kako su b_{hl} i elementi iz e_{hl} nezavisni sa normalnom raspodelom, $\mathbf{1}b_{hl}$ i $\boldsymbol{e}_{hl}$ su nezavisni vektori sa multivarijantnom normalnom raspodelom,

$$var(\boldsymbol{Y}_{hl}) = var(\mathbf{1}b_{hl}) + var(\boldsymbol{e}_{hl}) = \mathbf{1}var(b_{hl})\mathbf{1}' + var(\boldsymbol{e}_{hl}). \qquad (3.8)$$

Znamo da je $var(b_{hl}) = \sigma_b^2$, pa je

$$\mathbf{1}\mathbf{1}' = \boldsymbol{J}_n = \begin{bmatrix} 1 & \dots & 1 \\ 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix} \quad \text{i} \quad var(\boldsymbol{e}_{hl}) = \sigma_e^2\boldsymbol{I}_n.$$

Primenjujući to na (3.8) dobijamo

$$var(\boldsymbol{Y}_{hl}) = \sigma_b^2 \boldsymbol{J}_n + \sigma_e^2 \boldsymbol{I}_n = \boldsymbol{\Sigma}. \quad (3.9)$$

Koristićemo oznaku $\boldsymbol{J}_n$ da označimo kvadratnu matricu svih jedinica, gde indeks n označava dimenziju matrice, a $\boldsymbol{I}$ je kvadratna matrica gde su jedinice na dijagonali, a nule van dijagonale.

3.1.1. Složena simetrija

Kod analize varijanse treba da je zadovoljena pretpostavka homogenih varijansi. Kod analize varijanse sa ponovljenim merenjima važi slična, ali složenija pretpostavka, takozvani uslov složene simetrije, što znači da su pored jednakih varijansi (dijagonalni elementi kovarijansne matrice) i kovarijanse između parova promenljivih takođe jednake (van-dijagonalni elementi kovarijansne matrice)[2].

U modelima datim u (3.7) i (3.9) kovarijansa matrica zadovoljava uslov složene simetrije. Vandijagonalni elementi su svi jednaki σ_b^2. Ako računamo korelacije, one su sve iste i jednake $\sigma_b^2/(\sigma_b^2 + \sigma_e^2)$. Korelacija je pozitivna jer su σ_b^2 i σ_e^2 pozitivne varijanse. Ranije smo videli da bez obzira koliko su dva elementa iz $\boldsymbol{Y}_{hl}$ vremenski udaljena, stepen povezanosti između njih je isti. Dijagonalni elementi u (3.7) su svi jednaki $(\sigma_b^2 + \sigma_e^2)$, što znači da je varijansa svakog elementa $\boldsymbol{Y}_{hl}$ ista.

Takođe, koristi se i uslov sferičnosti, koji je pretpostavka o strukturi kovarijansne matrice za ponovljena merenja. Sferičnost je pretpostavka da su varijanse razlika između svih rezultata u različitim vremenskim momentima jednake:

$$\sigma^2\left(X_{hlj} - X_{hlj'}\right) = const, \quad j, j' = 1,2, \dots, n.$$

Ukoliko je zadovoljen uslov složene simetrije zadovoljena je i sferičnost. U praksi možemo očekivati da varijanse razlika rezultata u različitim vremenskim momentima u posmatranom uzorku budu slične. Dakle, ako su u kovarijansnoj matrici slične kovarijanse i slične varijanse, sferičnost neće biti problem.

Složena simetrija je jači uslov od sferičnosti, pa ukoliko složena simetrija nije zadovoljena, ipak moramo proveriti da li je zadovoljen uslov sferičnosti.

Na primer, neka je data sledeća kovarijansna matrica:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 10 & 5 & 10 & 15 \\ 5 & 20 & 15 & 20 \\ 10 & 15 & 30 & 25 \\ 15 & 20 & 25 & 40 \end{bmatrix}$$

$$\sigma_{1-2}^2 = \sigma_1^2 + \sigma_2^2 - 2cov(1,2) = 10 + 20 - 2 \cdot 5 = 20$$
$$\sigma_{1-3}^2 = \sigma_1^2 + \sigma_3^2 - 2cov(1,3) = 10 + 30 - 2 \cdot 10 = 20$$
$$\sigma_{1-4}^2 = \sigma_1^2 + \sigma_4^2 - 2cov(1,4) = 10 + 40 - 2 \cdot 15 = 20$$
$$\sigma_{2-3}^2 = \sigma_3^2 + \sigma_3^2 - 2cov(2,3) = 20 + 30 - 2 \cdot 15 = 20$$
$$\sigma_{2-4}^2 = \sigma_3^2 + \sigma_4^2 - 2cov(2,4) = 20 + 40 - 2 \cdot 20 = 20$$
$$\sigma_{3-4}^2 = \sigma_3^2 + \sigma_4^2 - 2cov(3,4) = 30 + 40 - 2 \cdot 25 = 20$$

Ovaj primer ilustruje da je složena simetrija jači uslov od sferičnosti tj. pokazuje da nedostatak složene simetrije ne mora da znači da je narušena sferičnost. (Složena simetrija je dovoljan, ali ne i potreban uslov za sferičnost). Varijanse razlika rezultata u vremenu su

[2] Varijanse ne moraju biti jednake kovarijansama.

jednake, a ni dijagonalni, ni vandijagonalni elementi nisu jednaki (što je vrlo neobično za realne podatke).

Pretpostavka složene simetrije može biti restriktivna za longitudinalne podatke – jer naglasak stavlja na varijacije između jedinki. Ako korelacije unutar jedinki (zbog fluktuacija) nisu zanemarljive reprezentacija može biti loša. Dakle, model (3.1) implicira prilično restriktivnu pretpostavku o prirodi varijacije unutar vektora podataka.

Takođe, kovarijansna matrica (3.7) je ista za sve jedinke, bez obzira na grupe.

Uobičajeno je korišćenje modela (3.1) kao osnove za analizu longitudinalnih podataka, ali to nekad može biti neprikladno, jer model traži jake pretpostavke o korelaciji između posmatranja iste jedinke tokom vremena.

- Drugačija notacija

Možemo koristiti notaciju predstavljenu ranije (2.3). Indeks $h = 1, \dots, r_l$ označava jedinke unutar grupa, a l označava grupe; imamo $m = \sum_{l=1}^{q} r_l$ jedinki. Možemo promeniti oznake, koristeći jedan indeks, $i = 1, \dots, m$, za bilo koju jedinku, određen svojom jedinstvenom vrednošću h i l. Menjamo i b_{hl} i $\boldsymbol{e}_{hl}$ na isti način.

Sada $\boldsymbol{Y}_i, i = 1, \dots, m$,

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{bmatrix},$$

označava vektor $\boldsymbol{Y}_{hl}, h = 1, \dots, r_l, l = 1, \dots, q$, sa promenjenim indeksima. Slično, pišemo b_i i e_i. Da bismo predstavili model sa ovim oznakama, moramo na neki način naglasiti informacije o pripadnosti grupi, jer se ne vidi eksplicitno iz indeksa. Zato ćemo napisati model na sledeći način.

Neka $\boldsymbol{M}$ označava matricu svih očekivanih vrednosti μ_{lj} (kao u modelu (3.1)),

$$\boldsymbol{M} = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{q1} & \mu_{q2} & \dots & \mu_{qn} \end{bmatrix}. \tag{3.10}$$

l – ta vrsta matrice $\boldsymbol{M}$ je transponovani vektor $\boldsymbol{\mu}_l$ (dimenzije $(n \times 1)$), pa možemo pisati

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{\mu}_1' \\ \vdots \\ \boldsymbol{\mu}_q' \end{bmatrix}.$$

Koristeći novu notaciju, neka je za $l = 1, \dots, q$,

$$a_{il} = \begin{cases} 1, \text{ako je jedinka } i \text{ iz } l- \text{ te grupe}, \\ 0, \qquad\qquad\qquad\qquad \text{inače.} \end{cases}$$

a_{il} daje informaciju o pripadnosti grupi. Neka je dalje $\boldsymbol{a}_i$ vektor (dimenzije $(q \times 1)$) čiji su elementi a_{il} vrednosti koje odgovaraju i – toj jedinki,

$$\boldsymbol{a}_i' = [a_{i1} \quad a_{i2} \quad \dots \quad a_{iq}].$$

Kako svaka jedinka može pripadati samo jednoj grupi, $\boldsymbol{a}_i$ će biti vektor nula i jedne jedinice na mestu koje odgovara i – toj grupi.

Na primer, ako imamo $q = 3$ grupe i $n = 4$ vremenske tačke, imamo

$$\boldsymbol{M} = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} \\ \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{bmatrix}$$

i ako je *i* – ta jedinka iz druge grupe, tada je

$$\boldsymbol{a_i'} = [0 \quad 1 \quad 0]$$

Dalje dobijamo,

$$\boldsymbol{a_i'M} = [\mu_{21} \quad \mu_{22} \quad \mu_{23} \quad \mu_{24}] = \boldsymbol{\mu_i'},$$

što je vektor očekivanja *i* – te jedinke. Elementi vektora $\boldsymbol{\mu_i}$ su određeni grupom kojoj *i* – ta jedinka pripada i isti su za sve jedinke iste grupe.

Dalje, možemo model (3.3) i (3.4) predstaviti kao

$$\boldsymbol{Y_i'} = \boldsymbol{a_i'M} + \boldsymbol{1'}b_i + \boldsymbol{e_i'}, \qquad i = 1, \dots, m.$$

tj.

$$\boldsymbol{Y_i'} = \boldsymbol{a_i'M} + \boldsymbol{\varepsilon_i'}, \ i = 1, \dots, m, \tag{3.11}$$

gde je $\boldsymbol{\epsilon_i'} = [\epsilon_{i1}, \dots, \epsilon_{in}]$.

Ovo je standardni način predstavljanja modela kada se jedinke indeksiraju jednim indeksom (*i* u ovom slučaju).

Postoji još jedan način za predstavljanje ovog modela, koji ćemo koristiti u kasnijim razmatranjima. Neka je **β** vektor svih parametara modela (3.1) za sve grupe i sva vremena tj. sve μ, τ_l, γ_j i $(\tau\gamma)_{lj}, l = 1, \dots, q, j = 1, \dots, n$.

Na primer, za $q = 2$ grupe i $n = 3$ vremenske tačke, imamo

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ (\tau\gamma)_{11} \\ (\tau\gamma)_{12} \\ (\tau\gamma)_{13} \\ (\tau\gamma)_{21} \\ (\tau\gamma)_{22} \\ (\tau\gamma)_{23} \end{bmatrix}.$$

Sada je $E(\boldsymbol{Y_i}) = \boldsymbol{\mu_i}$. Ako je na primer, *i* u grupi 2, tada je

$$\boldsymbol{\mu_i} = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} = \begin{bmatrix} \mu + \tau_2 + \gamma_1 + (\tau\gamma)_{21} \\ \mu + \tau_2 + \gamma_2 + (\tau\gamma)_{22} \\ \mu + \tau_2 + \gamma_3 + (\tau\gamma)_{23} \end{bmatrix}$$

i ako definišemo

$$\boldsymbol{X_i} = \left[\begin{array}{c|cc|ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{array}\right. \left.\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right],$$

možemo napisati

$$\boldsymbol{\mu_i} = \boldsymbol{X_i\beta}.$$

Opšti model, ako adekvatno definišemo **β** i $\boldsymbol{X_i}$, možemo zapisati kao

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\beta + \boldsymbol{1}b_i + \boldsymbol{e_i} \text{ ili } \boldsymbol{Y_i} = \boldsymbol{X_i\beta} + \boldsymbol{\varepsilon_i}, i = 1, \dots, m.$$

$\boldsymbol{X_i}$ je odgovarajuća matrica nula i jedinica i može biti ista za svako *i* u istoj grupi.

Predstavljajući očekivanja μ_{lj} parametrima μ, τ_l, γ_j i $(\tau\gamma)_{lj}$ dolazimo do modela koji ima previše parametara. Naime, iako imamo dovoljno informacija da objasnimo kako se očekivanja μ_{lj} razlikuju, nemamo dovoljno informacija da objasnimo kako se ona razlažu na sve ove komponente. Na primer, ako smo imali dve tretman grupe, ne možemo reći koliki su μ, τ_1 i τ_2 samo na osnovu informacija koje imamo. Pretpostavimo da znamo da je $\mu + \tau_1 = 30$ i $\mu + \tau_2 = 20$. Ovo zadovoljeno ako je

$$\mu = 25,\ \tau_1 = 5,\ \tau_2 = -5;$$

ili

$$\mu = 21,\ \tau_1 = 9,\ \tau_2 = -1;$$

Možemo analogno napisati beskonačno mnogo takvih primera. Ovaj problem se može sagledati shvatanjem da matrica $\boldsymbol{X_i}$ nije punog ranga.

Iako je ovakva reprezentacija očekivanja μ_{lj} korišćena u kontekstu analize varijanse pogodna da nam pomogne u shvatanju efekata različitih faktora odstupanja od "ukupnog" očekivanja, ne možemo identifikovati sve ovakve komponente. U cilju da ih identifikujemo, nameću se ograničenja koja čine reprezentaciju jedinstvenom (primoravajući na samo jedan od beskonačno mnogo načina):

$$\sum_{l=1}^{q} \tau_l = 0, \qquad \sum_{j=1}^{n} \gamma_j = 0, \qquad \sum_{l=1}^{q} (\tau\gamma)_{lj} = 0 = \sum_{j=1}^{n} (\tau\gamma)_{lj}$$

za sve j, l. Nametanje ovih ograničenja je ekvivalentno redefinisanju vektora parametara $\boldsymbol{\beta}$ i matrice $\boldsymbol{X_i}$ tako da $\boldsymbol{X_i}$ uvek bude matrica punog ranga za sve *i*.

3.2. Važna pitanja i hipoteze

U ovom delu izložićemo kako pitanja koja nas interesuju mogu biti rešena primenom modela za longitudinalne podatake. Podsetimo se da model možemo napisati kao (3.11),

$$\boldsymbol{Y_i'} = \boldsymbol{a_i'M} + \boldsymbol{\varepsilon_i'}, \quad i = 1, \dots, m.$$

gde je

$$\boldsymbol{M} = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{q1} & \mu_{q2} & \cdots & \mu_{qn} \end{bmatrix}$$

i

$$\mu_{lj} = \mu + \tau_l + \gamma_j + (\tau\gamma)_{lj}. \tag{3.12}$$

pretpostavlja se da ograničenja ostaju ista,

$$\textstyle\sum_{l=1}^{q} \tau_l = 0,\ \sum_{j=1}^{n} \gamma_j = 0,\ \sum_{l=1}^{q} (\tau\gamma)_{lj} = 0 = \sum_{j=1}^{n} (\tau\gamma)_{lj}. \tag{3.13}$$

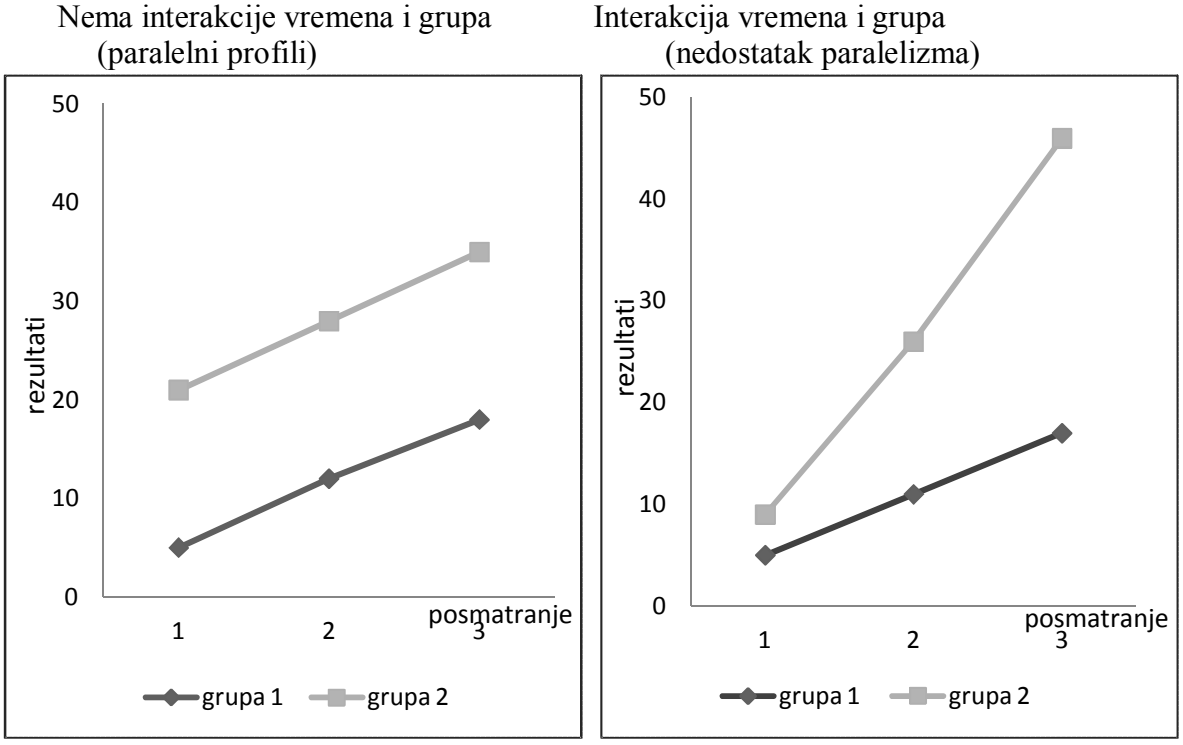
3.2.1 Interakcija grupa i vremena

Cilj u analizi longitudinalnih podataka je da se proveri da li se rezultat merenja menja tokom vremena za različite grupe. U tu svrhu se uobičajeno posmatraju očekivanja, proseci. Test za interakciju grupa i vremena naziva se *test paralelizma*.

Možemo grafički prikazati komponente sredine $\boldsymbol{\mu}$ svake od grupa posebno, počevši od prvog do n - tog merenja, pri čemu nacrtane tačke spajamo. Dobijene linije se nazivaju **profili grupa**.

Na primer, zanima nas koliko telesna težina razlikuje kod dečaka i kod devojčica tokom vremena? Ovo je ilustrovano na slici 3. Posmatrajmo dve grupe ($q = 2$) gde su merenja ponovljena tri puta ($n = 3$). Na slici 3 pokazana su dva moguća slučaja. Na svakom grafiku linija predstavlja očekivanu vrednost μ_{lj} rezultata za svaku grupu posebno. Možemo primetiti da je očekivana vrednost rezultata u svakom vremenskom momentu veća za grupu 2 nego za grupu 1. Na levom grafiku stopa promene očekivane vrednosti rezultata tokom vremena je ista za obe grupe, odnosno vremenski profili su paralelni, dok je na desnom grafiku stopa promene veća za grupu 2, što znači da profili nisu paralelni.

Slika 3. Interakcija vremena i grupa.



Na slici su predstavljena dva profila očekivanih vrednosti za dve grupe. Posmatranjem profila svake grupe, nameću se različita pitanja. Prvo pitanje je koje nas zanima jeste da li su profili paralelni?

Svaka tačka na slici je predstavljena forumulom (3.12),

$$\mu_{lj} = \mu + \tau_l + \gamma_j + (\tau\gamma)_{lj}.$$

Sabirci $(\tau\gamma)_{lj}$ su vrednosti kojima očekivane vrednosti l - te grupe u vremenskoj tački j mogu da se razlikuju od ukupne očekivane vrednosti. Razlika u očekivanim vrednostima između grupe 1 i grupe 2 u bilo kom posmatranom vremenu j je, prema modelu,

$$\mu_{1j} - \mu_{2j} = (\tau_1 - \tau_2) + \big((\tau\gamma)_{1j} - (\tau\gamma)_{2j}\big).$$

Dakle, sabirak $(\tau\gamma)_{lj}$ dopušta mogućnost da se razlike između očekivanih vrednosti grupa mogu razlikovati u različitim vremenskim tačkama, kao na desnom grafiku slike 3 (dok su na levom grafiku iste za sve vremenske momente). Iznos $\big((\tau\gamma)_{1j} - (\tau\gamma)_{2j}\big)$ je specifičan za vreme j.

Ukoliko su svi $(\tau\gamma)_{lj}$ isti, dobijamo da se razlika svodi na

$$\mu_{1j} - \mu_{2j} = (\tau_1 - \tau_2).$$

Ovde je razlika u očekivanoj vrednosti rezultata među grupama ista u svim vremenskim tačkama i jednaka je $(\tau_1 - \tau_2)$ (što ne zavisi od j), što je situacija levog grafika slike 3.

Ako su ograničenja

$$\sum_{l=1}^{q} (\tau\gamma)_{lj} = 0 = \sum_{j=1}^{n} (\tau\gamma)_{lj}$$

za sve l, j i ako je $(\tau\gamma)_{lj}$ jednako za sve l, j, tada je

$$(\tau\gamma)_{lj} = 0 \quad \text{za sve} \quad l, j.$$

Ukoliko želimo da razlikujemo paralelne profile (situacija na levom grafiku) i nedostatak paralelizma (desni grafik), postavljamo pitanje o jednakosti stope promene očekivane vrednosti rezultata tokom vremena, pa možemo formulisati nultu hipotezu kao

$$H_0: \text{svi } (\tau\gamma)_{lj} = 0.$$

Imamo $q \times n$ parametara $(\tau\gamma)_{lj}$. Ukoliko važe navedena ograničenja (3.13), tada ako je $(q-1)(n-1)$ vrednosti $(\tau\gamma)_{lj}$ jednako nuli, i ostali $(\tau\gamma)_{lj}$ moraju biti jednaki nula. Dakle, hipoteza je zapravo o ponašanju $(q-1)(n-1)$ parametara, pa postoji $(q-1)(n-1)$ stepeni slobode povezanih sa ovom hipotezom.

Specijalno, označavajući sa $\boldsymbol{M}$ matricu čije su vrste vektori očekivanih vrednosti za različite grupe, moguće je napisati formalne statističke hipoteze kao linearne funkcije elemenata matrice $\boldsymbol{M}$.

Neka je

- $\boldsymbol{C}$ - $(c \times q)$ matrica sa $c \leq q$ punog ranga.
- $\boldsymbol{U}$ - $(n \times u)$ matrica sa $u \leq n$ punog ranga.

Tada se nulta hipoteza o jednakosti očekivanih vrednosti može zapisati u sledećem obliku:

$$H_0: \boldsymbol{CMU} = \boldsymbol{0}.$$

U zavisnosti od izbora matrica $\boldsymbol{C}$ i $\boldsymbol{U}$, linearna funkcija $\boldsymbol{CMU}$ elemenata matrice $\boldsymbol{M}$ (individualna očekivanja za različite grupe u različitim vremenskim tačkama) može se formirati da se reše različita pitanja o jednakosti očekivanih vrednosti.

Analizirajmo prvo hipotezu H_0 o interakciji grupa i vemena. Zbog jednostavnosti, posmatrajmo slučaj kada imamo $q = 2$ grupe i $n = 3$ vremenske tačke. Neka je vektor

$$\boldsymbol{C} = [1 \quad -1],$$

Vidimo da je $c = 1 = q - 1$. Tada imamo,

$$\boldsymbol{CM} = [1 \quad -1]\begin{bmatrix}\mu_{11} & \mu_{12} & \mu_{13}\\ \mu_{21} & \mu_{22} & \mu_{23}\end{bmatrix} = [\mu_{11}-\mu_{21}, \mu_{12}-\mu_{22}, \mu_{13}-\mu_{23}]$$

$$= (\tau_1 - \tau_2 + (\tau\gamma)_{11} - (\tau\gamma)_{21},\ \tau_1 - \tau_2 + (\tau\gamma)_{12} - (\tau\gamma)_{22},\ \tau_1 - \tau_2 + (\tau\gamma)_{13} - (\tau\gamma)_{23})$$

Dakle, matrica $\boldsymbol{C}$ ima efekat pravljenja razlika između grupa.

Neka je dalje,

$$\boldsymbol{U} = \begin{bmatrix} 1 & 0\\ -1 & 1\\ 0 & -1\end{bmatrix},$$

pa je $c = 2 = n - 1$. Direktno sledi,

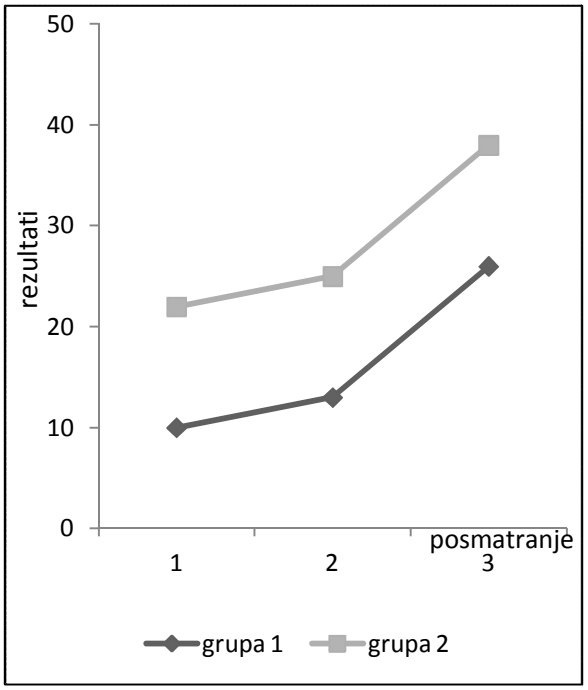
$$\boldsymbol{CMU} = (\mu_{11} - \mu_{21} - \mu_{12} + \mu_{22}, \mu_{12} - \mu_{22} - \mu_{13} + \mu_{23})$$

$$= ((\tau\gamma)_{11} - (\tau\gamma)_{21} - (\tau\gamma)_{12} + (\tau\gamma)_{22}, (\tau\gamma)_{12} - (\tau\gamma)_{22} - (\tau\gamma)_{13} + (\tau\gamma)_{23}).$$

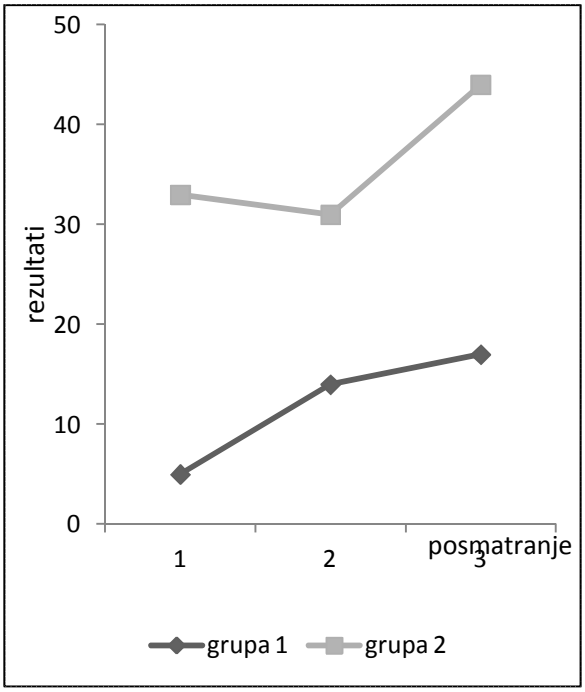
Može se proveriti da ako važe uslovi (3.13) i ako je svaki od ovih elemenata nula hipoteza H_0 je tačna. Važno je shvatiti da paralelizam ne znači da linija koja povezuje očekivane vrednosti tokom vremena izgleda kao prava linija za svaku grupu. Na slici 4, levi grafik predstavlja paralelizam, dok desni ne.

Slika 4. Interakcija grupa i vremena.

Nema interakcije grupa i vremena (paralelni profili)



Postoji interakcija grupa i vremena (nedostatak paralelizma)



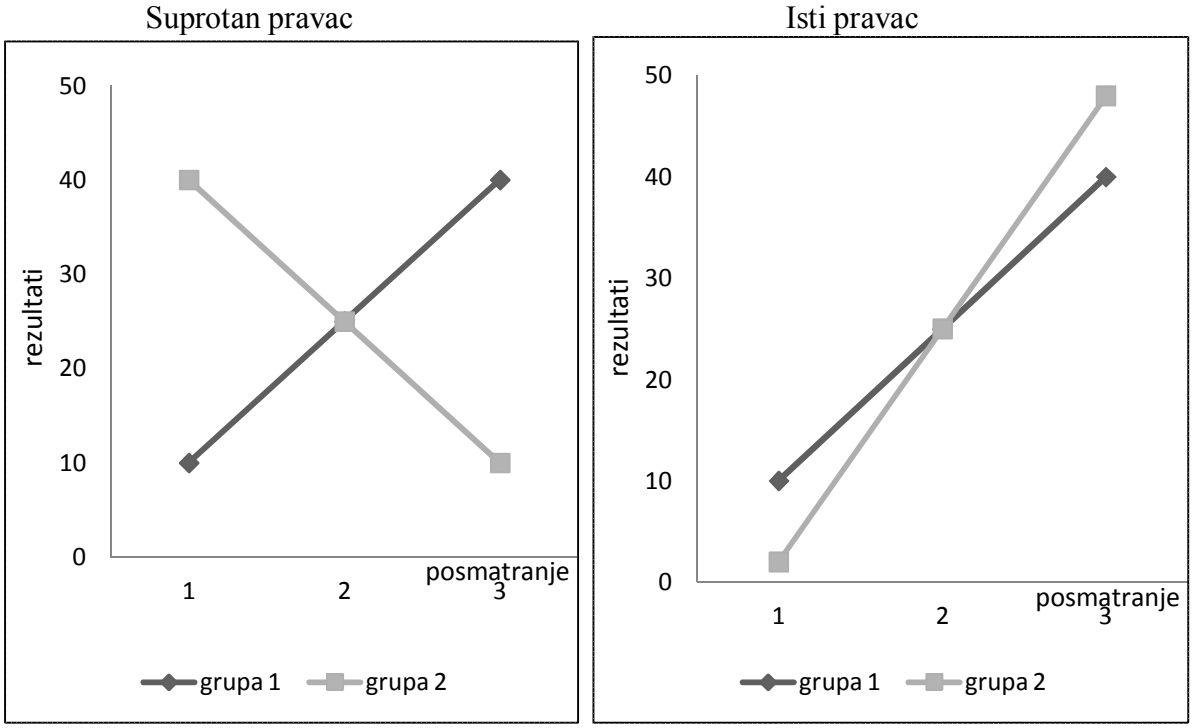
3.2.2. Glavni efekti grupa

Ukoliko su profili paralelni, sledeće pitanje koje se postavlja je da li se oni i podudaraju, odnosno, da li su u svakoj vremenskoj tački očekivane vrednosti rezultata posmatranih grupa jednake. Može se pokazati da, ukoliko su profili paralelni i ako se i poklapaju, tada su prosečne vrednosti tokom vremena očekivanih vrednosti rezultata iste za svaku grupu. Postavljanje pitanja da li su prosečne vrednosti tokom vremena očekivanih vrednosti rezultata iste za svaku grupu ukoliko profili nisu paralelni, može, a i ne mora biti od interesa.

Na primer, ako su očekivane vrednosti rezultata prikazane na desnim graficima slika 3 i 4, na pitanje da li se proseci očekivanih vrednosti rezultata tokom vremena razlikuju za dve grupe, dobili bismo odgovor da je očekivana vrednost rezultata grupe 2 veća u svim vremenskim tačkama. S druge strane, posmatrajmo levi grafik slike 5. U ovoj situaciji bi test za ovo pitanje bio beznačajan: promena očekivanih vrednosti rezultata tokom vremena je u suprotnom smeru za posmatrane dve grupe, pa je prosek tokom vremena od male važnosti.

Ukoliko promenljiva koja se proučava raste tokom vremena, analiziranje prosečnih rezultata tokom vremena može biti od malog značaja, već nas može na primer zanimati koliko se razlikuju očekivane vrednosti rezultata na kraju posmatranog perioda. Na desnom grafiku slike 5, očekivane vrednosti rezultata tokom vremena rastu za obe grupe različitim stopama, ali imaju isti prosek tokom vremena. Jasno je da grupa sa većom stopom rasta ima veću očekivanu vrednost rezultata merenja na kraju perioda.

Slika 5. Interakcija grupa i vremena.



Uopšte, najinteresantnije pitanje u analizi longitudinalnih podataka je da li su prosečne vrednosti (po vremenu) očekivanih vrednosti rezultata iste za različite grupe, ukoliko su profili tokom vremena približno paralelni. Razmotrimo slučaj kada imamo $q = 2$ grupe i $n = 3$ vremenske tačke. Zanima nas da li su prosečne vrednosti očekivanih vrednosti rezultata tokom vremena iste u svakoj grupi. Za l-tu grupu, prosek je,

$$\frac{\mu_{l1}+\mu_{l2}+\mu_{l3}}{3}=\mu+\tau_l+\frac{\gamma_1+\gamma_2+\gamma_3}{3}+\frac{(\tau\gamma)_{l1}+(\tau\gamma)_{l2}+(\tau\gamma)_{l3}}{3}.$$

Uzimajući razliku proseka između prve i druge grupe, $l=1$ i $l=2$, dobijamo

$$\tau_1-\tau_2+n^{-1}\sum_{j=1}^{n}(\tau\gamma)_{1j}-n^{-1}\sum_{j=1}^{n}(\tau\gamma)_{2j}.$$

Primetimo, međutim, da su ograničenja koja se stavljaju u model da bi model bio punog ranga oblika $\sum_{j=1}^{n}(\tau\gamma)_{lj}=0$ za svako l; pa sledi da su dve sume u ovom izrazu po pretpostavci jednake 0, pa nam ostaje izraz $\tau_1-\tau_2$. Prema tome, možemo hipotezu izraziti na sledeći način:

$$H_0\colon \tau_1-\tau_2=0.$$

Dalje, uz ograničenje $\sum_{l=1}^{q}\tau_l=0$, ako su τ_l jednaki kao što je to navedeno u H_0, tada moraju zadovoljavati $\tau_l=0$ za svako l. Zbog toga se hipoteza može zapisati kao

$$H_0\colon \tau_1=\tau_2=0.$$

Za opšte q i n, princip je isti,

$$H_0\colon \tau_1=\cdots=\tau_q=0.$$

Odgovarajuća nulta hipoteza koja se bavi ovim pitanjem može biti data u opštem obliku $H_0\colon \boldsymbol{CMU}=\boldsymbol{0}$ za odgovarajuće izbore $\boldsymbol{C}$ i $\boldsymbol{U}$. Oblik matrice $\boldsymbol{U}$ omogućuje njenu interpretaciju kao traženje proseka tokom vemena.

Neka opet imamo $q=2$ grupe i $n=3$ vremenske tačke. Neka je,

$$\boldsymbol{C}=[1\quad -1],$$

$c=1=q-1$. Tada je,

$$\boldsymbol{CM}=[1\quad -1]\begin{bmatrix}\mu_{11} & \mu_{12} & \mu_{13}\\ \mu_{21} & \mu_{22} & \mu_{23}\end{bmatrix}=[\mu_{11}-\mu_{21}\quad \mu_{12}-\mu_{22}\quad \mu_{13}-\mu_{23}]=$$

$$=[\tau_1-\tau_2+(\tau\gamma)_{11}-(\tau\gamma)_{21}\quad \tau_1-\tau_2+(\tau\gamma)_{12}-(\tau\gamma)_{22}\quad \tau_1-\tau_2+(\tau\gamma)_{13}-(\tau\gamma)_{23}]$$

Dalje, za $n=3$

$$\boldsymbol{U}=\begin{bmatrix}1/n\\ 1/n\\ 1/n\end{bmatrix},$$

Jednostavno se vidi da,

$$\boldsymbol{CMU}=\tau_1-\tau_2+n^{-1}\sum_{j=1}^{n}(\tau\gamma)_{1j}-n^{-1}\sum_{j=1}^{n}(\tau\gamma)_{2j}.$$

Ovaj izbor matrice $\boldsymbol{U}$ omogućuje njenu interpretaciju kao traženje proseka tokom vemena. Nametanjem ograničenja kao gore, hipotezu možemo izraziti na sledeći način:

$$H_0\colon \boldsymbol{CMU}=0$$

za odgovarajuće izbore za $\boldsymbol{C}$ i $\boldsymbol{U}$, gde matrica $\boldsymbol{U}$ vektor kolona jedinica što implicira traženje proseka tokom vremena i dovodi do opšte hipoteze $H_0\colon \tau_1=\cdots=\tau_q=0$.

3.2.3. Glavni efekti vremena

Pitanje od interesa može biti i da li su očekivane vrednosti rezultati konstantne tokom vremena. Ukoliko su profili paralelni, to se svodi na pitanje da li je prosek po grupama očekivanih vrednosti rezultata konstantan tokom vremena.

Ukoliko profili nisu paralelni, tada ovo može, a i ne mora biti od interesa. Na primer, primetimo da su na levom grafiku slike 5 proseci očekivanih vrednosti rezultata za grupe 1 i 2 isti u svakoj vremenskoj tački, ali očekivane vrednosti rezultata nisu konstantne tokom vremena ni za jednu grupu.

Razmotrimo ponovo slučaj kada je $q = 2$ i $n = 3$. Koristeći ograničenja $\sum_{l=1}^{q} \tau_l = 0$ i $\sum_{l=1}^{q} (\tau\gamma)_{lj} = 0$, prosek očekivanih vrednosti rezultata po grupama za j – to vreme je

$$q^{-1} \sum_{l=1}^{q} \mu_{lj} = \gamma_j + q^{-1} \sum_{l=1}^{q} \tau_l + q^{-1} \sum_{l=1}^{q} (\tau\gamma)_{lj} = \gamma_j$$

Pretpostavka da su svi ovi proseci isti u svakom posmatranom vremenskom trenutku, je ekvivalentna sa

$$H_0: \gamma_1 = \gamma_2 = \gamma_3.$$

Ako imamo i ograničenje $\sum_{j=1}^{n} \gamma_j = 0$, tada je $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0$.

Za opšte q i n, hipoteza je oblika

$$H_0: \gamma_1 = \cdots = \gamma_n = 0.$$

Možemo takođe napisati hipotezu u obliku $H_0: \boldsymbol{CMU} = \boldsymbol{0}$. Na primeru gde je $q = 2$ i $n = 3$, ako je

$$\boldsymbol{U} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}, \qquad \boldsymbol{C} = [1/2 \quad 1/2],$$

dobija se

$$\boldsymbol{MU} = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} \mu_{11} - \mu_{21} & \mu_{12} - \mu_{13} \\ \mu_{21} - \mu_{22} & \mu_{22} - \mu_{23} \end{bmatrix}$$

$$= \begin{bmatrix} \tau_1 - \tau_2 + (\tau\gamma)_{11} - (\tau\gamma)_{12} & \tau_2 - \tau_3 + (\tau\gamma)_{12} - (\tau\gamma)_{13} \\ \tau_1 - \tau_2 + (\tau\gamma)_{21} - (\tau\gamma)_{22} & \tau_2 - \tau_3 + (\tau\gamma)_{22} - (\tau\gamma)_{23} \end{bmatrix}.$$

Odavde jednostavno proizilazi, uzimajući u obzir ograničenja, da je

$$\boldsymbol{CMU} = [\gamma_1 - \gamma_2 \quad \gamma_2 - \gamma_3].$$

Izjednačavajući ovo sa nulom dobijamo $H_0: \gamma_1 = \gamma_2 = \gamma_3$. Za opšte q i n, možemo birati matrice $\boldsymbol{C}$ i $\boldsymbol{U}$ na sličan način. Primetimo da ovaj tip matrice $\boldsymbol{C}$ omogućuje njenu interpretaciju kao uzimanje proseka po grupama.

Testiranje ovih hipoteza sprovodi se na način uobičajen u analizi varijanse. Pod pretpostavkama za model (3.1) i pod pretpostavkom da zavisna promenljiva ima normalnu raspodelu, test statistike koje se koriste su uobičajeni količnici sa Fišerovom raspodelom.

3.3. Analiza varijanse kroz primere

Daćemo jednostavan numerički primer da prikažemo razlike između analize varijanse sa i bez ponovljenih merenja, a potom ćemo proširiti primer sa informacijom o pripadnosti grupi.

Tabela 2: Primer podataka sa ponovljenim merenjima.[3]

i	Y_{t1}	Y_{t2}	Y_{t3}	Y_{t4}	Prosek
1	31	29	15	26	25,25
2	24	28	20	32	26
3	14	20	28	30	23
4	38	34	30	34	34
5	25	29	25	29	27
6	30	28	16	34	27
Prosek	27	28	22,33	30,83	27

Ovo je hipotetički skup longitudinalnih podataka četiri merenja na šest subjekata.

1. Ako u primeru zanemarimo činjenicu da se radi o ponovljenim merenjima, odnosno da je ista jedinka izmerena četiri puta, na pitanje da li postoji razlika među rezultatima u različitim vremenskim tačkama, može se odgovoriti korišćenjem ANOVA, posmatrajući merenja u četiri vremenske tačke kao četiri nezavisne grupe. Faktor koji posmatramo je vreme, koji ima četiri nivoa. analiza varijanse se tada zasniva na poređenju sume kvadrata razlika između nivoa faktora i SS_B i sume kvadrata unutar nivoa faktora SS_w (poznata kao suma kvadrata grešaka). Suma kvadrata razlika između nivoa faktora se računa na sledeći način:

$$SS_B = m\sum_{j=1}^{n}\left(\bar{Y}_j - \bar{Y}\right)^2$$

gde je m broj jedinki, $j, j = 1, \dots, n$ broj nezavisnih merenja, $\bar{Y}_j$ prosek rezultata u vremenu j i $\bar{Y}$ ukupan prosek promenljive Y;

$$SS_w = \sum_{j=1}^{n}\sum_{i=1}^{m}\left(Y_{ij} - \bar{Y}_j\right)^2$$

gde je m broj jedinki, n broj merenja, Y_{ij} rezultat merenja jedinke i u vremenu j i $\bar{Y}_j$ prosek promenljive Y u vremenu j.

$$SS_B = 6[(27-27)^2 + (28-27)^2 + (22{,}33-27)^2 + (30{,}83-27)^2] = 224{,}79$$

$$SS_w = (31-27)^2 + (24-27)^2 + \ \dots + \ (29-30{,}83)^2 + (34-30{,}83)^2 = 676{,}17$$

Ove sume kvadrata se koriste u F – testu analize varijanse. Sredina sume kvadrata (MS) se definiše kao ukupna suma kvadrata podeljena sa stepenima slobode. Za SS_B stepeni slobode su $n-1$, a za SS_w stepeni slobode su $n \times (m-1)$. U ovom primeru

$$MS_B = 74{,}93 \quad \text{i} \quad MS_w = 33{,}81.$$

F statistika je $\frac{MS_B}{MS_w}$ i ima *Fišerovu* raspodelu sa $n-1$, $n(m-1)$ stepena slobode. Realizovana vrednost F statistike sa 3 i 20 stepeni slobode je $2{,}216$. Odgovarajuća p vrednost je $0{,}118$, što znači da ne postoji značajna razlika između četiri vremenske tačke.

[3] Podaci preuzeti iz knjige *Applied Longitudinal Data Analysis for Epidemiology*, Jos Twisk, Cambridge University Press.

U SPSS – u dobijamo:

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	224.792	3	74.931	2.216	.118
Within Groups	676.167	20	33.808		
Total	900.958	23			

2. Kao što je rečeno, u gornjim izračunavanjima smo zanemarili činjenicu da postoji zavisnost među posmatranjima - ignorisano je da je isti subjekat izmeren četiri puta.

U modelu sa ponovljenim merenjima izračunavamo individualnu suma kvadrata SS_i na sledeći način:

$$SS_i = n\sum_{i=1}^{m}(\bar{Y}_i - \bar{Y})^2$$

gde je m broj jedinki, n broj ponovljenih merenja, $\bar{Y}_i$ prosek svih merenja jedinke i i $\bar{Y}$ ukupan prosek promenljive Y. Dobijamo,

$$SS_i = 4[(25{,}25 - 27)^2 + (26 - 27)^2 + \ ... + (27 - 27)^2] = 276{,}21$$

Može se videti da se određeni deo (276.21/676.17) sume kvadrata greške (to jest sume kvadrata unutar jedinki, zbog faktora vremena) može objasniti individualnim razlikama. Dakle, u longitudinalnoj studiji, ukupna suma kvadrata greške 676,17 se deli na dve komponente: deo zbog pojedinačnih razlika (276,21) kao posledica delovanja ponovljenog faktora (vremena) i na deo od 399,96 (=676.17-276.27), kao posledica drugih faktora. Suma kvadrata SS_B je još uvek ista, jer ova suma kvadrata odražava razlike među četiri vremenske tačke.

U statističkom paketu SPSS, dobijamo sledeće rezultate (posmatramo rezultate kada je zadovoljena pretpostavka sferičnosti):

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
vreme	Sphericity Assumed	224.792	3	74.931	2.810	.075
	Greenhouse-Geisser	224.792	1.457	154.282	2.810	.131
	Huynh-Feldt	224.792	1.903	118.129	2.810	.111
	Lower-bound	224.792	1.000	224.792	2.810	.155
Error(vreme)	Sphericity Assumed	399.958	15	26.664		
	Greenhouse-Geisser	399.958	7.285	54.901		
	Huynh-Feldt	399.958	9.515	42.036		
	Lower-bound	399.958	5.000	79.992		

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	17550.042	1	17550.042	317.696	.000
Error	276.208	5	55.242		

3. Razmatrali smo longitudinalne studije kod kojih se neprekidna promenljiva meri više puta tokom vremena (dizajn unutar jedinki), a sada hoćemo da predstavimo situaciju kada se porede vrednosti promenljive Y i između grupa. Ovaj dizajn je poznat kao dizajn unutar jedinki i između jedinki grupa. Vreme je faktor unutar jedinki, a grupa je faktor između jedinki. Pitanje koje sad rešavamo je da li postoji razlika u promeni vrednosti promenljive Y između dve ili više grupa tokom vremena, odnosno testiramo hipoteze o efektima grupa, efektoma vremena i interakciji vremena i grupa.

Pod pretpostavkama modela (3.1) i normalne raspodele posmatranja, u tabeli 3 analize varijanse, navedene dalje u tekstu, koristi se notacija sa tri indeksa (Y_{hlj} označava rezultat h - te jedinke iz l - te grupe u trenutku j). Definišimo:

- $\bar{Y}_{hl\cdot} = n^{-1}\sum_{j=1}^{n} Y_{hlj}$, prosek tokom vremena za h - tu jedinku iz l - te grupe (za sva posmatranja te jedinke)
- $\bar{Y}_{\cdot lj} = r_l^{-1}\sum_{h=1}^{r_l} Y_{hlj}$, prosek svih jedinki iz l - te grupe u vremenu j
- $\bar{Y}_{\cdot l\cdot} = (r_l n)^{-1}\sum_{h=1}^{r_l}\sum_{j=1}^{n} Y_{hlj}$, prosek svih posmatranja l - te grupe
- $\bar{Y}_{\cdot\cdot j} = m^{-1}\sum_{l=1}^{q}\sum_{h=1}^{r_l} Y_{hlj}$, prosek svih posmatranja u vremenu j
- $\bar{Y}_{\cdots}$ = prosek svih $m \times n$ posmatranja.

Neka je

$$SS_G = \sum_{l=1}^{q} n r_l(\bar{Y}_{\cdot l\cdot} - \bar{Y}_{\cdots})^2, \quad SS_B = n\sum_{l=1}^{q}\sum_{h=1}^{r_l} r_l(\bar{Y}_{hl\cdot} - \bar{Y}_{\cdots})^2$$

$$SS_{EU} = SS_B - SS_G$$

$$SS_T = m\sum_{j=1}^{n}\left(\bar{Y}_{\cdot\cdot j} - \bar{Y}_{\cdots}\right)^2, \quad SS_{GT} = \sum_{j=1}^{n}\sum_{l=1}^{q} r_l\left(\bar{Y}_{\cdot lj} - \bar{Y}_{\cdots}\right)^2 - SS_T - SS_G$$

$$SS_{sve\ ukupno} = \sum_{l=1}^{q}\sum_{h=1}^{r_l}\sum_{j=1}^{n}\left(Y_{hlj} - \bar{Y}_{\cdots}\right)^2 = SS_B + SS_W$$

$$SS_W = SS_T + SS_{GT} + SS_E.$$

Navodimo tabelu analize varijanse.

Tabela 3: Analiza varijanse

Izvori varijacije	*Suma kvadrata*	*Stepeni slobode*	*Prosek sume kvadrata (ocena varijanse)*	*F statistike*
Između grupa	SS_G	$q-1$	MS_G	$F_G = \frac{MS_G}{MS_{EU}}$
Greška između jedinki	SS_{EU}	$m-q$	MS_{EU}	
Vreme	SS_T	$n-1$	MS_T	$F_T = \frac{MS_T}{MS_E}$
Grupa × Vreme	SS_{GT}	$(q-1)(n-1)$	MS_{GT}	$F_{GT} = \frac{MS_{GT}}{MS_E}$
Greška unutar jedinki	SS_E	$(m-q)(n-1)$	MS_E	
Ukupno	$SS_{sve\ ukupno}$	$nm-1$		

Termin greške između jedinki obuhvata varijacije zbog biološke varijacije među jedinkama, a termin greške unutar jedinki obuhvata varijacije zbog fluktuacija i grešaka merenja. Ukoliko podatke posmatramo u formi vektora, tada su pretpostavke sledeće:

$$\boldsymbol{Y}_i \sim \mathcal{N}_n(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \sigma_b^2 \boldsymbol{J}_n + \sigma_e^2 \boldsymbol{I}_n.$$

Ako vektori podataka imaju multivarijantnu normalnu raspodelu i zadovoljavaju uslov složene simetrija, tada *F* statistike imaju Fišerovu raspodelu sa stepenima slobode navedenim u tabeli.

Ovo ćemo ilustrovati koristeći ranije razmatrani primer, proširen sa informacijom o pripadnosti grupi.

Tabela 4: Podaci sa ponovljenim merenjima sa informacijom o pripadnosti grupi.

i	Grupa	$\boldsymbol{Y_{t1}}$	$\boldsymbol{Y_{t2}}$	$\boldsymbol{Y_{t3}}$	$\boldsymbol{Y_{t4}}$	Prosek
1	1	31	29	15	26	25,25
2	1	24	28	20	32	26
3	1	14	20	28	30	23
Prosek		23	25,67	21	29,33	24,75
4	2	38	34	30	34	34
5	2	25	29	25	29	27
6	2	30	28	16	34	27
Prosek		31	30,33	23,67	32,33	29,33

Da bismo procenili uticaj različitih efekata, uočimo da se deo ukupne sume kvadrata grešaka odnosi na razlike između dve grupe. Da bismo to dobili suma kvadrata jedinki SS_i se mora izračunati za svaku grupu posebno. Za grupu 1 je

$$SS_i = 3[(25{,}25 - 24{,}75)^2 + (26 - 24{,}75)^2 + (23 - 24{,}75)^2] = 19{,}5,$$

a za grupu 2 je

$$SS_i = 3[(34 - 29{,}33)^2 + (27 - 29{,}33)^2 + (27 - 29{,}33)^2] = 130{,}7.$$

Ova dva rezultata sabiramo da dobijemo ukupnu sumu kvadrata grešaka $150{,}2$. Ukoliko se zanemari pripadnost grupi, ukupna suma kvadrata grešaka je $276{,}2$. Ovo znači da je među jedinkama suma kvadrata uzrokovana odstupanjima u grupama 126. Dalje se dobija da je ukupna suma kvadrata grešaka unutar subjekta (ispravljena za grupe) $373{,}17$. Ne uzimajući u obzir "različitost" grupa, suma kvadrata grešaka unutar subjekta je $399{,}96$. Razlika između njih je suma kvadrata koja se odnosi na interakcije između faktora unutar jedinki – vremena i faktora među jedinkama – grupe. Ova suma kvadrata iznosi $26{,}79$.

Razlikujemo sledeće efekte: ukupni efekat vremena (da li postoje razlike u očekivanim vrednostima promenljive Y na populaciji tokom vremena), ukupni efekat grupa (da li u proseku postoji razlika postoje razlike u očekivanim vrednostima promenljive Y na populaciji između grupa) i efekat interakcije vremena i grupa (da li je promena tokom vremena ishoda promenljive Y različita za poređene grupe). Ove efekte testiramo sledećim testovima:

- Test interakcije vremena i grupa (paralelizam)

$$H_0\colon (\tau\gamma)_{lj} = 0 \text{ za sve } j, l \text{ protiv } H_1\colon \text{bar jedno } (\tau\gamma)_{lj} \neq 0.$$

Test odbacuje H_0 na nivou značajnosti α ako

$$F_{GT} > \mathcal{F}_{(q-1)(n-1),(n-1)(m-q),\alpha}$$ [4]

ili ekvivalentno: verovatnoća da je vrednost test statistike veća ili jednaka od realizovane vrednosti F_{GT} je manja od α , ako je H_0 tačna (p vrednost je manja od α).

- Test glavnog efekta vremena (konstantnost)

$$H_0\colon \gamma_j = 0 \text{ za sve } j \text{ protiv } H_1\colon \text{bar jedno } \gamma_j \neq 0.$$

Test odbacuje H_0 na nivou značajnosti α ako

$$F_T > \mathcal{F}_{n-1,(n-1)(m-q),\alpha}$$

ili ekvivalentno: verovatnoća da je vrednost test statistike veća ili jednaka od F_T je manja od α , ako je H_0 tačna.

- Test glavnog efekta grupa (slučajnost)

$$H_0\colon \tau_l = 0 \text{ za sve } j \text{ protiv } H_1\colon \text{bar jedno } \tau_l \neq 0.$$

Test odbacuje H_0 na nivou značajnosti α ako

$$F_G > \mathcal{F}_{q-1,(m-q),\alpha}$$

ili ekvivalentno: verovatnoća da je vrednost test statistike veća ili jednaka od F_G je manja od α , ako je H_0 tačna.

[4] $\mathcal{F}_{a,b,\alpha}$ je kvantil reda 1-α Fišerove raspodele sa a i b stepeni slobode.

U SPSS-u se dobija:

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
vreme	Sphericity Assumed	224.792	3	74.931	2.410	.118
	Greenhouse-Geisser	224.792	1.459	154.083	2.410	.174
	Huynh-Feldt	224.792	2.658	84.583	2.410	.128
	Lower-bound	224.792	1.000	224.792	2.410	.196
vreme * grupa	Sphericity Assumed	26.792	3	8.931	.287	.834
	Greenhouse-Geisser	26.792	1.459	18.364	.287	.695
	Huynh-Feldt	26.792	2.658	10.081	.287	.812
	Lower-bound	26.792	1.000	26.792	.287	.620
Error(vreme)	Sphericity Assumed	373.167	12	31.097		
	Greenhouse-Geisser	373.167	5.836	63.947		
	Huynh-Feldt	373.167	10.631	35.103		
	Lower-bound	373.167	4.000	93.292		

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	17550.042	1	17550.042	467.482	.000
grupa	126.042	1	126.042	3.357	.141
Error	150.167	4	37.542		

3.4. Kontrasti

Hipoteze o interakciji grupa i vremena (paralelizam) i glavnim efektima vremena odgovaraju na pitanja o tome šta se događa tokom vremena, jer je vreme faktor unutar jedinki. Ove hipoteze odgovaraju na pitanja opšteg karaktera, kao na primer: da li je način promene očekivanih vrednosti rezultata tokom vremena različit za različite grupe (hipoteze o interakciji grupa i vremena).

Često nas zanima detaljnija analiza kako se očekivane vrednosti rezultata ponašaju tokom vremena. Prvo dajemo sledeću definiciju.

Ako je $\boldsymbol{c}$ $(n \times 1)$ vektor i $\boldsymbol{\mu}$ je $(n \times 1)$ vektor očekivanja, tada se linearna kombinacija

$$\boldsymbol{c}'\boldsymbol{\mu} = \boldsymbol{\mu}'\boldsymbol{c}$$

naziva *kontrast* ako je $\boldsymbol{c}$ takvo da je suma njegovih elemenata nula.

Hipoteze o razlikama očekivanih vrednosti možemo izraziti preko kontrasta. Konkretno, ako je $\boldsymbol{c}'\boldsymbol{\mu} = \boldsymbol{0}$, tada ne postoje razlike.

Na primer, neka je $q = 2$ i $n = 3$. Kontrasti

$$\mu_{11} - \mu_{12} \text{ i } \mu_{21} - \mu_{22} \tag{3.14}$$

upoređuju očekivane rezultate u prvoj i drugoj vremenskoj tački za svaku od dve grupe, slično,

$$\mu_{12} - \mu_{13} \text{ i } \mu_{22} - \mu_{23} \tag{3.15}$$

porede očekivane rezultate u drugoj i trećoj vremenskoj tački za svaku grupu. Dakle, kontrasti govore o tome kako se očekivanja razlikuju od jedne do druge vremenske tačke u svakoj grupi.

Podsetimo se,

$$\boldsymbol{\mu}_1' = [\mu_{11} \quad \mu_{12} \quad \mu_{13}], \qquad \boldsymbol{\mu}_2' = [\mu_{21} \quad \mu_{22} \quad \mu_{23}] .$$

Vidimo da su kontrasti u (3.14) rezultat naknadnog množenja ovih vektora očekivanja svake od grupa sa vektorom

$$\boldsymbol{c} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix};$$

Odnosno za (3.15) sa,

$$\boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}.$$

Pomoću kontrasta se mogu postaviti specijalna pitanja od interesa koja se odnose na to kako se očekivanja razlikuju od jednog do drugog vremenskog momenta.

- Može nas zanimati da li se očekivanja razlikuju, od npr. vremena 1 do vremena 2 u različitim grupama. Ovo je očigledno deo ukupne interakcije vremena i grupa, i to između vremena 1 i 2. U ovakvom slučaju, ako postoje dve grupe, interesovali bi nas kontrasti u (3.14).
- Takođe može da nas zanima da li je način na koji se očekivanja razlikuju u vremenu 2 i 3, različit između dve grupe. Ovo je takođe deo interakcije vremena i grupa i to je formalno predstavljeno preko razlika u kontrastima (3.15).
- Može da nas zanima da li postoje razlike u očekivanjima, u momentu 1 i 2, ako se posmatraju proseci po grupama. Ovo je deo glavnog efekta vremena i može se predstaviti kao prosek kontrasta u (3.14). Za vremena 2 i 3, zanimao bi nas prosek kontrasta u (3.15).

Navođenje ovih kontrasta, a zatim razmatranje njihovih razlika među grupama ili uprosečenih po grupama, jeste način da analiziramo kako efekti interakcije vremena i grupa i glavni efekat vremena nastaju i time dobijemo dodatni uvid o tome kako i da li se promenljiva menja tokom vremena.

Kontraste možemo predstaviti preko reprezentacije $\boldsymbol{CMU}$. Da dobijemo kontraste u (3.14) i (3.15), kada je $q = 2$ i $n = 3$, razmotrimo $n \times (n-1)$ matricu

$$\boldsymbol{U} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}.$$

A zatim,

$$\boldsymbol{MU} = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} \mu_{11} - \mu_{12} & \mu_{12} - \mu_{13} \\ \mu_{21} - \mu_{22} & \mu_{22} - \mu_{23} \end{bmatrix}. \tag{3.16}$$

Svaki element dobijene matrice je jedan od kontrasta. Ovaj izbor matrice kontrasta $\boldsymbol{U}$ daje kontraste koji imaju veze sa razlikama u očekivanjima od jednog vremena do sledećeg. Svaka kolona predstavlja moguće kontraste ovog tipa.

Primetimo da će se ista matrica $\boldsymbol{U}$ moći primeniti za veće q, važno je da ima $n-1$ kolonu od kojih se svaka sadrži jedno od $n-1$ moguća poređenja očekivanja od jednog vremenskog momenta do sledećeg. Za opšte n, matrica će imati sledeći oblik

$$\boldsymbol{U} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 1 \\ 0 & \cdots & 0 & -1 \end{bmatrix} \tag{3.17}$$

sa n vrsta i $n-1$ kolonom.

Množenje matrice $\boldsymbol{M}$ sa desne strane, opštim oblikom matrice kontrasta $\boldsymbol{U}$ u (3.17) se često naziva *profil transformacija* matrice očekivanja unutar jedinki.

Od interesa mogu biti i drugi kontrasti. Umesto da se pitamo šta se dešava od jednog vremena do sledećeg, možemo se pitati kako se očekivanje u bilo kom momentu razlikuje od onih u kasnijim vremenima. Ovo može da nam pomogne da razumemo u kom vremenskom momentu počinju da se dešavaju promene (ako ih ima).

Na primer, neka je $q = 2$ i $n = 4$ i posmatrajmo kontrast

$$\mu_{11} - \frac{\mu_{12} + \mu_{13} + \mu_{14}}{3}.$$

Ovaj kontrast poredi za grupu 1 očekivanje u vremenu 1 sa prosekom očekivanja u svim ostalim vremenima. Slično,

$$\mu_{12} - \frac{\mu_{13} + \mu_{14}}{2}$$

poredi u grupi 1 očekivanja u vremenu 2 sa prosekom svih sledećih vremenskih momenata. Poslednji kontrast ovog tipa za grupu 1 je

$$\mu_{13} - \mu_{14},$$

koji poredi šta se dešava u vremenu 3 sa "prosekom" onoga šta sledi, a to je jedno očekivanje u vremenu 4. Možemo posmatrati odgovarajuće kontraste i za drugu grupu. Možemo izraziti sve takve kontraste sa drugačijom matricom $\boldsymbol{U}$. Neka je

$$\boldsymbol{U} = \begin{bmatrix} 1 & 0 & 0 \\ -1/3 & 1 & 0 \\ -1/3 & -1/2 & 1 \\ -1/3 & -1/2 & -1 \end{bmatrix}, \tag{3.18}$$

Tada, ako je $q = 2$,

$$\boldsymbol{MU} = \begin{bmatrix} \mu_{11} - \mu_{12}/3 - \mu_{13}/3 - \mu_{14}/3 & \mu_{12} - \mu_{13}/2 - \mu_{14}/2 & \mu_{13} - \mu_{14} \\ \mu_{21} - \mu_{22}/3 - \mu_{23}/3 - \mu_{24}/3 & \mu_{22} - \mu_{23}/2 - \mu_{24}/2 & \mu_{23} - \mu_{24} \end{bmatrix},$$

što predstavlja sve takve kontraste. Prva vrsta daje kontraste za grupu 1.

Uopšteno, $n \times (n-1)$ matrica čije kolone definišu kontraste ovog tipa naziva se *Helmert transformacija* matrice oblika

$$\boldsymbol{U} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1/(n-1) & 1 & 0 & \cdots & 0 \\ -1/(n-1) & -1/(n-2) & 1 & \cdots & 0 \\ \vdots & \vdots & -1/(n-3) & \vdots & \vdots \\ -1/(n-1) & -1/(n-2) & \vdots & \cdots & 1 \\ -1/(n-1) & -1/(n-2) & -1/(n-3) & \cdots & -1 \end{bmatrix}. \tag{3.19}$$

Množenje matrice $\boldsymbol{M}$ sa desne strane matricom oblika (3.19) predstavlja poređenja svakog očekivanja sa prosekom očekivanja u svim kasnijim vremenima. Za $q = 2$ i $n = 3$, jednostavno se dobija da ova transformacija vodi do

$$\boldsymbol{MU} = \begin{bmatrix} \mu_{11} - \mu_{12}/2 - \mu_{13}/2 & \mu_{12} - \mu_{13} \\ \mu_{21} - \mu_{22}/2 - \mu_{23}/2 & \mu_{22} - \mu_{23} \end{bmatrix}. \tag{3.20}$$

Sa "prostim" kontrastom vrednost svakog merenja se poredi sa prvim merenjem; sa kontrastom "razlika" vrednosti svakog merenja se porede sa prosekom svih prethodnih merenja; "Helmert" kontrast poredi merenja sa prosekom svih sledećih merenja; "ponovljeni" kontrast vrednosti svakog merenja poredi sa vrednostima prvog sledećeg merenja.

Kontraste možemo primeniti pri sledećim testovima:

1. Sveukupni testovi.

Videli smo primenu $\boldsymbol{CMU}$ reprezentacije za sveukupne testove interakcija grupa i vremena i glavne efekte vremena. I kontrast matrica $\boldsymbol{U}$ u (3.16) (profil) i matrica u (3.20) (Helmert) sadrže skup od $n - 1$ kontrasta koji predstavljaju sve moguće razlike u očekivanjima tokom vremena na različite načine. Intuitivno možemo očekivati da bilo koja od njih dovodi do sveukupnih testova o interakciji grupa i vremena i o glavnom efektu vremena za odgovarajuću matricu $\boldsymbol{C}$ (koja predstavlja razlike između grupa ili proseke grupa, redom). Može se pokazati da množenje s leve stranice bilo koje od matrica (3.16) i (3.20) istom matricom $\boldsymbol{C}$ vodi do istih sveukupnih hipoteza izraženih komponentama γ_j i $(\tau\gamma)_{lj}$. Na primer, videli smo da množenjem (3.16) s leve strane matricom $\boldsymbol{C} = (1,1)$, uz ograničenja $(\tau\gamma)_{lj}$, dobijamo

$$\boldsymbol{CMU} = [\gamma_1 - \gamma_2 \quad \gamma_2 - \gamma_3] = \boldsymbol{0}.$$

Može se pokazati da množenjem (3.20) s leve strane istom matricom $\boldsymbol{C}$ dobijamo

$$\boldsymbol{CMU} = [\gamma_1 - 0{,}5\gamma_2 - 0{,}5\gamma_3 \quad \gamma_2 - \gamma_3] = [0 \quad 0].$$

U oba slučaja sledi isti zaključak, da testiramo $\gamma_1 = \gamma_2 = \gamma_3$.

Ovo pokazuje da izbor matrice kontrasta $\boldsymbol{U}$ nije bitan za sveukupne testove glavnih efekata vremena i interakcija grupa i vremena. Izbor matrice $\boldsymbol{U}$ je bitan kada nas interesuje analiziranje ovih sveukupnih efekata.

Vratimo se na to kako da predstavimo hipoteze i sprovedemo testove za pitanja kao što su ranije navedena za datu kontrast matricu $\boldsymbol{U}$. Množenjem matrice $\boldsymbol{U}$ s leve strane matricom $\boldsymbol{M}$ dobijamo $q \times (n-1)$ matricu $\boldsymbol{MU}$ čija l-ta vrsta sadrži kontraste koji su od interesa (oni su određeni kolonama matrice $\boldsymbol{U}$) za grupu l. Ukoliko množimo $\boldsymbol{MU}$ s leve strane, $q \times (n-1)$ matricom

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & -1 \end{bmatrix}$$

(ranije smo razmatrali specijalni slučaj za $q = 2$), tada za svaki kontrast definisan u $\boldsymbol{U}$, razmatramo kako se kontrast razlikuje među grupama. Kontrast obuhvata specifičan način na koji se očekivani rezultati razlikuju u različitim vremenima, kao što je komponenta interakcija grupa i vremena (kako se način promene prosečnih vrednosti među grupama razlikuje u različitim vremenima).

Množenjem s leve strane matricom $\boldsymbol{C} = [1/q \quad 1/q \quad \cdots \quad 1/q]$ svaki od $n-1$ elementa dobijene $1 \times (n-1)$ matrice će odgovarati proseku svakog od ovih kontrasta po

grupama, a svi zajedno čine glavni efekat vremena. Ako razmatramo jedan od ovih elemenata, videćemo da on predstavlja kontrast koji poredi očekivanje rezultata u vremenu j sa prosekom očekivanja rezultata u svim vremenima posle j, po grupama. Ako je taj kontrast jednak nuli, očekivani rezultati u vremenu j su jednaki naknadnim prosečnim očekivanim rezultatima.

S druge strane, možemo posmatrati posebno svaki kontrast da bismo ispitali neke aspekte ponašanja očekivanih vrednosti tokom vremena.

2. Posebni testovi.

Testiranje posebnih hipoteza za svaki kontrast u matrici $\boldsymbol{U}$ može se vršiti na sledeći način. Razmotrimo k - tu kolonu matrice $\boldsymbol{U}$, $\boldsymbol{c}_k$, $k = 1, \dots, n-1$. Primenjujujemo funkciju određenu tom kolonom matrice $\boldsymbol{U}$ na svaki vektor ponovljenih merenja svake jedinke. Dakle, za svaki vektor $\boldsymbol{Y}_{hl}$, dobijamo

$$\boldsymbol{y}'_{hl}\boldsymbol{c}_k = \boldsymbol{c}'_k\boldsymbol{Y}_{hl}.$$

Ovo svodi ponovljena merenja određene jedinke na *jedan broj* koji predstavlja vrednosti kontrasta za tu jedinku. Ako svaki vektor podataka jedinke ima istu kovarijansnu matricu $\boldsymbol{\Sigma}$, tada svaka takva "svedena" vrednost podataka ima istu varijansu za sve jedinke.

Da bismo testirali da li je kontrast jednak nuli kada se traži prosek po grupama, testiramo da li je sveukupno očekivanje podataka jednako nuli, koristeći standardni t test (ili ekvivalentno, F test zasnovan na kvadratu t statistike).

Ovi testovi će biti dobri bez obzira da li važi složena simetrija; jedino je važno da je $\boldsymbol{\Sigma}$ ista za sve jedinke. Varijansa svedenih vrednosti $\boldsymbol{c}'_k\boldsymbol{Y}_{hl}$ za h- tu jedinku l-te grupe je

$$var(\boldsymbol{c}'_k\boldsymbol{Y}_{hl}) = \boldsymbol{c}'_k\boldsymbol{\Sigma}\boldsymbol{c}_k.$$

i konstantna je za sve h i l dogod je $\boldsymbol{\Sigma}$ ista. Uobičajena pretpostavka konstantne varijanse koja je neophodna za jednostranu analizu varijanse ispunjena je za podatke koji odgovaraju svakom kontrastu.

3.4.1. Ortogonalni kontrasti

U nekim slučajevima možemo da primetimo da kontrasti koji se pojavljuju u ovim transformacionim matricama, imaju dodatnu osobinu. Ako su $\boldsymbol{c}_1$ i $\boldsymbol{c}_2$ dve kolone matrice, takve da je

$$\boldsymbol{c}'_1\boldsymbol{c}_2 = 0;$$

tj. suma proizvoda koji odgovara elementima te dve kolone je nula, kaže se da su vektori $\boldsymbol{c}_1$ i $\boldsymbol{c}_2$ *ortogonalni*. Kontrasti koji odgovaraju ovim vektorima su ortogonalni kontrasti[5].

Razmotrimo kakva je prednost transformacija kada su kontrasti ortogonalni. Za skup ortogonalnih kontrasta, posebni testovi imaju lepu osobinu koju nemaju skupovi kontrasta koji nisu ortogonalni. Kao što je intuitivno, ako su kontrasti ortogonalni, treba da podele ukupnu interakciju grupa i vremena kao i sumu kvadrata grešaka unutar jedinki na $n-1$ različitih ili "nepreklapajućih" komponenti. Dakle, rezultat jednog od takvih testova može biti sagledan bez obzira na ishod ostalih. Može se pokazati da ako radimo sa odgovarajućom "normalizovanom" verzijom matrice $\boldsymbol{U}$, čije su kolone ortogonalne, tada se ova osobina može jasno uočiti. Specijalno, sume kvadrata za grupe u pojedinačnim ANOVA testovima za

[5]Napomena: Kontrasti koji čine profil transformaciju nisu ortogonalni, a kontrasti koji koji čine Helmert transformaciju jesu ortogonalni.

svaki kontrast se sabiraju do sume kvadrata SS_{GT}. Analogno, suma kvadrata grešaka se sabira do sume kvadrata SS_E.

Razmotrimo Helmert matricu u (3.18),

$$\boldsymbol{U} = \begin{bmatrix} 1 & 0 & 0 \\ -1/3 & 1 & 0 \\ -1/3 & -1/2 & 1 \\ -1/3 & -1/2 & -1 \end{bmatrix}.$$

Svaka kolona odgovara jednoj funkciji koja se primenjuje na vektor podataka svake jedinke tj., k- ta kolona opisuje k- ti kontrast funkcije $\boldsymbol{c}_k'\boldsymbol{Y}_{hl}$ vektora podataka. Konstante koje čine svaki vektor $\boldsymbol{c}_k$ su različite za svako k. Dakle, vrednosti $\boldsymbol{c}_k'\boldsymbol{Y}_{hl}$ za svako k su na različitim skalama merenja, nisu uporedive po svakom od $n-1$ kontrasta, pa ni sume kvadrata svake pojedinačne ANOVA-e nisu uporedive, jer svaka ima podatke na različitim mernim skalama.

Moguće je modifikovati svaki kontrast (bez uticanja na osobinu ortogonalnosti) tako da rezultati podataka budu na istim skalama merenja. Primetimo da su sume kvadrata elemenata svake kolone različite, tj. sume kvadrata prve, druge i treće kolone su

$$1^2 + (-1/3)^2 + (-1/3)^2 + (-1/3)^2 = 4/3,$$

$3/2$ i 2, redom. Ovo pokazuje da kontrasti zaista nisu na istim skalama, te je potrebna modifikacija.

Množimo svaki kontrast odgovarajućom konstantom da je suma kvadrata elemenata jednaka 1.

U našem primeru, ako množimo prvu kolonu sa $\sqrt{3/4}$, drugu sa $\sqrt{2/3}$, a treću sa $\sqrt{1/2}$, tada se dobija da je suma kvadrada modifikovanih elemenata 1, na primer:

$$\left(\sqrt{3/4}\cdot 1\right)^2 + \left(\sqrt{3/4}\,(-1/3)\right)^2 + \left(\sqrt{3/4}\,(-1/3)\right)^2 + \left(\sqrt{3/4}\,(-1/3)\right)^2 = 1.$$

Množenje svakog kontrasta konstantom ne menja odgovarajuće hipoteze; na primer, za prvu kolonu, testiranje hipoteze

$$H_0: \mu_{11} - \mu_{12}/3 - \mu_{13}/3 - \mu_{14}/3 = 0.$$

je isto kao testiranje hipoteze

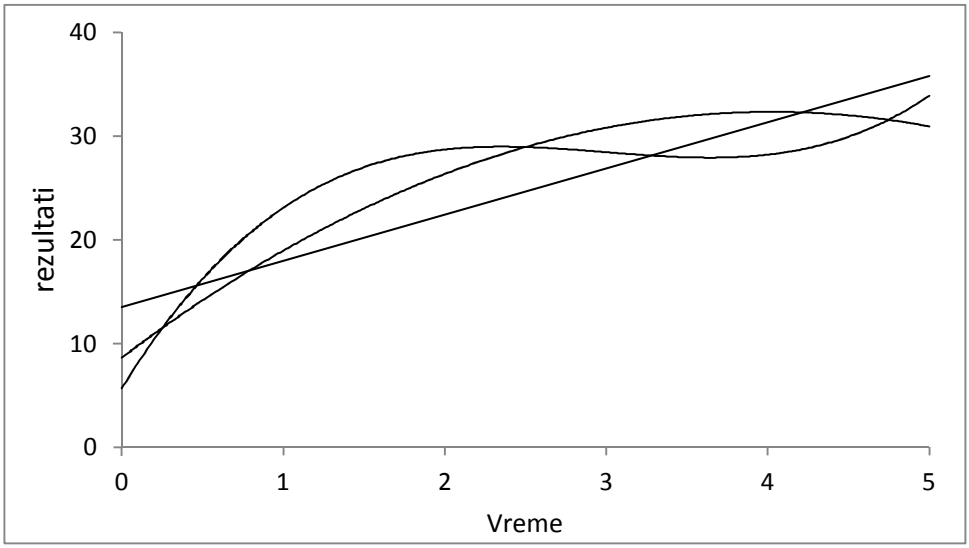
$$H_0: \sqrt{3/4}\,\mu_{11} - \sqrt{3/4}\,\mu_{12}/3 - \sqrt{3/4}\,\mu_{13}/3 - \sqrt{3/4}\,\mu_{14}/3 = 0.$$

Kada su svi kontrasti u ortogonalnoj transformaciji isto sklirani, tada se kaže da su oni ortonormirani.

Ako kontrasti nisu ortogonalni, tumačenje posebnih testova je teže jer odvojeni testovi nisu više "nepreklapajući". Ukupna suma kvadrata za interakciju grupe i vremena nije više podeljena kao ranije. Dakle, rezultat koji daje jedan test je povezan sa onim koji daje drugi.

Zajednička karakteristika longitudinalnih podataka jeste da izgleda da svaka jedinka ima glatku vremensku trajektoriju. U nekim slučajevima trajektorije su prave linije. U drugim slučajevima trajektorije su krive. Dakle, ako razmatramo trajektoriju jedne jedinke, razumno je da razmišljamo o njoj kao o linearnoj, kvadratnoj, kubnoj ili uopšteno o polinomnoj funkciji vremena. Na slici 6 su primeri takvih trajektorija.

Slika 6. Polinomne trajektorije (linearna, kvadratna i kubna).



Polinomni kontrasti predstavljaju specijalni skup ortogonalnih kontrasta koji testiraju polinomni oblik podataka. Ako imamo merenja svake jedinke ponovljena u n perioda, tada je u principu moguće imati najviše polinom stepena $n-1$. Tada je, moguće definisati $n-1$ ortogonalni polinomni kontrast koji meri jačinu linearnog, kvadratnog, kubnog polinoma $n-1$ stepena. Ovo je moguće za vremenske tačke koje su i na jednakim međusobnim rastojanjima i na nejednakim rastojanjima tokom vremena. Za jednaka vremenska rastojanja između ponovljenih merenja, koeficijenti $n-1$ ortogonalnog polinoma su dostupni u tabelama mnogih statističkih tekstova, dok za nejednaka rastojanja između merenja, proračuni zavise od međusobne udaljenosti merenja.

3.5 Narušavanje pretpostavke o obliku kovarijansne matrice i prilagođeni testovi

Naglasili smo da postupci koji se temelje na analizi varijanse važe samo ako važi pretpostavka složene simetrije za kovarijansnu matricu vektora podataka. U stvarnosti ti postupci važe i sa opštijim uslovima. Važan zahtev je da kovarijansna matrica mora biti posebnog oblika, a ako to nije zadovoljeno navedeni testovi će biti neodgovarajući i mogu dovesti do pogrešnih zaključaka. Odnosno, F količnici F_T i F_{GT} više neće imati Fišerovu F raspodelu.

Matica $\mathbf{\Sigma}$ dimenzija $(n \times n)$ se naziva tipa H ako se može predstaviti u sledećem obliku:

$$\mathbf{\Sigma} = \begin{bmatrix} \lambda + 2\alpha_1 & \alpha_1 + \alpha_2 & \cdots & \alpha_1 + \alpha_n \\ \alpha_2 + \alpha_1 & \lambda + 2\alpha_2 & \dots & \alpha_2 + \alpha_n \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_n + \alpha_1 & \alpha_n + \alpha_2 & \dots & \lambda + 2\alpha_n \end{bmatrix}.$$

Može se proveriti da matrica koja zadovoljava složenu simetriju je tipa H.

Moguće je pokazati da će razmatrani F testovi biti opravdani sve dok vektori podataka $\boldsymbol{Y}_i$ imaju multivarijantnu normalnu raspodelu sa zajedničkom kovarijansnom matricom $\boldsymbol{\Sigma}$ tipa H. Ukoliko kovarijansna matrica $\boldsymbol{\Sigma}$ nije tipa H, a F testovi se ipak sprovode, biće previše „slobodni“, jer će težiti da odbace nulte hipoteze češće nego što bi trebalo. Dakle, jedna od mogućih posledica korištenja postupka analize varijanse kada nije baš primereno jeste zaključak da interakcija grupa i vremena postoji kada to nije tačno.

Pretpostavimo da imamo razloga da sumnjamo da je kovarijansna matrica podataka $\boldsymbol{\Sigma}$ tipa H. To se može desiti ako ne verujemo da je uslov sferičnosti ispunjen i odbacili smo nultu hipotezu da je $\boldsymbol{\Sigma}$ tipa H. Dakle, dovodimo u pitanje pretpostavku o H obliku kovarijansne matrice, a time i da ne važi složena simetrija, naš model zahteva. Tada su uobičajeni F testovi za interakciju vremena i grupa su neodgovarajući. Postoji nekoliko načina prilagođavanja uobičajenih F testova.

Definišimo koeficijent sferičnosti

$$\varepsilon = \frac{tr^2(\boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U})}{(n-1)tr(\boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U}\boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U})},$$

gde je $\boldsymbol{U}$ $(n \times (n-1))$ $(u = n-1)$ matrica sa kolonama koje su normalizovani ortogonalni kontrasti. Može se pokazati da tako definisana konstanta ε zadovoljava

$$\frac{1}{n-1} \leq \varepsilon \leq 1$$

i važi da je $\varepsilon = 1$ ako i samo ako $\boldsymbol{\Sigma}$ ima H oblik. Dakle, u idealnoj situaciji (zadovoljena sferičnost) koeficijent sferičnosti je jednak 1, a ako pretpostavka nije u potpunosti ispunjena koeficijent će biti manji od jedan. U ovom slučaju se stepeni slobode F – testa mogu promeniti tako da a, b stepene slobode zamenimo sa $\varepsilon a,\ \varepsilon b$ stepena slobode brojioca i imenioca. To će učiniti stepene slobode manjim nego što je uobičajeno. Kritičnie vrednosti F raspodele postaju veće što su stepeni slobode brojioca i imenioca manji, Dakle, efekat ovog prolagođavanja jeste poređenje realizovanih vrednosti F količnika sa većim kritičnim vrednostima, pa je teže odbaciti nultu hipotezu, i testovi postaju “liberalniji”.

Koeficijent ε je nepoznat, jer zavisi od nepoznate $\boldsymbol{\Sigma}$ matrice. Moguće je oceniti matricu $\boldsymbol{\Sigma}$ iz uzorka, pa na osnovu toga dati ocenu za ε. Dve ocene su poznate kao *Greenhouse-Geisser* i *Huynh-Feldt* korekcije. Svaka od njih procenjuje ε na različite načine. Moguće je kao ocenu uzeti i prosek ovih prilagođavanja. Dakle, ove korekcije nam služe da smanjimo grešku prve vrste. F pokazatelj ne menja rezultat primenom ovih korekcija, već se menjaju samo stepeni slobode. U SPSS-u su navedene dve korekcije, kao i *korekcija niže granice*. Kada je $\varepsilon > 0{,}75$ Greenhouse-Geisser korekcija govori da će doći do pogrešnog odbacivanja nulte hipoteze. Predloženo je da ukoliko je $\varepsilon > 0{,}75$ da se primeni Huynh-Feldt korekcija, a kada je $\varepsilon < 0{,}75$ ili se ništa ne zna o složenoj simetriji, da se primeni Greenhouse-Geisser korekcija.

Može se testirati značajnost koeficijenta sferičnosti (nulta hipoteza bi bila da je $\epsilon = 1$). Ukoliko je uzorak jako veliki, test sferičnosti će najčešće davati značajan rezultat, dok u studiji sa malim uzorkom test najčešće neće davati značajan rezultat. U prvoj situaciji test je precenjen, što znači da će se uočiti čak i malo narušavanje pretpostavke sferičnosti, dok u slučaju sa malim uzorkom, test je podcenjen tj. snaga uočavanja narušavanja pretpostavke sferičnosti je vrlo niska.

Mauchly-ev test sferičnosti testira da li je zadovoljena pretpostavka sferičnosti odnosno testira nultu hipotezu da su jednake varijanse razlika između svih rezultata u različitim vremenskim momentima. Kada je vrednost Mauchly-eve test statistike veća ili jednaka od 0,05 ne odbacujemo nultu hipotezu, pa se zaključuje da je pretpostavka

zadovoljena. Međutim, ukoliko je ta vrednost manja od 0,05 ne može se pretpostaviti složena simetrija, odnosno postoje značajne razlike između varijansi.

Kada je utvrđena složena simetrija, a time i sferičnost, F test je validan. Ukoliko uslov sferičnosti nije ispunjen radi se korekcija za broj stepena slobode F raspodele, čime se dobija tačnija F vrednost. F pokazatelj se mora tumačiti sa rezervom, jer narušavanje ove pretpostavke može dovesti do povećanja greške prvog tipa i utiče za zaključke analize.

U sledećoj tabeli prikazan je Mauchly-ev test sferičnosti za podatke u primeru na strani 39.

Mauchly's Test of Sphericity[b]

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
vreme	.112	8.165	5	.160	.486	.634	.333

Mauchly-ev test je jedan od najčešće korišćenih testova za procenu sferičnosti, ali ovaj test slabo detektuje nedostatak ove osobine u malim uzorcima, a u velikim uzorcima previše naglašava nedostatak sferičnosti. Možemo zaključiti da veličina uzorka ima uticaj na analizu podataka.

Dakle, statistički model koji smo posmatrali zasniva se na pretpostavci složenoj simetriji, odnosno o povezanosti posmatranja na istoj jedinki. Kada je ova pretpostavka zadovoljena, onda se koriste poznate metode analize varijanse. Ova pretpostavka se može testirati, ali testovi nisu pozdani. U slučaju da nismo sugurni da je ispunjena pretpostavka o složenoj simetriji, na raspolaganju su približne. Međutim, ove korekcije takođe nisu pouzdane. To sugeriše da umesto da "namećemo" uslov složene simetrije, bolji je da počnemo ispočetka sa realnijim statističkim modelom. U nastavku razmatraćemo i druge modele za analizu longitudinalnih podataka koji ne traže pretpostavku složene simetrije (ili generalno da je matrica $\boldsymbol{\Sigma}$ tipa H).

4. Multivarijantna analiza varijanse sa ponovljenim merenjima

Statistički model koji je u osnovi univarijantne analize varijanse sa ponovljenim merenjima koji smo razmatrali u Glavi 3 uključuje veoma restriktivne pretpostavke o obliku kovarijansne matrice vektora podataka. Ponovimo, ako je $\boldsymbol{Y}_i$ vektor podataka posmatranja u n vremenskih trenutaka za i- tu jedinku, model možemo predstaviti u obliku

$$\boldsymbol{Y}_i' = \boldsymbol{a}_i'\boldsymbol{M} + \boldsymbol{\varepsilon}_i', i = 1, \dots, m. \tag{4.1}$$

gde su $\boldsymbol{a}_i'$ i $\boldsymbol{M}$ ranije definisani u (3.10) i (3.11), kao $(1\times q)$ vektor pokazatelj pripadnosti grupi i $(q \times n)$ matrica čije su vrste transponovani vektori očekivanja svake grupe. Vektor greške $\boldsymbol{\varepsilon}_i$ pridružen i - toj jedinki, zbog načina na koji je konstruisan model, ima kovarijansnu matricu

$$\boldsymbol{\Sigma} = \sigma_b^2\boldsymbol{J}_n + \sigma_e^2\boldsymbol{I}_n;$$

to jest model implicira pretpostavku složene simetrije. Sa pretpostavkom o normalnoj raspodeli, model takođe implicira da svaki vektor podataka ima multivarijantnu normalnu raspodelu:

$$\boldsymbol{Y}_i \sim \mathcal{N}_n(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \qquad \boldsymbol{\mu}_i = \boldsymbol{a}_i'\boldsymbol{M}.$$

Elementi $\boldsymbol{\mu}_i$ u ovom modelu imaju specifičnu formu; ako je jedinka i iz l- te grupe, j – ti element ovog vektora, $j = 1, \dots, n$, ima oblik

$$\mu + \tau_l + \gamma_j + (\tau\gamma)_{lj}.$$

Testove zasnovane na poznatim tehnikama analize varijanse možemo koristiti dok važi pretpostavka složene simetrije. Od velikog interesa je test postojanja interakcije grupa i vremena, kojim se proverava da li se promena očekivane vrednosti rezultata razlikuje među grupama tokom vremena (paralelizam). Sve dok su pretpostavke o složenoj simetriji i normalnoj raspodeli zadovoljene, uobičajena test statistika zasnovana na količniku dve sume kvadrata ima Fišerovu raspodelu, pa se vrednost statistike može uprediti sa odgovarajućim kvantilom pri sprovođenju testa. Međutim, ako pretpostavka složene simetrije nije zadovoljena, primena ovoga testa može dovesti do pogrešnih zaključaka. Jedan pristup, predstavljen u Glavi 3, za rešavanje ovog problema su "prilagođeni" testovi. Međutim, takav pristup nije zadovoljavajući, jer može prikriti stvarni problem, da pretpostavka složene simetrije nije odgovarajuća. Činjenica je da je ova pretpostavka previše restriktivna da okarakteriše korelacije koje se pojavljuju kod longitudinalnih podataka. Dakle, bolja alternativa od "prilagođenih" testova je da se napravi statistički model sa manje restriktivnom pretpostavkom i da se razviju nove odgovarajuće procedure za model pod tom pretpostavkom.

Najopštija alternativa za složenu simetriju je da se pretpostavi vrlo malo o prirodi kovarijansne strukture vektora podataka. Podsetimo se da je odstupanje $\boldsymbol{\varepsilon}_i'$ imalo poseban oblik:

$$\boldsymbol{\varepsilon}_i' = \boldsymbol{1}'b_i + \boldsymbol{e}_i',$$

što je impliciralo strukturu složene simetrije. Alternativni pristup je da posmatramo model (4.1) kao polaznu tačku i postavimo uslove direkno za kovarijansnu strukturu povezanu sa $\boldsymbol{\varepsilon}_i'$. Možemo i dalje smatrati da je kovarijansna matrica vektora podataka $\boldsymbol{Y}_i$ ista za sve i, bez

obzira na pripadnost određenoj grupi, ali ne moramo smatrati da matrica zadovoljava strukturu složene simetrije. Predstavimo to formalno preko modela

$$\boldsymbol{Y}_i' = \boldsymbol{a}_i'\boldsymbol{M} + \boldsymbol{\varepsilon}_i',\ i = 1, \dots, m, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \tag{4.2}$$

gde je sada $\boldsymbol{\Sigma}$ proizvoljna kovarijansna matrica za koju se ne pretpostavlja da ima bilo kakvu određenu strukturu. Sve što tražimo je da $\boldsymbol{\Sigma}$ bude simetrična matrica nestruktuiranog oblika

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

I da ista je za sve i.

- Ovakav model ne prepoznaje kako izvori varijacije unutar jedinki i među jedinkama doprinose ukupnoj varijaciji vektora podataka. Umesto toga, pretpostavlja se da se spajanjem oba izvora varijacije dobija proizvoljna kovarijansna struktura nestuktuiranog oblika. Na pretpostavlja se ništa o tome kako se ova dva izvora varijacije kombinuju.
- Dobijena nestruktuirana matrica zavisi od $n(n+1)/2$ parametra (a ne od dva parametra σ_b^2 i σ_e^2 kao kod pretpostavke o složenoj simetriji). Tako je potrebno mnogo više parametara da bi se opisalo kako merenja u vektoru podataka mogu varirati i kovarirati.

Multivarijantni postupci: Polazeći od modela datog u (4.2) moguće je razviti dobre postupke za testiranje hipoteza koje nas interesuju. Međutim, model je komplikovanji jer nemamo više lepu i jednostavnu pretpostavku o kovarijansama. Više nije moguće da model gradimo na osnovu pojedinačnog merenja, pa zbog toga ne možemo da dobijemo, kao ranije lepe rezultate o raspodeli količnika sredina sume kvadrata, pa se poznata procedura zasnovana na *F* količnicima ne može više koristiti.

Sada je neophodno da podatke posmatramo u obliku vektora. Zbog toga su postupci koje ćemo sada razmatrati su poznati kao multivarijantna analiza varijanse sa ponovljenim merenjima (MANOVA). Podaci nisu, kao do sada, skalari, nego vektori.

Možemo reći da je multivarijantna analiza varijanse skup statističkih metoda koje simultano analiziraju višedimenziona merenja dobijena za svaku jedinku iz skupa objekata koje ispitujemo. Dakle, multivarijantna analiza varijanse (MANOVA) je statitička procedura poređenja očekivanih vrednosti nekoliko grupa koristeći varijansu i kovarijansu između promenljivih.

4.1. Opšti multivarijantni problem

Razmatramo opšti slučaj multivarijantnog modeliranja longitudinalnih podataka. Koristimo notaciju sa dva indeksa radi praktičnosti.

- Jedinke su nasumice raspoređene u q grupa.
- Vektor podataka $\boldsymbol{Y}_{hl}$ je rezultat za h- tu jedinku iz l- te grupe.
- Pretpostavlja se da $\boldsymbol{Y}_{hl}$ zadovoljava

  $$\boldsymbol{Y}_{hl} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}),$$

  gde je $\boldsymbol{\mu}_l$ vekor očekivanih vrednosti za l- tu grupu, a $\boldsymbol{\Sigma}$ je proizvoljna kovarijansna matrica, za koju se pretpostavlja da je ista za svaku grupu.
- Ima r_l jedinki u svakoj grupi, tj. za svaku grupu l važi, $h = 1, \dots, r_l$.

- Ne mora se pretpostaviti da su komponente $\boldsymbol{Y}_{hl}$ rezultati merenja istih karakteristika. Umesto toga, svaka komponenta $\boldsymbol{Y}_{hl}$ može predstavljati merenje različite karakteristike. Može se meriti n različitih karakteristika i prikazati preko vektora $\boldsymbol{Y}_{hl}$. Na primer, karakteristike: y_{hl1} - visina, y_{hl2} - obim struka, y_{hl3} – BMI itd.
- Longitudinalni podaci su specijalan slučaj ovog modela: $\boldsymbol{Y}_{hlj}$ su merenja iste karakteristike tokom vremena.

Naravno, najviše nas interesuje poređenje grupa na osnovu dobijenih podataka.

- Kod univarijantnih metoda primetili smo da kada su sva merenja u vektoru podataka ista, merenja iste karakteristike, prirodni pristup je da razmišljamo o pravljenju proseka rezultata tokom vremena i poređenju tih proseka. Tako su interpretirane hipoteze razvijene za testiranje glavnog efekata grupa. (Ovo može da bude problem ako profili nisu paralelni).
- Jasno je da bi ovde pravljenje proseka za sve rezultate i poređenje proseka grupa bilo besmisleno. Za prethodni primer, pravili bismo prosek visina, obima struka, BMI, itd., dakle promenjivih koje mere potpuno različite karakteristike na različitim skalama.
- Dakle, najbolje čemu se možemo nadati jeste da na određen način uporedimo sve različite rezultate "istovremeno". Pri tome, prirodno je da se uzme u obzir da su rezultati dobijeni na istoj jedinki u korelaciji.

U našem statističkom modelu $\boldsymbol{\mu}_l$ je vektor očekivanih vrednosti (sastavljen od n različitih rezultata) dobijen za jedinke l- te grupe. Formalno, našu želju je da uporedimo n rezultata "istovremeno" možemo da formulišemo kao želji da uporedimo q vektora očekivanih vrednosti $\boldsymbol{\mu}_l, l = 1, \dots, q$, na osnovu svih njihovih komponenti. Zainteresovani smo za testiranje nulte hipoteze

$$H_0: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q \tag{4.3}$$

protiv alternativne da H_0 nije tačno. Sve dok se n rezultata u vektoru podataka razlikuju i nisu uporedivi (npr. Ne može im se tražiti prosečna vrednost) ovo je najbolje što možemo uraditi a da se odnosi na naše opšte pitanje.

4.2. Hotellingova T^2 statistika

Standardni metodi za testiranje nulte hipoteze (4.3) su generalizacije standardnih metoda u slučaju gde su podaci svake jedinke samo skalari y_{hl}, pa je tada $\boldsymbol{Y}_{hl}$ vektor dužine $n = 1$.

Razmotrimo slučaj kada postoje $q = 2$ grupe.

Skalarni slučaj: Ako su rezultati skalari, a ne vektori, interesuje nas poređenje dva skalarna očekivanja $\mu_l, l = 1,2$, pa se hipoteza H_0 svodi na

$$H_0: \mu_1 = \mu_2 \text{ ili } \mu_1 - \mu_2 = 0.$$

Dalje, nepoznata kovarijansna matrica $\boldsymbol{\Sigma}$ bi se svela na skalarnu varijansu σ^2. Pod pretpostavkom normalne raspodele, standardni test za H_0 bio bi t test za testiranje dve očekivane vrednosti.

- Kako je σ^2 nepoznata, mora se oceniti. Ovo se postiže ocenjujući σ^2 na osnovu rezultata u svakoj grupi, a zatim ujedinjavanjem rezultata. Ako $\bar{Y}_{\cdot l}$ označava

aritmetičku sredinu uzorka za r_l posmatranja y_{hl} za l- tu grupu, nalazimo varijansu uzorka

$$S_l^2 = (r_l - 1)^{-1} \sum_{h=1}^{r_l} (Y_{hl} - \bar{Y}_{\cdot l})^2$$

i formiramo ocenu za σ^2 iz podataka obe grupe kao "ponderisani prosek"

$$S^2 = (r_1 + r_2 - 2\,)^{-1}\big((r_1 - 1)S_1^2 + (r_2 - 1)S_2^2\big).$$

Sada, formiramo test statistiku

$$t = \frac{\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2}}{\sqrt{(r_1^{-1} + r_2^{-1})S^2}}.$$

Ova statistika t ima Studentovu t raspodelu sa $r_1 + r_2 - 2$ stepena slobode.

Multivarijantni slučaj: Sada se testira hipoteza

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ ili } \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}. \tag{4.4}$$

Potrebno je naći multivarijantni analogon *t* testa.

U skalarnom slučaju smo pretpostavljali da grupe imaju zajedničku varijansu σ^2, a sada pretpostavljamo zajedničku kovarijansnu matricu $\boldsymbol{\Sigma}$ (koja je nepoznata) i želimo proceniti ovu matricu za svaku grupu, a zatim objediniti rezultate.

Specijalno, možemo izračunati objedinjenu kovarijansnu matricu uzorka. Ako prikažemo aritmetičke sredine uzorka $\bar{\boldsymbol{Y}}_{\cdot lj}, j = 1, \dots, n$ u obliku vektora

$$\bar{\boldsymbol{Y}}_{\cdot l} = \begin{bmatrix} \bar{y}_{\cdot l1} \\ \vdots \\ \bar{y}_{\cdot ln} \end{bmatrix}, \text{ odnosno } \quad \bar{\boldsymbol{Y}}_{\cdot l} = \frac{1}{r_l} \sum_{l=1}^{r_l} \boldsymbol{Y}_{hl}$$

Tada je uzoračka kovarijansna matrica za l- tu grupu $(n \times n)$ matrica

$$\hat{\boldsymbol{\Sigma}}_l = (r_l - 1)^{-1} \sum_{h=1}^{r_l} (\boldsymbol{Y}_{hl} - \bar{\boldsymbol{Y}}_{\cdot l})(\boldsymbol{Y}_{hl} - \bar{\boldsymbol{Y}}_{\cdot l})'. \tag{4.5}$$

Suma u (4.5) se naziva matrica sume kvadrata i uzajamnog proizvoda (SS&CP).

Ukupna objedinjena kovarijansa uzorka, ocena za kovarijansnu matricu uzorka $\boldsymbol{\Sigma}$ je tada

$$\hat{\boldsymbol{\Sigma}} = (r_1 + r_2 - 2\,)^{-1} \Big((r_1 - 1)\hat{\boldsymbol{\Sigma}}_1 + (r_2 - 1)\hat{\boldsymbol{\Sigma}}_2\Big).$$

U višedimenzionalnom slučaju koristimo test statistiku analognu kvadratu t statistike, poznatu kao Hotellingova T^2 statistika, koja je data je sa

$$T^2 = (r_1^{-1} + r_2^{-1})^{-1} \left(\bar{\boldsymbol{Y}}_{\cdot 1} - \bar{\boldsymbol{Y}}_{\cdot 2}\right)' \hat{\boldsymbol{\Sigma}}^{-1} (\bar{\boldsymbol{Y}}_{\cdot 1} - \bar{\boldsymbol{Y}}_{\cdot 2}).$$

Može se pokazati da važi:

$$\frac{r_1 + r_2 - n - 1}{(r_1 + r_2 - 2)n} T^2 \sim \mathcal{F}_{n, r_1 + r_2 - n - 1},$$

odnosno da ima Fišerovu raspodelu sa $n \; i \; r_1 + r_2 - n - 1$ stepeni slobode.

Testiranje H_0 može se sprovesti na nivou značajnosti α upoređujući ovu realizovanu vrednost test statistike T^2 sa odgovarajućim kvantiolm Fišerove raspodele. Na nivou značajnosti α odbacujemo hipotezu

$$H_0: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$$

ako je

$$\frac{r_1 + r_2 - n - 1}{(r_1 + r_2 - 2)n} T^2 > \mathcal{F}_{n, r_1 + r_2 - n - 1; \alpha}$$

gde je $\mathcal{F}_{n, r_1 + r_2 - n - 1; \alpha}$ kvantil reda $(1 - \alpha)$ Fišerove $\mathcal{F}$ raspodele sa n i $r_1 + r_2 - n - 1$ stepeni slobode.

Imajmo na umu da ako je $n = 1$, multiplikativni faktor je jednak 1 i statistika ima *F* raspodelu sa 1 i $r_1 + r_2 - 2$ stepenom slobode, što je kvadrat $t_{r_1 + r_2 - 2}$ distribucije. Tako se multivarijantni test svodi na skalarni t test, ukoliko je dimenzija vektora podataka $n = 1$.

Primer:[6] Prodaja gaziranog i negaziranog pića beleži se u dva regiona. Na osnovu slučajnog uzorka od $r_1 = 40$ prodavnica iz prvog regiona i $r_2 = 50$ prodavnica iz drugog regiona dobijeni su uzorački pokazatelji:

$$\overline{\boldsymbol{Y}}_{\cdot 1} = \begin{bmatrix} 28 \\ 28 \end{bmatrix}, \overline{\boldsymbol{Y}}_{\cdot 2} = \begin{bmatrix} 25 \\ 30 \end{bmatrix}, \widehat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 10 & 5 \\ 5 & 12 \end{bmatrix}, \widehat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 10 & 4 \\ 4 & 8 \end{bmatrix}.$$

Pitamo se da li postoji razlika u prodaji pića u ova dva regiona.

Uzoračka sredina je

$$\overline{\boldsymbol{Y}}_{\cdot 1} - \overline{\boldsymbol{Y}}_{\cdot 2} = \begin{bmatrix} 3 \\ -2 \end{bmatrix},$$

a objedinjena kovarijansna matrica uzorka je

$$\widehat{\boldsymbol{\Sigma}} = (40 + 50 - 2)^{-1} \left((40 - 1)\widehat{\boldsymbol{\Sigma}}_1 + (50 - 1)\widehat{\boldsymbol{\Sigma}}_2 \right) = \begin{bmatrix} 10 & 4.4431 \\ 4.4431 & 9{,}7727 \end{bmatrix}.$$

Vrednost T^2 statistike je:

$$T^2 = \left(\frac{1}{40} + \frac{1}{50} \right)^{-1} \left(\overline{\boldsymbol{Y}}_{\cdot 1} - \overline{\boldsymbol{Y}}_{\cdot 2} \right)' \widehat{\boldsymbol{\Sigma}}^{-1} \left(\overline{\boldsymbol{Y}}_{\cdot 1} - \overline{\boldsymbol{Y}}_{\cdot 2} \right) =$$

$$= \frac{40 \cdot 50}{40 + 50} \begin{bmatrix} 3 & -2 \end{bmatrix} \begin{bmatrix} 0.12532 & -0.05697 \\ -0.05697 & 0.12823 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} = 51{,}6543.$$

Kritična vrednost testa na nivou značajnosti $\alpha = 0{,}05$ je

$$\mathcal{F}_{2, 40 + 50 - 2 - 1; 0{,}05} = \mathcal{F}_{2, 87; 0{,}05} = 3{,}1013;$$

Dalje računamo test statistiku, zasnovanu na statistici T^2 :

$$\frac{40 + 50 - 2 - 1}{(40 + 50 - 2)2} 51{,}6543 = \frac{87}{176} \cdot 51{,}6543 = 25{,}5337$$

Kako je $25{,}5337 > 3{,}1013$ hipotezu H_0 odbacujemo na nivou značajnosti od 5% i zaključujemo da postoji razlika u prodaji dve vrste pića s obzirom na region.

Hipoteza (4.4) se može izraziti u formi koju smo koristili ranije. Konkretno, ako definišemo $\boldsymbol{M}$ kao i ranije, kao i $(2 \times n)$ matricu čije su vrste transponovani vektori očekivanja $\boldsymbol{\mu}_1'$ i $\boldsymbol{\mu}_2'$, tj.,

$$\boldsymbol{M} = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1n} \\ \mu_{21} & \cdots & \mu_{2n} \end{bmatrix},$$

[6] Podaci su preuzeti iz knjige Multivarijaciona analiza, Z. Kovačić, Univerzitet u Beogradu, Ekonomski fakultet, Beograd.

a ukoliko definišemo $\boldsymbol{C} = [1 \quad -1]$, dobijamo

$$\boldsymbol{CM} = [\mu_{11} - \mu_{21} \quad \cdots \quad \mu_{1n} - \mu_{2n}] = [\boldsymbol{\mu}_1' - \boldsymbol{\mu}_2']'.$$

Dakle, hipotezu možemo izraziti u sledećem obliku:

$$H_0: \boldsymbol{CMU} = \boldsymbol{0}, \qquad \boldsymbol{U} = \boldsymbol{I}_n.$$

4.3. Jednofaktorska MANOVA

Ispitujemo uticaj jednog faktora na varijabilnost više karakteristika, pa se stoga odgovarajući model analize naziva jednofaktorska MANOVA. Rezultate jedinki jedne grupe (jednog tretmana) smatramo uzorkom iz višedimenzione populacije određenih karakteristika. MANOVA sa ponovljenim merenjima se zasniva na nekoliko pretpostavki. Te pretpostavke su manje ili više uporedive sa onima kod t – test parova i glase:

1. Posmatranja različitih jedinki svakog ponovljenog merenja su nezavisna.
2. Posmatranja treba da imaju multivarijantnu normalnu raspodelu.

Kao što se poređenje očekivanja za skalarne rezultate merenja dve grupe može generalizovati za $q > 2$ grupe koristeći tehniku analize varijanse tako se i multivarijantna analiza takođe može generalizovati.

Skalarni slučaj: Ako su rezultati skalari, bili bismo zainteresovani za poređenje q skalarnih očekivanja $\mu_l, l = 1, \dots, q$, pa se H_0 svodi na

$$H_0: \mu_1 = \cdots = \mu_q,$$

i nepoznata kovarijansna matrica $\boldsymbol{\Sigma}$ bi se svela na skalarnu varijansu σ^2. Testiramo hipotezu H_0 protiv alternativne hipoteze da je kod bar jedne grupe očekivana vrednost rezultata različita u odnosu na vrednosti drugih grupa. Nulta hipoteza o jednakosti očekivanih vrednosti grupa ekvivalentna je hipotezi izraženoj u obliku:

$$H_0: \mu_1 - \mu_2 = \mu_2 - \mu_3 = \cdots = \mu_{q-1} - \mu_q = 0.$$

Pod pretpostavkom normalne raspodele, standardni test za H_0 u univarijantnoj analizi varijanse zasniva se na količniku dve ocene za σ^2. U tabeli 5 je data uobičajena univarijantna analiza varijanse; podsetimo se da je $m = \sum_{l=1}^{q} r_l$ ukupan broj jedinki:

Tabela 5: Analiza varijanse

Izvori varijacije	*Zbir kvadrata*	*Stepeni slobode*	*Sredina zbira kvadrata (varijansa)*	*F*
Između grupa	$SS_G = \sum_{l=1}^{q} r_l(\bar{Y}_{\cdot l} - \bar{Y}_{\cdot\cdot})^2$	$q - 1$	MS_G	MS_G/MS_E
Unutar grupa,greške	$SS_E = \sum_{l=1}^{q} \sum_{h=1}^{r_l} (\bar{Y}_{hl} - \bar{Y}_{\cdot l})^2$	$m - q$	MS_E	
Ukupno	$\sum_{l=1}^{q} \sum_{h=1}^{r_l} (\bar{Y}_{hl} - \bar{Y}_{\cdot\cdot})^2$	$m - 1$		

Uočimo da suma kvadrata grešaka SS_E može biti zapisana kao

$$SS_E = (r_1 - 1)S_1^2 + \ \ldots + \big(r_q - 1\big)S_q^2, \quad S_l^2 = (r_l - 1)^{-1} \sum_{h=1}^{r_l} (Y_{hl} - \bar{Y}_{\cdot l})^2,$$

gde je S_l^2 varijansa uzorka za l - tu grupu, pa MS_E se tumači kao objedinjena ocena varijanse uzorka za σ^2 svih q grupa. MS_G je ocena za σ^2 na osnovu odstupanja očekivanja grupe od ukupnog očekivanja i koja će preceniti σ^2 ukoliko su očekivanja različita. Količnik F ima Fišerovu raspodelu sa $(q-1)$ i $(m-q)$ stepeni slobode, pa je test sproveden na nivou značajnosti α, upoređujući dobijenu vrednost sa odgovarajućim kvantilom $\mathcal{F}_{q-1,m-q,\alpha}$.

Multivarijantni slučaj: Hipoteze su sada

$$H_0: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q.$$

Kao u slučaju kada postoje $q = 2$ grupe, multivarijantna generalizacija podrazumeva činjenicu da ne tražimo ocenu za jednu varijansu, nego za čitavu kovarijansnu matricu $\boldsymbol{\Sigma}$ (koja je ista za sve grupe).

Neka je $\bar{Y}_{\cdot\cdot j}$ aritmetička sredina uzorka svih posmatranja na svim jedinikama i grupama za j - ti element i definišimo ukupni vektor očekivanja.

$$\bar{Y}_{\cdot\cdot} = \begin{bmatrix} \bar{Y}_{\cdot\cdot 1} \\ \vdots \\ \bar{Y}_{\cdot\cdot n} \end{bmatrix}.$$

Tabela 6: MANOVA tabela

Izvori varijacije	*SS&CP*	*Stepeni slobode*
Između grupa	$Q_H = \sum_{l=1}^{q} r_l(\bar{Y}_{\cdot l} - \bar{Y}_{\cdot\cdot})(\bar{Y}_{\cdot l} - \bar{Y}_{\cdot\cdot})'$	$q-1$
Između jedinki,greška	$Q_E = \sum_{l=1}^{q} \sum_{h=1}^{r_l} (\bar{Y}_{hl} - \bar{Y}_{\cdot l})(\bar{Y}_{hl} - \bar{Y}_{\cdot l})'$	$m-q$
Ukupno	$Q_H + Q_E = \sum_{l=1}^{q} \sum_{h=1}^{r_l} (\bar{Y}_{hl} - \bar{Y}_{\cdot\cdot})(\bar{Y}_{hl} - \bar{Y}_{\cdot\cdot})'$	$m-1$

Izrazi u tabeli su matrice i svaka se može posmatrati kao pokušaj ocene matrice $\boldsymbol{\Sigma}$.

Matrica sume kvadrata grešaka između jedinki i uzajamnog proizvoda $\boldsymbol{Q}_E$ može se zapisati na sledeći način:

$$\boldsymbol{Q}_E = (r_1 - 1)\widehat{\boldsymbol{\Sigma}}_1 + \ \ldots + \big(r_q - 1\big)\widehat{\boldsymbol{\Sigma}}_q,$$

gde je $\widehat{\boldsymbol{\Sigma}}_l$ ocena za $\boldsymbol{\Sigma}$ zasnovana vektorima podataka l - te grupe. Dakle, ova veličina podeljena sa svojim stepenima slobode ima interpretaciju kao "objedinjena" ocena $\boldsymbol{\Sigma}$ svih grupa.

Multivarijantni testovi:

Kako su ovi izrazi matrice, nije jednostavno napraviti jedinstvenu generalizaciju za F količnik koji se može koristiti za testiranje hipoteze H_0. Jasno je da želimo uporediti "veličinu" "SS&CP" matrica $\boldsymbol{Q}_H$ i $\boldsymbol{Q}_E$, ali nema jedinstvenog načina da se to uradi. Postoji nekoliko statistika koje bi se mogle koristiti. Jedna od njih je *Wilksova lambda*. U skalarnom slučaju, F količnik je

$$\frac{SS_G/(q-1)}{SS_E/(m-q)};$$

Dakle, u skalarnom slučaju, H_0 se odbacuje kada je količnik SS_G/SS_E velik. Ovo je ekvivalentno odbacivanju hipoteze H_0 za velike vrednosti izraza

$$1+\frac{SS_G}{SS_E},$$

odnosno za male vrednosti izraza

$$\frac{1}{1+SS_G/SS_E}=\frac{SS_E}{SS_G+SS_E}. \tag{4.6}$$

Kod MANOVA sa jednim faktorom navedene sume kvadrata zamenjujemo odgovarajućim vrednostima generalizovanih varijansi odnosno matricama sume kvadrata i uzajamnih proizvoda unutar i između grupa $\boldsymbol{Q}_H$ i $\boldsymbol{Q}_E$. Za multivarijantni problem, statistika *Wilksova lambda* u oznaci Λ, analogna je veličini (4.6), i glasi:

$$\Lambda=\frac{|\boldsymbol{Q}_E|}{|\boldsymbol{Q}_H+\boldsymbol{Q}_E|};$$

Ovde je za svaku matricu SS&CP uzeta njena determinanta. Wilksova lambda Λ se često naziva uproštena uzoračka varijansa. Hipoteza H_0 se odbacuje za male vrednosti Λ.

U izvesnim specifičnim slučajevima egzaktna raspodela Wilksove lambde je poznata i navedena je u narednoj tabeli. U ostalim slučajevima se koriste aprkosimacije[7].

Tabela 7: Transformacija Wilksove lambde u statistiku

Broj promenljivih (dimenzija)	Broj grupa	Uzorački raspored za višedimenzione normalne podatke
$\boldsymbol{n=1}$	$q\geq 2$	$\frac{1-\Lambda}{\Lambda}\frac{m-q}{q-1}\sim\mathcal{F}_{q-1,m-q}$
$\boldsymbol{n=2}$	$q\geq 2$	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{m-q-1}{q-1}\sim\mathcal{F}_{2(q-1),m-q-1}$
$\boldsymbol{n\geq 1}$	$q\geq 2$	$\frac{1-\Lambda}{\Lambda}\frac{m-n-1}{q-1}\sim\mathcal{F}_{n,m-n-1}$
$\boldsymbol{n\geq 1}$	$q=3$	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{m-n-2}{q-1}\sim\mathcal{F}_{2n,2(m-n-2)}$

Pored Wilksove lambde koriste se i druge statistike. Spomenimo neke: Raova statistika, Hotelling trag, Pillaijev trag i Royov najveći koren. Prva od njih, Raova statistika, predstavlja transformaciju Wilksove lambde i neki njeni specijalni slučajevi prikazani su u Tabeli 7. Neke od statistika predstavljaju transformaciju Wilksove lambde, dok su druge bazirane na matricama $\boldsymbol{Q}_H$ i $\boldsymbol{Q}_E$, odnosno na njihovim karakterističnim korenima ili tragu.

[7] Najpoznatija je Bartlettova aproksimacija.

Hotellingov trag se zasniva na tragu matrice $\boldsymbol{Q}_H\boldsymbol{Q}_E{}^{-1}$ (odbacuje H_0 za velike vrednosti $tr(\boldsymbol{Q}_H\boldsymbol{Q}_E^{-1})$), dok se Royov najveći koren zasniva na najvećem karakterističnom korenu matrice $\boldsymbol{Q}_H\boldsymbol{Q}_E{}^{-1}$.

Ispitivanja moći različitih testova ukazuje da njihova moć zavisi od tipa alternativne hipoteze i da nema testa koji je superioran u odnosu na ostale. Vrednost navedenih statistika, kao i aproksimaciju vrednosti F – statistike na osnovu njihove transformacije možemo dobiti među izlaznim rezultatima računarskih statističkih programa, a time možemo izabrati test u zavisnosti od tipa alternativne hipoteze.

Značajno je istaći da se hipoteza H_0: $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q$ može izraziti kao $H_0 : \boldsymbol{CMU} = \boldsymbol{0}$ za odgovarajući izbor $\boldsymbol{C}$ i za $\boldsymbol{U} = \boldsymbol{I}_n$. Na primer, neka je $q = 3$ i

$$\boldsymbol{M} = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1n} \\ \mu_{21} & \cdots & \mu_{2n} \\ \mu_{31} & \cdots & \mu_{3n} \end{bmatrix},\ \boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \tag{4.7}$$

$$\boldsymbol{CM} = \begin{bmatrix} \mu_{11} - \mu_{21} & \cdots & \mu_{1n} - \mu_{2n} \\ \mu_{11} - \mu_{31} & \cdots & \mu_{1n} - \mu_{3n} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \\ (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)' \end{bmatrix}.$$

Izjednačavanje $\boldsymbol{CM}$ sa $\boldsymbol{0}$ je ekvivalentno tome da su svi vektori očekivanja $\boldsymbol{\mu}_l$ jednaki.

Videli smo da u situacijama u kojima se vektori podataka sastoje od n zapažanja o eventualno različitim karakteristikama na različitim skalama, moguće je testirati da li su svi vektori očekivanja za svaku grupu isti, koristeći takozvane jednofaktorske MANOVA metode. Ako je nulta hipoteza (4.3) odbačena, to ukazuje da se najmanje jedan od q vektora očekivanja razlikuje od drugih u najmanje jednoj od n komponenti. Ovo nije naročita informacija, posebno ako su q i/ili n veliki. Pored toga, intuitivno deluje da će se teško otkriti takva razlika - sa q vektora i n komponenti.

Štaviše, metodi zahtevaju procenu svih $n(n+1)/2$ elemenata kovarijansne matrice **Σ,** za koju pretpostavljamo da je zajednička za sve grupe.

Predstavimo još interesantnu interpretaciju Wilksove lambde. Koeficijent determinacije (R^2) meri jačinu veze između zavisne i nezavisnih promenljivih, odnosno, pokazuje koliki je deo varijacije zavisne promenljive objašnjen modelom višestruke regresije. Koeficijent determinacije uzima vrednosti $0 \leq R^2 \leq 1$. Definišemo ga kao:

$$R^2 = 1 - \frac{\text{suma kvadrata reziduala}}{\text{ukupna suma kvadrata}}.$$

U višedimenzionom slučaju meru analognu R^2 definišemo sa $1 - \Lambda$, gde je $\Lambda = \frac{|\boldsymbol{Q}_E|}{|\boldsymbol{Q}_H + \boldsymbol{Q}_E|}$. Wilksova lambda meri udeo varijanse unutar tretmana u ukupnoj varijansi. Što je veće $1 - \Lambda$, veći je deo ukupne generalizovane varijanse koja se može pripisati varijacijama između grupa, odnosno tretmana. Na primer, ukoliko je veličina $1 - \Lambda$ iznosi 0,83, tada je 83% ukupne varijacije između posmatranja rezultat varijacija između sredina.

4.4. Analiza profila

U svakoj od q grupa, svaki subjekat je podvrgnut dejstvu n tretmana. Svi rezultati različitih tretmana odnose se na istu promenljivu, pa su iskazani u istim jedinicama mere. Ovde nas može interesovati kako poređenje sredina n tretmana za ovih q grupa, tako i poređenje sredina q grupa za ovih n tretmana. Dakle, analizu profila možemo vršiti i po grupama i po tretmanima. Sa stanovišta analize profila svejedno je koji eksperimentalni plan ćemo razmatrati.

Mogu se sprovesti multivarijantni testovi koji nemaju nikakvu posebnu pretpostavku o formi $\boldsymbol{\Sigma}$ (naglasimo da važi da je kovarijansna matrica ista za sve grupe). Slučajni uzorci uzeti iz svake grupe su međusobno nezavisni. Podsetimo se da se MANOVA test hipoteze u (4.3), $H_0: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q$ može posmatrati kao test posebnih hipoteza oblika

$$H_0: \boldsymbol{CMU} = \boldsymbol{0}$$

za odgovarajući izbor $\boldsymbol{C}$ i sa $\boldsymbol{U} = \boldsymbol{I_n}$. Takođe je moguće da se razviju multivarijantne procedure za opštiji izbor matrica $\boldsymbol{C}$ i $\boldsymbol{U}$.

Zanimaju nas odgovori na sledeća pitanja preko kojih formulišemo pitanje o jednakosti sredina. Ta pitanja glase:

1. Da li su profili paralelni?
2. Pretpostavivši da su profili zaista paralelni, da li su istovremeno i podudarni?
3. Pretpostavivši da se profili zaista podudaraju, da li su profili na istom nivou?

Od posebnog interesa u analizi longitudinalnih podataka je test paralelizma ili interakcije grupa i vremena. Videli smo da nulta hipoteza koja odgovara paralelizmu može da se izrazi preko elemenata vektora očekivanja $\boldsymbol{\mu}_l$ ili ekvivalentno preko izraza $(\tau\gamma)_{lj}$:

$$H_0: \text{svi } (\tau\gamma)_{lj} = 0.$$

Posebno, ukoliko je $q = 2$ i $n = 3$, ovaj test možemo predstaviti sa

$$\boldsymbol{C} = [1 \quad -1], \qquad \boldsymbol{U} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}, \qquad \boldsymbol{M} = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \end{bmatrix}.$$

Za opšte q i n, možemo dati drugačiji zapis. Neka j_p označava vektor kolonu jedinica dužine p, pa biranjem

$$\boldsymbol{C} = [j_{q-1} \quad -\boldsymbol{I}_{q-1}]_{(q-1\times q)}, \ \boldsymbol{U} = \begin{bmatrix} j'_{n-1} \\ -\boldsymbol{I}_{n-1} \end{bmatrix}_{(n\times n-1)} \tag{4.8}$$

dobijamo nultu hipotezu paralelizma.

Multivarijantni test za paralelizam

Univarijantni test za ovu nultu hipotezu u Glavi 3 zahteva pretpostavku složene simetrije. Ovde tražimo test koji nema pretpostavke o obliku matrice $\boldsymbol{\Sigma}$. Da bismo ovo razumeli, razmotrimo prvo multivarijantni test (4.3). Podsetimo se da se u MANOVA tabeli ovaj test svodi na poređenje dve SS&CP matrice, $\boldsymbol{Q}_H$ i $\boldsymbol{Q}_E$. Matrica $\boldsymbol{Q}_E$ u stvari meri udaljenost pojedinačnih vektora podataka od očekivanja njihovih grupa. Matrica $\boldsymbol{Q}_H$ meri udaljenost vektora očekivanja grupa od ukupnog vektora očekivanja. Za očekivati je da je $\boldsymbol{Q}_H$ „mnogo veće" u odnosu na $\boldsymbol{Q}_E$ ako stvarno postoji razlika između q očekivanja $\boldsymbol{\mu}_l, l = 1, \dots, q$.

Osnovni test koji nas interesuje je test paralelizma ili test interakcije grupa i vremena. Ovo se može predstaviti kao H_0: $\boldsymbol{CMU} = \mathbf{0}$, sa $\boldsymbol{C}$ i $\boldsymbol{U}$, kao u (4.8), pa odgovarajući $\boldsymbol{Q}_H$ i $\boldsymbol{Q}_E$ mogu da se izračunaju. Dakle, test statistike kao Wilksova lambda, Pillaijev trag i tako dalje, mogu da se koriste u sprovođenju testa. U zavisnosti od dimenzija q i n, ovi testovi mogu biti tačni ili približni i mogu ali ne moraju da se poklapaju.

Naredni test se sprovodi samo ako hipoteze paralelizma nisu odbačene. Test, H_0: $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q$ možemo predstaviti i u formi H_0: $\boldsymbol{CMU} = \mathbf{0}$ sa $\boldsymbol{C}$ kao u (4.8) $\boldsymbol{U} = \boldsymbol{I_n}$. Ovo je uobičajeni MANOVA test razmatran ranije; kada su tu ponovljena merenja, test se često zove *test slučajnosti*. Očito, ako profili nisu paralelni, ne savetuje se testiranje slučajnosti, jer nije jasno šta bi to uopšte značilo.

Ukoliko su profili paralelni, možemo unaprediti ovaj test. Može se pokazati da je testiranje H_0 sa dodatnom pretpostavkom da su profili paralelni ekvivalentno testiranju hipoteza H_0: $\boldsymbol{CMU} = \mathbf{0}$ sa $\boldsymbol{C}$ kao u (4.8), ali sa $\boldsymbol{U} = j_n/n$. Testiranje da li se profili stvarno podudaraju je isto kao testiranje da li su proseci očekivanja tokom vremena isti za svaku grupu; to zovemo glavnim efektom grupa.

Ispostavlja se da za testiranje ovih hipoteza su svi multivarijantni testovi ekvivalentni. Dalje, oni smanjuju (uprošćavaju) univarijantni F test za glavne efekte grupa (razmatrane u Glavi 3). Sve dok je matrica $\boldsymbol{\Sigma}$ ista za sve vektore podataka, svi podaci imaju istu varijansu, pa možemo primeniti običan F odnos. Ovi testovi se sprovede samo ako hipoteze paralelizma nisu odbačene.

Takođe nas zanima da li su profili zapravo konstantni tokom vremena. Može se pokazati da se ovo može predstaviti u formi H_0: $\boldsymbol{CMU} = \mathbf{0}$ sa izborima $\boldsymbol{U}$ kao u (4.8) i $\boldsymbol{C} = j_q'/q$, $(1 \times q)$ vektora $1/q$ - ova. Primetimo da su ovo potpuno iste hipoteze kao one za glavni efekat vremena, razmatran u Glavi 5 – ako znamo da su profili paralelni, tada postavljanje pitanja da li su očekivanja konstantna tokom vremena je isto kao postavljanje pitanja da li su prosečni očekivani rezultati za grupe isti tokom vremena.

Važno je prepoznati da iako i univarijantne i multivarijantne test statistike imaju F uzoračku raspodelu, to su različite statistike koje su zasnovane na različitim pretpostavkama o obliku matrice $\boldsymbol{\Sigma}$. Koja je prikladnija zavisi od stvarnog oblika $\boldsymbol{\Sigma}$.

4.5. Oblik odnosa između zavisne promenljive i vremena

Kada rešimo pitanje od interesa da li postoji ili ne promena vrednosti zavisne promenljive Y tokom vremena i zaključimo da takva promena postoji, dalje hoćemo da ispitamo oblik veze između zavisne promenljive Y i vremena.

Jasno je da je to pitanje od značaja kada postoji više od dva merenja. Kada postoje samo dva merenja, jedini moguć odnos je linearan. Pitanje o obliku veze može se rešiti primenom MANOVA za ponovljena merenja, tako što se poredi odnos zavisne promenljive Y i vremena sa hipotetičkim linearnom vezom, hipotetičkim kvadratnom vezom itd. Kada postoji n ponovljenih merenja, može se testirati $n - 1$ moguća funkcija vremena. Iako se može testirati svaki mogući odnos, važno je imati odgovarajuću ideju ili hipotezu o obliku odnosa zavisne promenljive i vremena. Za svaku moguću vezu računa se F – statistika koja ima F raspodelu sa $1, (m - 1)$ stepeni slobode.

Razmotrimo hipotetički primer naveden u Glavi 3.3. Da bismo odgovorili na pitanje: "Koji je oblik odnosa između zavisne promenljive Y i vremena?", moramo transformisati zavisnu promenljvu. Kako postoje četiri ponovljena merenja, Y se može transformsati u linearnu, kvadratnu ili kubnu komponentu.

Ove transformacije se vrše prema "faktorima" transformacije koji su dobijeni u SPSS – u i prikazani su u sledećoj tabeli:

vreme[a]

Zavisna promenljiva	vreme		
	Linearna	Kvadratna	Kubna
vreme1	-.671	.500	-.224
vreme2	-.224	-.500	.671
vreme3	.224	-.500	-.671
vreme4	.671	.500	.224

Ovako transformisane promenljive se dalje koriste da se testiraju moguće veze sa vremenom.

Svaka vrednost iz originalnog skupa podataka se množi odgovarajućim transformacionim množiocem da se dobije transformisani skup podataka.

Sledeća tabela predstavlja linearnu transformaciju skupa podataka.

Tabela 8: Transformisani podaci sa ponovljenim merenjima.

i	$Y_{t1}{}^*$	$Y_{t2}{}^*$	$Y_{t3}{}^*$	$Y_{t4}{}^*$	Prosek
1	-20,8	-6,5	3,4	17,5	-1,62
2	-16,1	-6,3	4,5	21,5	0,89
3	-9,4	-4,5	6,3	20,1	3,13
4	-25,5	-7,6	6,7	22,8	-0,9
5	-16,8	-6,5	5,6	19,5	0,45
6	-20,1	-6,3	3,6	22,8	0
Prosek					0,33

Testiraćemo da li postoji linearna veza promenljive vremena.

Prvi korak je da izračunamo individualnu sumu kvadrata transformisanih promenljivih. Dobijamo:

$$SS_T{}^* = 4[(-1{,}62 - 0{,}33)^2 + (0{,}89 - 0{,}33)^2 + \ \ldots + (0 - 0{,}33)^2] = 54{,}43.$$

Sledeći korak je da izračunamo individualnu sumu kvadrata kada pretpostavimo da je ukupna očekivana vrednost nula. Primenjujući to na transformisani skup podataka dobijamo

$$SS_T{}^0 = 4[(-1{,}62 - 0)^2 + (0{,}89 - 0)^2 + \ \ldots + (0 - 0)^2] = 56{,}96.$$

Razlike između ove dve individualne sume kvadrata je odsečak. U ovom primeru odsečak je 2,546 i ta vrednost se koristi za testiranje linearne zavisnosti tokom vremena. Što je razlika bliža nuli, manje je verovatno da postoji linearni odnos s vremenom. U ovom primeru dobija se da je p vrednost odsečka 0,65, što znači da ne postoji značajan linearni odnos između zavisne promenljive i vremena.

Analogna procedura se sprovodi kada se testira moguća kvadratna ili kubna veza sa vremenom.

4.6. Razlike univarijantnog i multivarijantnog pristupa

Poslednje dve decenije metode univarijantne i multivarijantne analize podataka primenjuju se u skoro svim naučnim oblastima i to iz dva osnovna razloga. Prvi je taj što je razvoj kompjuterske tehnike i softverskih proizvoda omogućio relativno jednostavnu primenu ovih metoda, a drugi je sagledavanje potreba mnogih naučnih istraživanja da se analiziraju odnosi između promenljivih. Multivarijantni pristup se razvio kasnije u odnosu na univarijantni pristup.

U okviru analize ponovljenih merenja možemo napraviti razliku između multivarijantnog pristupa (višefaktorsko proširenje t – test parova) i univarijantnog pristupa (proširenje ANOVA). Problem je što ova dva pristupa ne daju iste rezultate pa se postavlja pitanje koji pristup treba koristiti. Jedna od razlika između ova dva pristupa je pretpostavka sferičnosti. Za multivarijantni pristup ova pretpostavka nije neophodna, dok je kod univarijantnog pristupa vrlo važna. Ograničenje pretpostavke složene simetrije (jednake korelacije i jednake varijacije tokom vremena) dovodi do povećanja stepena slobode, odnosno povećanje moći univarijantnog pristupa, koje postaje važno kada uzorak postaje manji.

Kada pretpostavka sferičnosti nije zadovoljena moguća su dva pristupa. Prvi je da koristimo prilagođene testove univarijantnog pristupa (korekcije ANOVA testova), a drugi je da koristimo multivarijantni pristup. Ponekad se tvrdi da kada je broj posmatranih subjekata veći od broja ponovljenih merenja plus deset i kada je epsilon manje od 0,7 treba koristiti MANOVA jer je ima veću moć. U svakoj drugoj situaciji, preporučljivo je da se rezultati i univarijantnog i multivarijantnog pristupa koriste za rešavanje pitanja od interesa i tek kada oba pristupa daju iste rezultate prilično je izvesno da postoji ili ne postoji značajna promena tokom vremena. Kada ova dva pristupa daju različite rezultate, zaključci se moraju izvući sa mnogim ograničenjima i velikim oprezom.

Jedan od problema MANOVA sa ponovljenim merenjima je da su vremenski periodi koji se posmatraju jednako udaljeni. Neznačajna promena tokom kratkog vremenskog perioda može biti relativno veća od značajne promene tokom dužeg vremenskog perioda. Dakle, kada su vremenski periodi između merenja nejednako udaljeni, rezultati MANOVA sa ponovljenim merenjima ne mogu biti dobro interpretirani. Mora se uzeti u obzir i dužina vremenskih intervala. Drugi veliki problem MANOVA sa ponovljenim merenjima je da uzima u obzir samo jedinke sa kompletnim podacima tj. jedinke koje su izmerene u svim vremenskim tačkama (bez izostavljenih merenja). Ukoliko jedinka nije dostupna za neko merenje njeni ostali podaci se brišu iz analize.

4.7. Nedostaci i ograničenja univarijantnog i multivarijantnog pristupa

Univarijantne i multivarijantne klasične metode koje smo razmatrali mogu biti proširene na komplikovanije situacije, jer ovako imaju ograničenja, od kojih smo neka primetili do sada. Kad smo upoznati sa ovim tzv. klasičnim metodama i statističkim modelima u osnovi njih, u poziciji smo da budemo precizniji sa tim ograničenjima. Navedimo pretpostavke klasičnih metoda i ograničenja koja se nameću.

1. Istaknuta karakteristika univarijantnih i multivarijantnih klasičnih modela i metoda je potreba da sve jedinke budu posmatrane u *istim* "vremenskim" tačkama (n merenja). Dakle, ne samo da svaki vektor podataka $\boldsymbol{Y}_i$ mora biti iste veličine n, za sve jedinke, nego se svaki element $Y_{ij}, j = 1, \dots, n$ mora posmatrati na *istom* skupu vremenskih tačaka $t_1, \dots, t_n$.

U nekim slučajevima, ovo nisu velika ograničenja. Na primer, u poljoprivrednim i industrijskim eksperimentima, gde je moguće imati dobru kontrolu nad eksperimentalnim uslovima, eksperiment može biti pažljivo planiran i izvršen. Skroz je razumno očekivati da će posmatranja u određenim trenucima biti dostupna.

Ipak i u stvarnosti postoji slučaj da stvari krenu naopako. Uzorci mogu biti zagubljeni, pogrešno odbačeni, da dođe do grešaka pri merenju ili uzorci mogu biti nedostupni u određenim vremenskim tačkama, čime se kvari ravnoteža neophodna u klasičnim modelima i metodama gde se primenjuju što dovodi do grešaka pri sprovođenju testa.

Kada su jedinke ljudi to postaje veći problem čak i ako je studija pažljivo dizajnirana. Na primer, pretpostavimo da je studija sprovedena da se uporedi nekoliko lekova za snižavanje holesterola u krvi. Subjektima se nasumično dodeljuju redovne doze jednog od lekova i potrebno je da dolaze na tromesečnim intervalima tokom dve godine tako da se može uzeti mera seruma holesterola iz uzoraka krvi pri svakoj poseti. Subjekat treba da ima merenja seruma holesterola $n = 8$ puta (meseci $3, 6, 9, 12, \dots, 21$ i 24 u odnosu na početak studije). Međutim, u praksi se mogu desiti nepredviđene situacije:

- Subjekti mogu da se odsele tokom studije, tako da su dostupna samo merenja do njihove poslednje posete.
- Subjekat može biti van grada i propustiti svoju posetu na primer u 6. mesecu, ali umesto toga, dolazi na kliniku u 7. mesecu.
- Uzorci krvi mogu biti neoznačeni ili da su ispali u laboratoriji, tako da je rezultati o serumu holestrola za neka vremena pojedinih subjekata nemoguće dobiti.
- Greške tehničara pri obavljanju analitičkih laboratorijskih tehnika potrebnih za merenje nivoa holestrola mogu učiniti i druga merenja pogrešnim ili nedostupnim.

Suština je da stvarni život često čini da uslov ravnoteže postaje nedostižan za mnoge longitudinalne studije, jer istraživači često ne mogu da kontrolišu sve okolnosti dobijenih rezultata, tako da podaci postaju neuravnoteženi ili nepotpuni. Neki istraživači su raspravljali o načinima da se klasični pristupi "prilagode" rukovanju situacijama neravnoteže (kao prilagođeni F testovi kod univarijantne analize, ali ova prilagođavanja skrivaju pravi problem, a to je da je model koji zahteva ravnotežu previše restriktivan da predstavlja stvarni život).

2. Klasične univarijantne i multivarijantne procedure koje smo razmatrali pretpostavljaju da je kovarijansna matrica svakog vektora podataka $\boldsymbol{Y}_i, j = 1, \dots, m$ ista za sve i, bez obzira na pripadnost grupi. Pod uslovom da verujemo da je ova pretpostavka opravdana i da je $\boldsymbol{\Sigma}$ zajednička $(n \times n)$ kovarijansna matrica i dalje smo suočeni sa pitanjem šta pretpostavljamo o strukturi $\boldsymbol{\Sigma}$.

Univarijantne metode pretpostavljaju složenu simetriju, što podrazumeva veoma specifičan vid korelacije između posmatranja uzetih na istoj jedinki u različitim vremenima, koja može biti prilično nerealna za longitudinalne podatake. Ovaj model kaže da je korelacija svih zapažanja date jedinke ista bez obzira na to koliko su zapažanja blizu ili daleko u odnosu na vreme. Dakle, univarijantne metode su zasnovane na pretpostavci o strukturi kovarijanse koja može biti previše restriktivna ako izvori korelacije unutar jediniki nisu zanemarljivi.

Multivarijantne metode ne traže da je zadovoljena pretpostavka složene simetrije**.** Dakle, ove metode ne pokušavaju uzeti u obzir sve načine na koje nastaju rezultati u longitudinalnoj analizi. Razlikujemo sledeće izvore varijacije:

- Slučajna (biološka) varijacija među jedinkama.
- Varijacije unutar jedinke, zbog načina na koji su merenja uzeta sa jednike (greška u merenju brzine, korelacija zbog vremenskog razdvajanja, itd).

Multivarijantni modeli ne priznaju eksplicitno ova dva izvora. Umesto toga dozvoljavaju mogućnost da kovarijansna struktura bude bilo šta što uključuje i moguće strukture koje verovatno neće predstavljati podatke jedinke koji podležu ovim izvorima iznad. Dakle, multivarijantne metode su zasnovane na pretpostavci o kovarijansnoj strukturi koja je verovatno previše neodređena.

3. Univarijantni i multivarijantni pristupi pretpostavljaju da je kovarijansna matrica vektora podataka ista za sve jedinke bez obzira na grupu. (Ovo liči na uobičajenu pretpostavku linearne regresije ili skalarne analize varijanse gde je varijansa ista za sva skalarna posmatranja). Ovo se često usvaja bez mnogo razmišljanja, ali sasvim je realno očekivati da ovo može biti loša pretpostavka.

Na primer, pretpostavimo da su jedinke ljudi i da su grupe određene primenom ili leka hipertenzije ili placebo. Zajedničko zapažanje ovakvih podataka jeste da osobe sa "visokim" sistolnim krvnim pritiskom često teže da mnogo više variraju u njihovom okviru pojedinačnih merenja pritiska nego osobe sa "niskim" sistolnim krvnim pritiskom. Fluktuacije unutar jedinki za subjekte sa visokim krvnim pritiskom imaju tendenciju da budu veće od onih kod subjekata sa niskim krvnim pritiskom. Dakle, $var(e_{1ij})$ je manja za subjekate sa niskim krvnim pritiskom nego za one sa visokim krvnim pritiskom.

Pretpostavimo da je lek prilično efikasan u snižavanju sistolnog krvnog pritiska. Možemo očekivati da posmatranja na subjektima grupe koja koristi lek (naročito prema kraju studije) budu "niži" od onih u placebo grupi. Simbolima, ako je $\boldsymbol{Y}_i$ vektor podataka za subjekte u grupi koja uzima lekove (1), mogli bismo očekivati

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{bmatrix}, \quad var(Y_{in}) = \sigma^2_{n(0)},$$

$$\sigma^2_{n(1)} < \sigma^2_{n(0)}.$$

Pod ovim uslovima, pretpostavka da $\boldsymbol{Y}_i$ iz obe grupe imaju istu matricu kovarijanse $\boldsymbol{\Sigma}$, je neprikladna, jer sumnjamo da je (n, n) element isti za vektore podataka u obe grupe. Bolji model je kad postoje dve različite kovarijansne matrice, tj. $var(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_0$ za subjekt i u placebo grupi i $var(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_1$ za subject i u grupi sa lekovima.

Moguće je izmeniti klasične modele i metode za prilagođavanje ovoj situaciji. Jedan od uobičajenih pristupa jeste da se radi na transformisanoj skali na kojoj se veruje da odstupanja mogu biti slična; npr. modelovanje logaritamski transformisanih podatka. Problem sa ovim pristupom je da rezultati mogu biti teški za tumačenje. Alternativno, metodi (kao što je Hotellingova T^2 statistika) mogu da se modifikuju da dozvole različite kovarijansne matrice za svaku grupu. Međutim, ovo može da napravi statističku snagu još nižom jer, moramo dati ocenu matrice kovarijanse posebno za svaku grupu.

4. Sledeća karakteristika koju dele i univarijantne i multivarijantne klasične metode o kojima smo govorili je da zbog toga što pretpostavljaju ravnotežu, vreme samo po sebi se ne pojavljuje eksplicitno u modelima vektora očekivanih podataka. Umesto toga, "vreme" ulazi u model samo kroz specifikaciju parametara γ_j i $(\tau\gamma)_{lj}$. Ovaj problem se može delimično rešiti korišćenjem, na primer, ortogonalnih polinoma kontrasta u vremenu, ali je direktna zastupljenost vremena u modelu mnogo korisnija.

Pored toga, možda ćemo želeti da uključimo i druge kovarijansne informacije. Na primer, u studiji holesterola, možemo smatrati da starost subjekta na početku studije igra ulogu na to kakvi će biti rezultati korišćenja lekova za snižavanja holesterola. Ili možemo smatrati da na rezultat tokom vremena može uticati sistolni krvni pritisak jednog subjekta, koji takođe može da se menja tokom vremena. Baš kao što je obična analiza varijanse modifikovana da uključi nezavisne promenljive (promenljive koje su u relaciji sa zavisnom promenljivom i utiču na varijaciju unutar grupa) analize kovarijanse, možemo želeti da uradimo nešto slično u slučaju ponovljenih merenja. Stvari su komplikovanije.

U prvom primeru, nezavisna promenljiva, uzrast na početku studije, je vremenski nezavisna ili fiksirana tokom vremenskih tačaka u kojima se jedinka posmatra i koja je merena samo jednom (na početku studije). I univarijantne i multivarijantne analize se mogu modifikovati da uzmu u obzir vremenski nezavisne promenljive.

Modeli koje smo razmatrali predstavljaju očekivani rezultat u svakoj vremenskoj tački kao funkciju informacija kao što je pripadnost grupi, odnosno, eventualno različita očekivanja svake grupe. Ako uzmemo u obzir modele koji sadrže promenljive informacije, javljaju se važna pitanja. Na primer, da li očekivani holesterol u određenom trenutku zavisi samo od sistolnog krvnog pritiska u tom vremenu? Ili da li to zavisi od sistolnog krvnog pritiska u prethodnih nekoliko puta?

Iako je moguće uvesti vremenski zavisne nezavisne promenljive u modelovanju ponovljenih merenja, ključno pitanje je ova koncepcija. Moguće je izmeniti univarijantnu analizu da inkorporira vremenski zavisne nezavisne promenljive; međutim, modifikacija MANOVA analize nije moguće.

5. Analize zasnovane na klasičnim metodama fokusirane su na testiranje hipoteza, odnosno opštih pitanja od interesa, izražavaju se terminima modela i utvrđuje se da li podaci daju dovoljno dokaza za odbacivanje nulte hipoteze. Tada se radi izricanje (nulta hipoteza se odbacuje ili ne odbacuje). Međutim, u mnogim situacijama, ciljevi istraživača su složeniji. Na primer, razmotrimo ranije opisanu studiju holesterola. Istražitelji možda žele da učine više od samog tvrđenja da je način na koji se menja prosek nivoa holesterola tokom vremena, pri različitim lekovima, drugačiji. Oni zapravo žele da koriste rezultate svog istraživanja da daju preporuke o tome kako tretirati buduće pacijente. Dakle, oni su možda želeli da dođu do specijalizovanijih zaključaka.

- Koliko se razlikuje stopa smanjenja holesterola među lekovima? Na primer ako su znali da lek 1 spušta holesterol sa stopom od 5 mm Hg mesečno, a lek 2 spušta holesterol sa stopom 15 mm Hg mesečno, ova informacija može da im pomogne da odluče koji lek (blagi lek 1 ili agresivni lek 2) bi više odgovarao određenom pacijentu.

Tako su možda istražitelji zainteresovani da zapravo procene brzinu promene očekivanih rezultata tokom vremena za svaku grupu.

- Kako bi putanja holesterola izgledala za novog muškog pacijenta starosti 45 godina posle 8 meseci sa jednim od lekova? Dakle, pre tretmana, istražitelji će možda želeti da predvide kako profil holesterola izgleda tokom 8 meseci za pacijenta sa specifičnim karakteristikama i koliki njegov nivo holesterola možda bude na kraju tog perioda na osnovu njegovog merenja u početnom trenutku. Imajte na umu da 8. mesec nije ni jedan od vremenskih tačaka (svaka tri meseca se vrše merenja) uključenih u originalnoj studiji.

Jasno, u cilju rešavanja takvih pitanja potreban je fleksibilniji model koji obuhvata vreme i brzinu promene na eksplicitniji način.

Klasični modeli i metode imaju ograničenja u vezi važnih pitanja od interesa. Ozbiljan nedostatak jeste potreba za ravnotežom, a zatim sledi neuspeh modela da eksplicitno predstave važne funkcije kao što su stope promena kroz vreme i korišćenje zajedničke kovarijansne matrice.

5. Opšti linearni modeli za longitudinalne podatke

Videli smo univarijantne i multivarijantne analize varijanse ponovljenih merenja, a sada ćemo analizirati nekoliko opštih linearnih modela longitudinalnih podataka, kada su podaci balansirani i kada podaci nisu balansirani.

U narednim primerima koristićemo model:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{5.1}$$

gde je

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{bmatrix}_{n\times 1}, \quad \boldsymbol{X}_i = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}_{n\times p}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1}, \quad \boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

uz ograničenja: $E(\boldsymbol{\epsilon}_i) = 0,\ var(\boldsymbol{\epsilon}_i) = \sigma^2,\ var(\boldsymbol{\epsilon}_i) = [\sigma_{ij}] = \boldsymbol{\Sigma}$.

5.1. Modeli za balansirane podatke

Razmotrićemo jednostavan slučaj longitudinalnih podataka gde ćemo videti da li model može da sadrži informacije o vremenskim posmatranjima za svaku jedinku. U ovom slučaju, mi ćemo nastaviti sa pretpostavkom da su podaci *balansirani* (svi subjekti imaju ponovljena merenja u istih n trenutaka u vremenu). Razmotrićemo sledeću situaciju:

Pretpostavimo da su svi $\boldsymbol{Y}_i,\ i = 1, \dots, m$ formata $(n \times 1)$, gde je j-ti element Y_{ij} posmatran u vremenu t_j. Vremena $t_1, \dots, t_n$ su ista za sve jedinke.

Primer 1: *Stomatološka studija*[8]

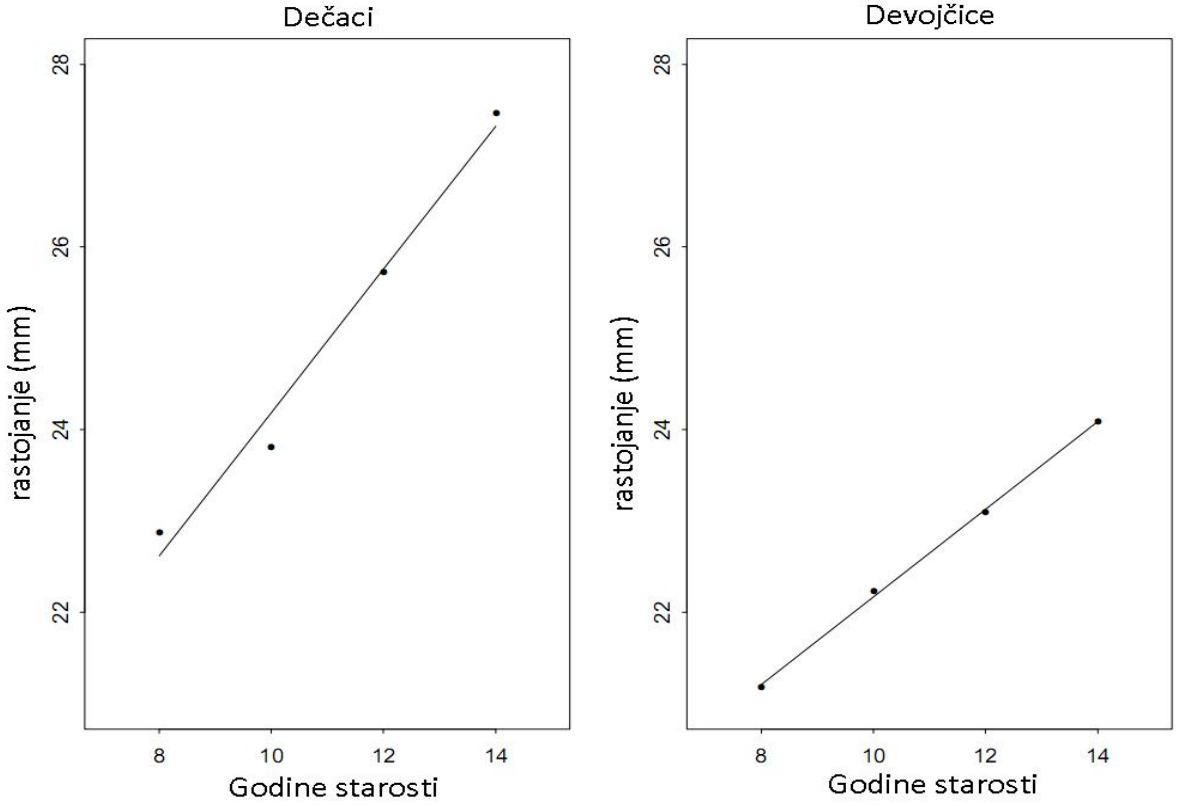
U primeru stomatološke studije istraživanje uključuje 27 dece: 16 dečaka i 11 devojčica. Na svakom detetu se meri rastojanje (mm) od hipofize do pukotine u lobanji iznad vilice u uzrastima od 8, 10, 12 i 14 godina.

Prvo predstavimo model za jednu grupu. Posmatrajmo samo jednu grupu gde se sve jedinke ponašaju na sličan način. Vektor očekivanja je jednostavan (nije neophodan opis preko grupa):

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}.$$

[8] Podaci su preuzeti iz rada Potthoff i Roy, *Generalizovani model višefaktorske analize varijanse posebno korisno za probleme krive rasta* (1964.)

Slika 7. Očekivane vrednosti uzorka u posmatranim starosnim godinama i procenjene pravolinijske putanje za podatke svakog pola.



Očekivanja za uzorak sugerišu na to da očekivane vrednosti μ_j u svakom trenutku prate pravu liniju. Umesto da gledamo na vektor očekivanja kao na skup n nepovezanih očekivanja μ_j, možemo da gledamo očekivane vrednosti tako da zadovoljavaju jednakost

$$\mu_j = \beta_0 + \beta_1 t_j.$$

Dakle, očekivane vrednosti padaju na liniju sa odsečkom β_0 i nagibom β_1. Možemo pisati

$$Y_{ij} = \beta_0 + \beta_1 t_j + \epsilon_{ij}$$

$$E(\boldsymbol{\epsilon}_i) = 0, \;\; var(\boldsymbol{\epsilon}_i) = \sigma^2, \;\; cov(\epsilon_i, \epsilon_j) = 0 \qquad (5.2)$$

Model (5.2) govori da u j - tom momentu t_j, vrednosti Y_{ij} imaju očekivanu vrednost $\beta_0 + \beta_1 t_j$ i variraju na osnovu ukupnih grešaka ϵ_{ij}. Tako da ovaj model predstavlja očekivanu vrednost eksplicitno zavisnu od vremena merenja t_j (pri posmatranju jedne grupe l, τ_l je isto za sve jedinke u modelu i očekivanja zavise od vremena kroz γ_j i $(\tau\gamma)_{lj}$). Prema modelu (5.2), zaključujemo da putanja očekivanih vrednosti treba biti prava linija. Naša najbolja procena za tu putanju bi bila zasnovana na proceni odsečka i nagiba β_0 i β_1.

Model (5.2) može biti zapisan i u matričnom obliku. Ako sa $\boldsymbol{Y}_i$ označimo vektor podataka dimenzije $(n \times 1)$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},$$

možemo zapisati model kao

$$\boldsymbol{Y}_i = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_i. \qquad (5.3)$$

Ovako zapisan model ima oblik modela (5.1). Matrica $\boldsymbol{X}$ je ista za sve jedinke jer su sve jedinke posmatrane u istih n trenutaka.

Da bismo dovršili model, takođe moramo da damo pretpostavku o kovarijansnoj matrici za slučajni vektor $\boldsymbol{\epsilon}_i$. Kao u klasičnim modelima, možemo pretpostaviti da je ova matrica ista za sve vektore podataka:

$$var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}.$$

Pošto ćemo razmatrati situaciju samo jedne populacije, prirodno je da ova matrica bude ista za sve jedinke.

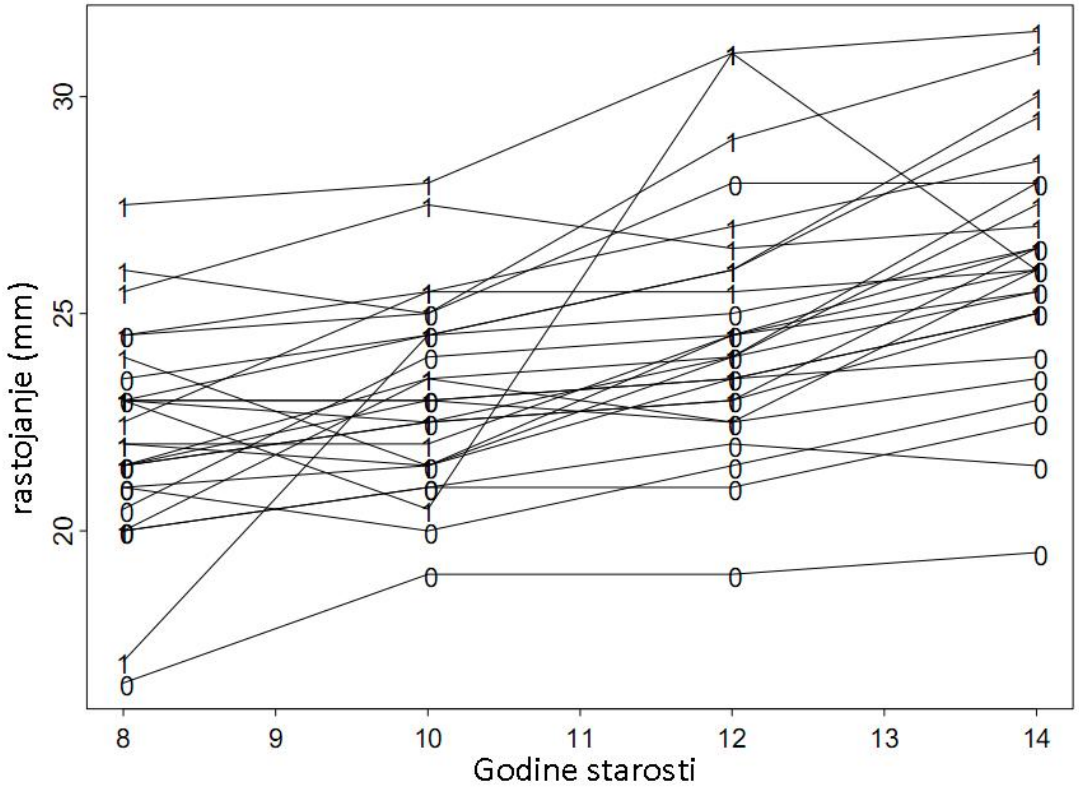
Pretpostavimo da rezultati Y_{ij} imaju normalnu raspodelu u svakom momentu, tako da vektor $\boldsymbol{Y}_i$ sadrži više slučajnih promenljiih sa normalnom raspodelom. Dakle, možemo predstaviti model kao

$$\boldsymbol{Y}_i \sim \mathcal{N}_n(\boldsymbol{X\beta}, \boldsymbol{\Sigma}),$$

gde su $\boldsymbol{X}$ i $\boldsymbol{\beta}$ već definisani.

Sad razmotrimo model ovog istraživanja koji obuhvata obe grupe. Dakle, na svakom detetu se meri rastojanje (mm) od hipofize do pukutine u lobanji iznad vilice u uzrastima od 8, 10, 12 i 14 godina i rezultati su dati na slici 8.

Slika 8. Merenja udaljenosti za svako dete po starosnim godinama. Sa (0) su označene devojčice, a sa (1) dečaci.



Imamo $q = 2$ grupe, devojčice i dečake. Sa slike 7 vidimo da za svaku grupu, očekivane vrednosti za svaku starosnu godinu padaju na pravu liniju, ali je prava linija drugačija u zavisnosti od grupe (pol). Ovo sugeriše na to da ako je jedinka i devojčica, možemo pisati

$$Y_{ij} = \beta_{0,devojčice} + \beta_{1,devojčice} t_j + \epsilon_{ij}, \tag{5.4}$$

gde su $\beta_{0,devojčice}$ i $\beta_{1,devojčice}$ odsečak i nagib koji opisuju očekivanu vrednost u svakom momentu za devojčice, kao funkcija od vremena. Slično, ako je jedinka i dečak, možemo pisati

$$Y_{ij} = \beta_{0,dečaci} + \beta_{1,dečaci} t_j + \epsilon_{ij}, \tag{5.5}$$

gde su $\beta_{0,dečaci}$ i $\beta_{1,dečaci}$ odsečak i nagib, verovatno drugačiji od $\beta_{0,devojčice}$ i $\beta_{1,devojčice}$. Definišemo za i - tu jedinku:

$$\delta_i = \begin{cases} 0, & \text{ako je jedinka } i \text{ devojčica} \\ 1, & \text{ako je jedinka } i \text{ dečak} \end{cases}.$$

Sada jednakosti (5.4) i (5.5) možemo pisati zajedno kao

$$Y_{ij} = (1-\delta_i)\beta_{0,devojčice} + \delta_i\beta_{0,dečaci} + (1-\delta_i)t_j\beta_{1,devojčice} + \delta_i t_j \beta_{1,dečaci} + \epsilon_{ij} \tag{5.6}$$

Ovo se može prikazati i u matričnom zapisu. Kompletan skup odsečaka i nagiba $\beta_{0,devojčice}$, $\beta_{1,devojčice}$, $\beta_{0,dečaci}$, i $\beta_{1,dečaci}$ karakteriše očekivane vrednosti za obe grupe. Tako da ćemo definisti parametarski vektor

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{0,devojčice} \\ \beta_{1,devojčice} \\ \beta_{0,dečaci} \\ \beta_{1,dečaci} \end{bmatrix}. \tag{5.7}$$

Dalje ćemo definisati

$$\boldsymbol{X}_i = \begin{bmatrix} (1-\delta_i) & (1-\delta_i)t_1 & \delta_i & \delta_i t_1 \\ \vdots & \vdots & \vdots & \vdots \\ (1-\delta_i) & (1-\delta_i)t_n & \delta_i & \delta_i t_n \end{bmatrix}; \tag{5.8}$$

jednostavno se vidi da ako je i devojčica imamo

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & 0 & 0 \end{bmatrix},$$

a ako je i dečak

$$\boldsymbol{X}_i = \begin{bmatrix} 0 & 0 & 1 & t_1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & t_n \end{bmatrix}.$$

$\boldsymbol{X}_i\boldsymbol{\beta}$ daje vektor dimenzije $(n \times 1)$ čiji su elementi $\beta_{0,devojčice} + \beta_{1,devojčice} t_j$ ili $\beta_{0,dečaci} + \beta_{1,dečaci} t_j$, u zavisnosti od toga da li je i devojčica ili dečak. Tako da ovaj model možemo pisati kao

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i,$$

gde je $\boldsymbol{\beta}$ definisan u (5.7), a $\boldsymbol{X}_i$ u (5.8).

Primetimo da je matrica $\boldsymbol{X}_i$ drugačija u zavisnosti od članstva u grupi i $\boldsymbol{\beta}$ sadrži sve parametre koji opisuju putanju očekivane vrednosti za obe grupe. Matrica $\boldsymbol{X}_i$ nije punog ranga (dečaci nemaju informacije o očekivanoj vrednosti devojčica i obrnuto).

Pretpostavimo da $\boldsymbol{Y}_i$ sadrži više slučajnih promenljiih sa normalnom raspodelom. Sa dodatnom pretpostavkom o kovarijansnoj matrici $var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}$ za svako i, važi

$$\boldsymbol{Y}_i \sim \mathcal{N}_n(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

Opisani model može da bude još uopšteniji. Razmotrimo primere za nebalansirane podatke.

5.2. Modeli za nebalansirane podatke

Opisan slučaj za modelovanje očekivane vrednosti se može generalizovati. Razmotrićemo mogućnosti kada imamo:

- Različite trenutke posmatranja za svaku jedinku,
- Različite brojeve posmatranja za svaku jedinku,

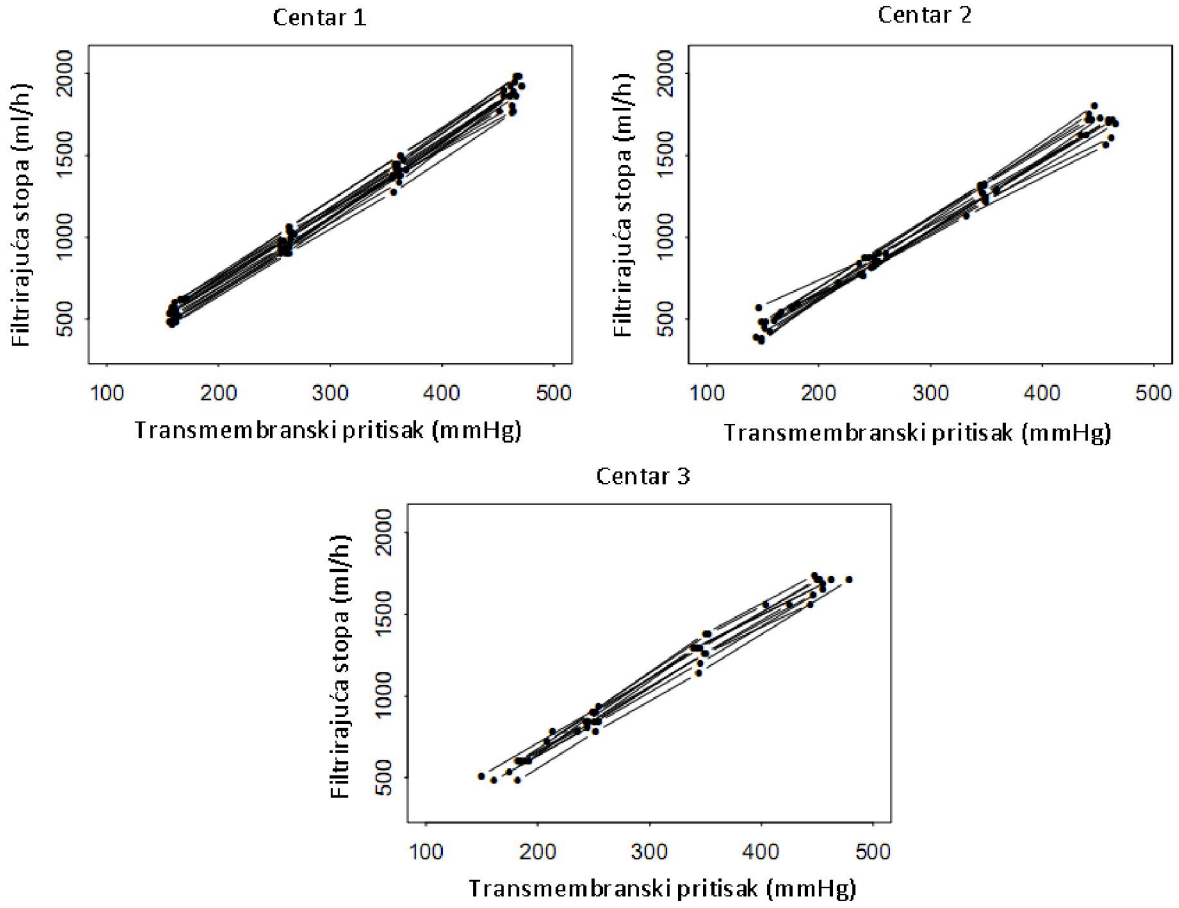
5.2.1. Slučaj kada su različiti trenuci posmatranja za svaku jedinku

Primer 2: *Filtriranje podataka za dijalizatore koji imaju membranu sa malim porama*[9]

Dijalizatori sa malim porama se koriste za lečenje pacijenata koji su u kasnoj fazi bubrežne bolesti, da bi uklonili višak tečnosti i otpada iz njihove krvi. Odnos filtrirajuće stope (ml/h) kojom je tečnost uklonjena i transmembranskog pritiska (mmHg) primenjenog preko membrane dijalizatora, prati pravu liniju. Istraživanje je sprovedeno da uporedi prosečne filtrirajuće stope dijalizatora za tri centra. Posmatramo $m = 41$ dijalizator i to samo dijalizatore koji su korišćeni na pacijentima. Eksperiment uključuje snimanje filtrirajućih stopa na nekoliko transmembranskih pritisaka za svaki dijalizator. Slika 9 pokazuje individualni profil dijalizatora u svakom centru. Značajno svojstvo slike 9 je to da četiri pritiska (trenuci posmatranja), $n = 4$, u kojima se posmatra svaki dijalizator, ne moraju biti isti. Dakle, i-ti dijalizator ima svoj skup trenutaka posmatranja $t_{ij},\ j = 1, 2, 3, 4$. Dakle, ne možemo izračunati očekivane vrednosti uzorka zato što svaki dijalizator vidi različite pritiske. Ipak, ako zamislimo očekivane vrednosti za svaku sliku tokom svih trenutaka, očekivane vrednosti će verovatno biti približno pravolinijske.

[9] Podaci su preuzeti iz knjige Vonesh, Chinchilli, *Linearni i nelinearni modeli za analizu ponovljenih merenja* (1997.)

Slika 9. Profil dijalizatora (za 41 dijalizator) u 3 centra



$$\begin{aligned} Y_{ij} &= \beta_1 + \beta_2 t_{ij} + \epsilon_{ij}, \quad \text{dijalizator } i \text{ u centru 1} \\ Y_{ij} &= \beta_3 + \beta_4 t_{ij} + \epsilon_{ij}, \quad \text{dijalizator } i \text{ u centru 2} \\ Y_{ij} &= \beta_5 + \beta_6 t_{ij} + \epsilon_{ij}, \quad \text{dijalizator } i \text{ u centru 3} \end{aligned} \tag{5.9}$$

β_1, β_3, β_5 su odsečci, a β_2, β_4, β_6 nagibi za očekivane vrednosti (prave linije) svakog centra. Definišimo

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_6 \end{bmatrix},$$

i posebne $\boldsymbol{X}_i$ matrice za svaku jedinku formata $(n \times 1)$, na osnovu članstva u grupi i jedinstvenog skupa trenutaka t_{ij}. Na primer, za jedinku i iz centra 1 matrica je

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in} & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Na taj način model (5.9) možemo zapisati kao

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i,$$

gde je $\boldsymbol{X}_i$ definisano za svaku jedinku posebno.

Ovaj model je moguće predstaviti i na drugačiji način. Možemo, na primer, da uporedimo stope promene očekivane vrednosti rezultata tokom vremena (pritisak) između grupa. U našem slučaju, želimo da uporedimo tri nagiba β_2, β_4 i β_6. Definisaćemo:

$$\delta_{i1} = \begin{cases} 1, & \text{jedinka } i \text{ iz centra 1} \\ 0, & \text{inače} \end{cases}$$

$$\delta_{i2} = \begin{cases} 1, & \text{jedinka } i \text{ iz centra 2} \\ 0, & \text{inače} \end{cases}.$$

Sada možemo zapisati model kao

$$Y_{ij} = \beta_1 + \beta_2\delta_{i1} + \beta_3\delta_{i2} + \beta_4 t_{ij} + \beta_5\delta_{i1}t_{ij} + \beta_6\delta_{i2}t_{ij} + \epsilon_{ij}. \quad (5.10)$$

Još uvek ima ukupno 6 parametara, ali parametri iz (5.10) imaju potpuno drugačiju interpretaciju od onih iz prvog modela. Jednostavnim uključivanjem u vrednosti δ_{i1} i δ_{i2} za svaki centar, važi sledeće:

Centar	Odsečak	Nagib
1	$\beta_1 + \beta_2$	$\beta_4 + \beta_5$
2	$\beta_1 + \beta_3$	$\beta_4 + \beta_6$
3	β_1	β_4

Primetimo da β_2 i β_3 predstavljaju razlike u odsečcima između Centara 1 i 3, tj. između Centara 2 i 3, a β_1 je odsečak za Centar 3. Slično tome, β_5 i β_6 predstavljaju razlike u nagibima između Centara 1 i 3, tj. između Centara 2 i 3, a β_4 je nagib za Centar 3. Ova parametrizacija nam omogućava da direktno procenimo razlike koje nas interesuju. Isti tip parametrizacije se koristi u običnoj linearnoj regresiji iz sličnih razloga. Različite parametrizacije nam daju iste modele, tako da je jedina stvar koja se razlikuje upravo interpretacija parametara.

Ovde su stvarne vrednosti vremena posmatranja različita za svaku jedinku. Razmotrimo modele za kovarijansnu matricu:

- Kod nestruktuiranog kovarijansnog modela sve jedinke koje su posmatrane u različitom skupu vremena, ne mogu deliti iste kovarijansne parametre, pa matrica $\mathbf{\Sigma}_i$ mora zavisiti od potpuno različitih veličina za svako i.
- Složeni simetrični kovarijansni model ne obraća pažnju na stvarne vrednosti vremena.
- Kod modela "jedna zavisna" možemo smatrati da su u korelaciji posmatranja koja su susedna u vremenu, bez obzira na to koliko ona mogu biti udaljena, dok ona koja su udaljenija nisu u korelaciji. Međutim, moguće je da je rastojanje za susedna posmatranja za jednu jedinku duže od rastojanja za nesusedna posmatranja za drugu jedinku, tako da ovo izgleda prilično besmisleno.
- Pretpostavka Autoregresivnog modela reda 1 (posmatranja su ravnomerno raspoređena u vremenu) je neprikladna zbog istog razloga.
- Pretpostavka Markovog modela (posmatranja su neravnomerno raspoređena u vremenu) više obećava u ovoj situaciji. Korelacija između posmatranja u okviru jedinke će zavisiti od vremenskih udaljenosti između posmatranja u okviru jedinke.

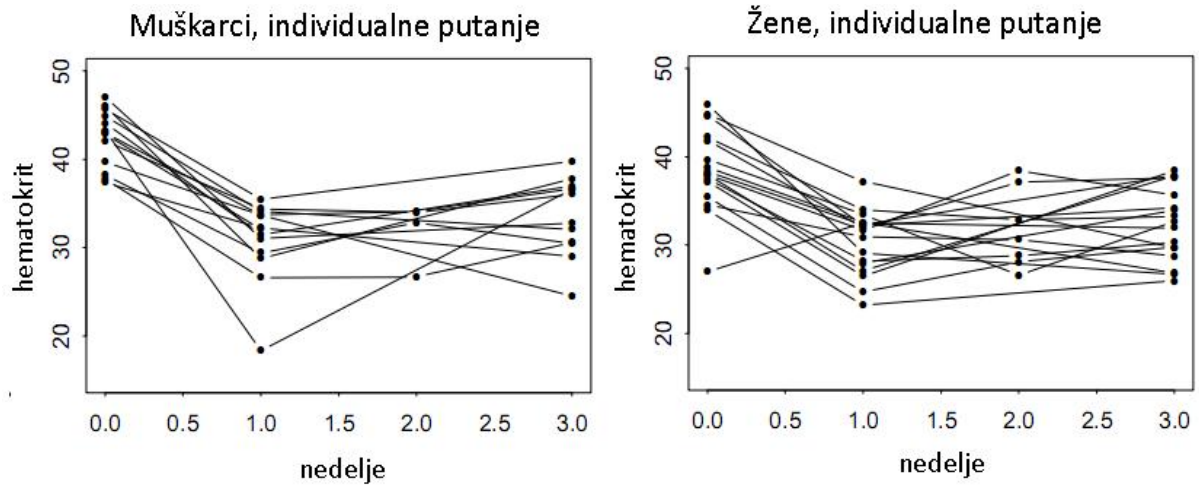
5.2.2. Slučaj kada izostaju pojedina merenja

Primer 3: *Studija zamene kuka*[10]

30 pacijenata je podvrgnuto operaciji zamene kuka, 13 muškaraca i 17 žena. Hematokrit, odnos količine crvenih krvnih zrnca prema količini ukupne krvi snimljeni u procentima, treba da se meri za svakog pacijenta pre zamene kuka u nedelji 0, pa onda u nedeljama 1 i 2, zatim posle zamene kuka u nedelji 3.

Potrebno je da se utvrdi da li postoje razlike u očekivanoj vrednosti rezultata, ako pratimo zamene kuka za muškarce i žene. Na slici 10 je prikazan špageti grafik za profil svakog pacijenta.

Slika 10. Putanje hematokrita za pacijente zamene kuka (Individualni profili po polu)



Sa slike 10 se može videti da određen broj i ženskih i muških pacijenata propušta merenja u nedelji 2. Mi ćemo pretpostaviti da razlog za ovaj nedostatak posmatranja nema nikakve veze sa našim primarnim interesom, tj. polom.

Dakle, imamo situaciju u kojoj su vektori podataka $\boldsymbol{Y}_i$ različite dužine za različite subjekte. Konkretno, sada imamo da je $\boldsymbol{Y}_i$ dimenzije $(n_i \times 1)$, gde je n_i broj posmatranja za subjekat i. Prema tome, ukupan broj posmatranja svih subjekata je

$$N = \sum_{i=1}^{m} n_i.$$

Da bismo utvrdili odgovarajuću reprezentaciju za očekivanu vrednost vektora podataka za svaku grupu, možemo izračunati očekivane vrednosti uzorka u svakom trenutku za muškarce i žene. Međutim, zbog nedostatka posmatranja, očekivane vrednosti uzorka u različitim trenucima će biti različitog kvaliteta.

Sa slike vidimo da očekivane vrednosti za oba pola ne padaju na pravu liniju. Za skoro sve pacijente, očitavanje pre zamene je visoko. Posle zamene, hematokrit opada, a zatim polako skače tokom naredne tri nedelje, tako da odnos između očekivanih vrednosti i vremena može izgledati više kao kvadratna funkcija od vremena, pa model ima oblik:

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \epsilon_{ij}, \quad \text{za muškarce} \\ Y_i &= \beta_4 + \beta_5 t_{ij} + \beta_6 t_{ij}^2 + \epsilon_{ij}, \quad \text{za žene.} \end{aligned} \tag{5.11}$$

[10] Podaci su dobijeni iz knjige Crowder, M.J. i Hand, D.J. *Analiza ponovljenih merenja* (1990.)

U (5.11) je moguće da vremena za svako i nisu ista, pa pišemo t_{ij}. Za naš skup podataka, vremena su ista za svaki subjekat. Međutim, kao što smo videli u primeru sa dijalizatorima, ovo ne mora biti slučaj.

Da bismo zapisali model u matričnom obliku, prvo ćemo definisati

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_6 \end{bmatrix}.$$

Jasno, matrica $\boldsymbol{X}_i$ za datu jedinku će zavisiti od vremena posmatranja te jedinke i imaće broj vrsta n_i, svaka vrsta odgovara jednom od n_i elemenata od Y_{ij}. Na primer, za muškarca sa n_i posmatranja imamo

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & 0 & 0 & 0 \end{bmatrix}.$$

Sada možemo zapisati model kao

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad (n_i \times 1),$$

gde je $\boldsymbol{X}_i$ matrica formata $(n_i \times 6)$ definisana za odgovarajuću jedinku i.

Za definisanje kovarijansne matrice ovde moramo biti malo oprezniji. Zato što se sada nosimo sa vektorima podataka $\boldsymbol{Y}_i$ različitih dužina n_i. Odgovarajuće kovarijansne matrice moraju da budu dimenzije $(n_i \times n_i)$. Dakle, nije moguće pretpostaviti da je kovarijansna matrica svih vektora podataka identična za svako i, pa ćemo pisati

$$var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_i.$$

Indeks i ukazuje na to da kovarijansna matrica zavisi od i preko svoje dimenzije n_i.

Na primer, pretpostavimo da postoji složena simetrija takva da sva posmatranja Y_{ij} imaju istu ukupnu varijansu σ^2, recimo da su i svi u podjednakoj korelaciji, bez obzira na to u kom vremenskom trenutku su uzete. Dakle, ovo će biti dobar izbor čak i za situaciju u kojoj su vremena različita za različite jedinke, kao u primeru sa dijalizatorima ili zbog nedostatka posmatranja. Da bi ovo predstavili, imali bi drugi parametar ρ. Za vektor podataka dužine n_i, bez obzira kada su uzeta ta n_i posmatranja u vremenu, matrica $\boldsymbol{\Sigma}_i$ bi bila matrica formata $(n_i \times n_i)$:

$$\boldsymbol{\Sigma}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

Bez obzira na dimenziju ili vremenske trenutke, pod ovom pretpostavkom, matrica $\boldsymbol{\Sigma}_i$ bi zavisila od 2 parametra: σ^2 i ρ, za sve i i zavisi od i samo zbog dimenzije.

Sa pretpostavkom da $\boldsymbol{Y}_i$ sadrži više slučajnih promenljiih sa normalnom raspodelom, možemo pisati model kao

$$\boldsymbol{Y}_i \sim \mathcal{N}_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i).$$

Zaključak

U ovom radu bavili smo se modelima i metodama analize longitudinalnih podataka (podaci u vidu ponovljenih merenja na istom subjektu tokom vremena ili u odnosu na neki drugi uslov). Predstavili smo neke modele korisne u analizi takvih podataka. Razmatrali smo varijacije između jedinki i unutar jedinki. Odstupanja između jedinki se mogu izolovati, pa da se pažnja usredsredi na analizu efekta tretmana.

Jedan od razloga zašto su longitudinalne studije danas popularne jeste što se veruje da rešavaju problem uzročnosti. To je međutim samo delimično tačno, jer longitudinalne studije zadovoljavaju samo jedan od bitnih kriterijuma za uzročnost, a to je pravilo prolaznosti (uzročnost koja se odnosi na efekat vremena). Longitudinalna studija je s jedne strane skupa, često vremenski zahtevna i teška za analizu, a s druge strane je ekonomična jer svaka jedinka ima značajan doprinos pa je potrebno manje jedinki za analizu. Ipak, glavna prednost longitudinalne studije u odnosu na trenutna posmatranja jeste da se može analizirati individualni razvoj određene promenljive u odnosu na određeni faktor, a takođe se individualni razvoj određene promenljive može uporediti sa individualnim razvojem drugih promenljivih.

Prve dve glave obuhvataju osnovne pojmove i modele, kao uvod u modelovanje longitudinalnim podacima, što nam je kasnije u radu bilo od koristi, i modele za korelacionu i kovarijansnu matricu, kako za balansirane, tako i za nebalansirane podatke.

Bavili smo se klasičnim metodama analize normalno raspoređenih, balansiranih ponovljenih merenja: metodama univarijantne i multivarijantne analize varijanse.

Univarijantne metode su opisane u Glavi 3. Obuhvaćene su osnovne situacije, statistički model za njih, pitanja od interesa i statističke hipoteze, analiza varijanse kroz primere, zatim šta se dešava ukoliko je narušena pretpostavka o obliku kovarijansne matrice i kada koristimo prilagođene testove. Ove metode imaju previše ograničavajuće pretpostavke o kovarijansi za većinu problema longitudinalnih podataka, dok multivarijantni metodi imaju opštije pretpostavke.

U Glavi 4 je predstavljena multivarijantna analiza varijanse koja obuhvata veći broj obeležja i posmatranje simultanih međuzavisnosti među promenljivima. Ovim metodama smo postigli pojednostavljivanje složene strukture kovarijansne matrice u cilju lakše interpretacije, a koristimo ih u procesu zaključivanja, tako što ocenjujemo stepen međuzavisnosti promenljivih i/ili testiramo njihovu statističku značajnost. U ovoj glavi je objašnjena Hotellingova T^2 statistika, MANOVA sa jednim faktorom i analiza profila. Objašnjena je osnovna razlika univarijantnih i multivarijantnih metoda analize varijanse, a zatim i nedostaci i ograničenja takvih metoda.

U poslednjoj glavi predstavljena je analiza nekoliko slučajeva realnih longitudinalnih podataka, koristeći regresionu analizu. Analizirali smo modele kada su podaci balansirani (kada su jedinke iz jedne grupe i kada su jedinke iz dve grupe) i kada su podaci nebalansirani (kada imamo različite trenutke posmatranja za svaku jedinku i različite brojeve posmatranja jedinki – izostavljena merenja).

Ne koriste sve longitudinalne studije iste statističke metode, a izbor odgovarajućeg metoda analize longitudinalnih podataka zavisi od mnogih faktora i opredeljen je pre svega vrstom problema, tipom podataka, karakteristikama metode i naravno ciljem istraživanja.

U velikom broju disciplina povećana je svest o značaju longitudinalnih studija za proučavanje promena i faktora koji utiču na promene. Analiza longitudinalnih podataka nastavlja da predstavlja mnogo zanimljivih metodoloških izazova koji će se verovatno rešiti u doglednoj budućnosti.

LITERATURA

[1] M. Davidian, *Applied Longitudinal Data Analysis,* Department of Statistics North Carolina State University, 2005.

[2] Michikazu Nakai, Weiming Ke, *Statistical Models for Longitudinal Data Analysis,* Applied Mathematical Sciences, 2009.

[3] Jos W. R. Twisk, *Applied Longitudinal Data Analysis for Epidemiology*, Cambridge University Press, 2003.

[4] Judith D. Singer, John B. Willett, *Applied Longitudinal Data Analysis – Modeling Change and Event Occurrence*, Oxford University Press, 2003.

[5] http://en.wikipedia.org/wiki/Mauchly%27s_sphericity_test

[6] Zlatko J. Kovačić, *Multivarijaciona analiza*, Univerzitet u Beogradu, Ekonomski fakultet, Beograd 1994.

[7] Biljana Bujić, *Analiza longitudinalnih podataka*, diplomski rad, Prirodno – matematički fakultet Novi Sad, 2010.

[8] http://www.oxfordjournals.org/our_journals/tropej/online/ma_chap10.pdf

[9] http://en.wikipedia.org/wiki/Longitudinal_study

[10] C. S. Davis, *Statistical Methods for the Analysis of Repeated Measurements*, New York: Springer, 2002.

[11] http://en.wikipedia.org/wiki/Contrast_%28statistics%29

[12] http://en.wikipedia.org/wiki/Repeated_measures_design

[13] Potthoff i Roy, *Generalizovani model višefaktorske analize varijanse posebno korisno za probleme krive rasta,* 1964.

[14] Vonesh, Chinchilli, *Linearni i nelinearni modeli za analizu ponovljenih merenja*, 1997.

[15] Crowder, M.J. i Hand, D.J. *Analiza ponovljenih merenja,* 1990.

[16] Prem S. Mann, *Uvod u statistiku*, Univerzitet u Beogradu, Ekonomski fakultet, Beograd, 2009.

[17] http://en.wikipedia.org/wiki/Multivariate_analysis_of_variance

[18] http://en.wikipedia.org/wiki/Statistical_model

[19] http://en.wikipedia.org/wiki/Analysis_of_variance#Characteristics_of_ANOVA

[20] http://www.psych.utoronto.ca/courses/c1/chap14/chap14.html

[21] http://homepages.gold.ac.uk/aphome/spheric.html

Kratka biografija

Zovem se Maja Zobenica i rođena sam 9. 4. 1988. godine u Somboru. Osnovnu školu "Nikola Vukićević" u Somboru sam završila 2003. g. i to kao nosilac diplome "Vuk Karadžić" i đak generacije. Srednju Ekonomsku školu u Somboru završila sam 2007. godine kao nosilac diplome "Vuk Karadžić" i đak generacije. Iste godine upisala sam Prirodno – matematički fakultet u Novom Sadu, smer matematika finansija. Osnovne studije sam završila u junu 2011. godine sa prosekom 9,12. U septembru 2011. godine zaposlila sam se na Pedagoškom fakultetu u Somboru, gde i sad radim kao saradnik u nastavi matematike. Master studije primenjene matematike sam upisala u oktobru 2011.

Novi Sad, 30. 5. 2013. Maja Zobenica

UNIVERZITET U NOVOM SADU
PRIRODNO MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
RBR

Identifikacioni broj:
IBR

Tip dokumentacije: *Monografska dokumentacija*
TD

Tip zapisa: Tekstualni štampani materijal
TZ

Vrsta rada: *Master rad*
VR

Autor: *Maja Zobenica*
AU

Mentor: *dr Zagorka Lozanov - Crvenković*
MN

Naslov rada: *Analiza logitudinalnih podataka*
NR

Jezik publikacije: *srpski (latinica)*
JP

Jezik izoda: *s/en*
JI

Zemlja publikovanja: *R Srbija*
ZP

Uže geografsko područje: *Vojvodina*
UGP

Godina: *2013.*
GO

Izdavač: *Autorski reprint*
IZ

Mesto i adresa: *Prirodno – matematički fakultet, Novi Sad, Trg D. Obradovića 4*
MA

Fizčki opis rada: *(5/82/21/8/10/0/0)*
FO *(broj poglavlja, strana, lit.citata, tabela, slika, grafika,priloga)*

Naučna oblast: *Matematika*
NO

Naučna disciplina: *Statistika*
ND

Predmetne odrednice, ključne reči: *Analiza longitudinalnih podataka, balansirani i nebalansirani podaci, slučajna varijacija među jedinkama, slučajna varijacija unutar jedinke, kovarijansna matrica, korelaciona matrica, složena simetrija, interakcija grupa i vremena, glavni efekti grupa, glavni efekti vremena, analiza varijanse, kontrasti, analiza profila, Hotellingova T^2 statistika.*
PO
UDK

Čuva se: *U biblioteci Departmana za matematiku i informatiku*
ČU

Važna napomena:
VN

Izvod: *Cilj ovog rada je da pruži pregled statističkih modela i metoda koji se koriste za analizu longitudinalnih podataka, a to su podaci u obliku ponovljenih merenja na istim jedinkama tokom vremena ili u odnosu na neki drugi skup uslova. Pitanja od naučnog interesa često uključuju ne samo uobičajena pitanja kao što su: kako se očekivanja rezultata merenja razlikuju po tretmanima, nego i kako se promena očekivanja rezultata razlikuje tokom vremena, kao i druga pitanja koja razmatraju odnose između rezultata i vremena. Opisane su različite statističke metode za rešavanje takvih pitanja i predstavljena je njihova primena u nekoliko primera. Glavne teme koje su razmatrane su različiti izvori varijacije longitudinalnih podataka, različite korelacione structure, univarijantni i multivarijantni pristupi analize longitudinalnih podataka.*
IZ

Datum prihvatanja teme od strane NN veća: *10. 05. 2012.*
DP

Datum odbrane:
DO

Članovi komisije:
Predsednik: *Dr Ljiljana Gajić, redovni professor Prirodno-matematičkog fakulteta u Novom Sadu*

Član: *Dr Zagorka Lozanov – Crvenković, redovni professor Prirodno-matematičkog fakulteta u Novom Sadu*

Član: *Dr Ivana Štajner Papuga, vanredovni professor Prirodno-matematičkog fakulteta u Novom Sadu*
KO

UNIVERSITY OF NOVI SAD
FACULTY OF NATURAL SCIENCES & MATHEMATICS
KEY WORDS DOCUMENTATION

Accession number:
ANO

Identification number:
INO

Document type: *Monograph documentation*
DT

Type of record: *Textual printed material*
TR

TR
Contents code: *Master thesis*
CC

Author: *Maja Zobenica*
AU

Mentor: *Dr Zagorka Lozanov - Crvenković*
MN

Title: *Analysis of longitudinal data*
TI

Language of text: *Serbian (Latin)*
LT

Language of abstract: *en/s*
LT

Country of publication: *R Serbia*
CP

Locality of publication: *Vojvodina*
LP

Publication year: *2013*
PY

Publisher: *Author's reprint*
PU

Publ. place: *Faculty of Natural Sciences and Mathematics Novi Sad, Trd D. Obradovića 4*
PP

Physical description: *(5/82/21/8/10/0/0)*
PD

Scientific field: *Mathematics*
SF

Scientific discipline: *Statistics*
SD

Subject Key words: *Analysis of longitudinal data, balanced data, unbalanced data, among-unit variation, within-units variation, covariance matrix, correlation matrix, compound symmetry, group by time interaction, main effect of groups, main effect of time, analysis of variance, contrasts, profile analysis, Hotelling's* T^2 *statistics.*
SKW
UC

Holding data: *Library of the Department of Mathematics and Computer Sciences*
HD

Note:
N

Abstract: *The goal of this paper is to provide an overview of statistical models and methods that are useful in the analysis of longitudinal data, and that is data in the form of repeated measurements on the same units over time or over some other set of conditions. The scientific questions of interest often involve not only the usual kinds of questions, such as how the mean response differs across treatments, but also how the change in mean response over time differs and other issues concerning the relationship between response and time. It is described different statistical methods for addressing such questions of scientific interest and its application to some examples. The main topics discussed are different sources of variation in longitudinal data, different correlation structures, univariate and multivariate approach for analysis of the longitudinal data.*
AB

Accepted on Scientific board on: *10. 05. 2012.*
AS

Defended: *2013.*
DE

Thesis Defend board:
President: *Dr Ljiljana Gajić, full profesor, Faculty of Natural Sciences and Mathematics, Novi Sad*

Member: *Dr Zagorka Lozanov - Crvenković, full profesor, Faculty of Natural Sciences and Mathematics, Novi Sad*

Member: *Dr Ivana Štajner Papuga, assistant professor, Faculty of Natural Sciences and Mathematics, Novi Sad*
DB