Matematički fakultet,
Univerzitet u Beogradu

Linearni statistički modeli

Regresioni modeli na longitudinalnim podacima

Natalija Lazić
Dejana Miladinović

decembar 2022.

Sadržaj

1 Longitudinalno istraživanje **3**
1.1 Istorijski razvoj metodologije . . . 4

2 Longitudinalni podaci **5**
2.0.1 Primer longitudinalne baze podataka . . . 6
2.1 Osnovne karakteristike longitudalnih podataka . . . 6
2.2 Matematička notacija . . . 7
2.2.1 Srednja vrednost . . . 8
2.2.2 Nezavisnost i korelacija . . . 8
2.3 Uzroci korelacije u longitudinalnim podacima . . . 10
2.4 Longitudinalni podaci i podaci preseka . . . 12

3 Linearni modeli na neprekidnim longitudinalnim podacima **14**
3.1 Matematička notacija . . . 14
3.1.1 Linearna regresija i promena srednjeg odgovora . . . 15
3.2 Pretpostavke o raspodeli modela . . . 16
3.2.1 Pretpostavke o raspodeli vektora slučajnih grešaka e_i . . . 16
3.2.2 Pretpostavke o raspodeli vektora neprekidnih odgovora Y_i . . . 16
3.3 Jednostavne opisne metode analize longitudinalnih podataka . . . 18
3.3.1 Grafička reprezentacija podataka . . . 19
3.3.2 Smooting techniques . . . 22
3.4 Modelovanje srednje vrednosti . . . 23
3.5 Modelovanje kovarijanse . . . 24
3.5.1 Nestruktuirana kovarijansa . . . 24
3.5.2 Modeli kovarijansnog uzorka - *CPM* . . . 25
3.5.3 Strukture kovarijanse sa slučajnim efektima . . . 25
3.6 Ocenjivanje regresionih koeficijenata i matrice kovarijanse greške . . . 26
3.6.1 Matrica kovarijanse greške . . . 26
3.6.2 Metod maksimalne verodostojnosti - *MLE* . . . 27
3.6.3 Metod ograničene maksimalne verodostojnosti - *REML* . . . 27
3.6.4 Ostali pristupi ocenjivanja parametara . . . 28

4 Regresioni modeli za analizu longitudinalnih podataka **29**
4.1 Modeli mešovitih efekata . . . 29
4.1.1 Ograničenja modela . . . 30
4.1.2 Korišćenje u R-u . . . 31
4.2 Model krive rasta . . . 33
4.2.1 Model slučajnih koeficijenata . . . 33
4.2.2 Ograničenja modela . . . 33
4.2.3 Korišćenje u R-u . . . 34
4.3 Metode za popravljanje modela . . . 36
4.3.1 Nedostajući podaci . . . 36
4.3.2 Uzorkovanje u različitim vremenskim intervalima . . . 38

5 Reziduali **39**
5.0.1 Transformisanje reziduala 39
5.1 Ispitivanje reziduala u R-u 40

6 Primer obrade longitudinalnih podataka u *R*-u **43**

7 Zaključak **49**

1 Longitudinalno istraživanje

Longitudinalna studija je istraživanje u okviru kog se posmatra jedna grupa ispitanika kroz vreme i na osnovu opservacija donose zaključci. Definišuća karakteristika longitudinalnog dizajna studije je da se merenja varijable odgovora vrše na istim pojedincima u nekoliko navrata.

Učesnici longitudinalne studije nazivaju se pojedincima ili subjektima i mogu biti ljudi ili životinje (npr. Laboratorijski miševi ili pacovi).

U longitudinalnoj studiji pojedinci se mere više puta u različitim prilikama ili vremenima. Broj ponovljenih opservacija i njihovo vreme mogu da variraju od jedne longitudinalne studije do druge. Na primer, kliničko ispitivanje može preduzeti ponovljene mere u određenim intervalima, dok opservaciona studija može da vrši merenja u nepravilnim intervalima. Kada su broj i vreme ponovljenih merenja isti za sve pojedince, bez obzira na to da li su prilike merenja podjednako ili nejednako raspoređene tokom trajanja studije, to se naziva *„uravnotežena"* longitudinalna studija. S druge strane, kada neki pojedinci propuste svoju zakazanu posetu ili datum posmatranja, to dovodi do *„neuravnotežene"* longitudinalne studije gde se ponovljena merenja ne dobijaju u uobičajenom nizu prilika. Neuravnoteženi longitudinalni dizajni su uobičajeni kada studija uključuje retrospektivno prikupljene podatke ili kada je vreme merenja definisano u odnosu na neki referentni događaj koji se dešava tokom perioda praćenja.

Iako dizajni koji su neuravnoteženi tokom vremena često nastaju zbog slučajnosti, ponekad ih planiraju istražitelji. U dizajnu studije *„rotirajući panel"* , koji se obično koristi u zdravstvenim istraživanjima kako bi se smanjio opterećenje odgovora, pojedinci se rotiraju u studiju i izlaze iz nje nakon davanja unapred određenog broja ponovljenih mera. Primarna motivacija za ovu vrstu dizajna studije je smanjenje troškova i ukupnog tereta učešća u studiji za svakog pojedinca, uz pružanje zapažanja u svakoj prilici za neki unapred određeni deo uzorka.

Podaci koji nedostaju su čest i izazovan problem u longitudinalnim studijama. Kada neka zapažanja nedostaju, podaci su nužno neuravnoteženi tokom vremena, pošto nemaju svi pojedinci isti broj ponovljenih merenja dobijenih u uobičajenom skupu prilika. Međutim, da bi se razlikovali podaci koji nedostaju u longitudinalnoj studiji od drugih vrsta neuravnoteženih podataka, takvi skupovi podataka se često nazivaju *„nepotpunim"*. Prilikom obrade podataka, biće potrebno pažljivo ispitati pretpostavke i prikladnost analize kako bi se utvrdila validnost zaključaka sa neuravnoteženim dizajnom i/ili podacima koji nedostaju.

Takođe, aspekt longitudinalnih podataka koji je istaknut u njihovoj statističkoj analizi je da su ponovljene mere na istoj osobi obično pozitivno korelisane. Korelisana zapažanja su pozitivna karakteristika longitudinalnih podataka jer daju preciznije procene stope promene ili efekta kovarijata na tu stopu promene nego što bi se dobilo iz jednakog broja nezavisnih posmatranja različitih pojedinaca. Ipak, korelacija između ponovljenih mera narušava osnovnu pretpostavku nezavisnosti koja je kamen temeljac velikog broja standardnih tehnika regresije.

1.1 Istorijski razvoj metodologije

Iako su naznake za korišćenje ove metodologije postojale i ranije, istraživanje statističkih metoda za dizajn i analizu ljudskih istraživanja eksplozivno se proširilo u drugoj polovini dvadesetog veka. Od ranih 1950-tih, američka vlada krenula je da ulaže u biomedicinska istraživanja, epidemiološke studije i klinička ispitivanja.

Fokus ranijih studija ovog tipa bio je identifikacija uzroka rane smrti pojedinaca i procena efikasnih tretmana za odlaganje iste.

Jedan od najpoznatijih primera longitudinalne studije koja je ostavila trag u istoriji nauke jeste *Heart Study Framingham*, tj. Framinghamska studija srca koja se zasnivala na istraživa- nju srca, srčanih bolesti i njihovih uzroka. U okviru ove studije učesnici su posećivali bolnicu u intervalima od dve godine. Važno je napomenuti da su ishodi preživljavanja tokom uzastopnih dvogodišnjih perioda tretirani kao nezavisni događaji,te modelovani korišćenjem višestruke logističke regresije.

Vremenom se, pored traženja uzroka za hronične bolesti u okviru rada sa longitudalnim podacima, pojavilo interesovanje i za nivoe i promene karakteristika koje dovode do istih. Naučna zajednica krenula je da analizira ponašanje faktora rizika. U okviru Framinghamske studije srca postavilo se pitanje da li je nivo krvnog pritiska u detinjstvu indikator hipertenzije u odraslom dobu. Takođe, istraživanje se proširilo na bolesti poput astme, artritisa i drugih koje nisu opasne po život, tj. na efekte tretmana tokom vremena u merama težine bolesti. Počelo je praćenje stanovništva svih uzrasta tokom vremena u okviru opservacionih i kliničkih ispitivanja u cilju razumevanja razvoja i postojanosti bolesti, te identifikacije faktora koji menjaju tok razvoja bolesti.

Usled razvoja kompjutera, došlo je do novih pristupa statističkoj analizi, te se predlaže upotreba EM algoritma za *fittovanje* klase mešovitih modela koji su adekvatni za analizu ponovljenih merenja, kao i drugih algoritama poput Fišerovog bodovanja ili Njutnove metode. Liang i Zeger su uveli generalizovane jednačine za procenu u bio-statističku literaturu i predložili porodicu generalizovanih linearnih modela za uklapanje ponovljenih posmatranja binarnih i izbrojanih podataka.

Mnogi drugi istraživači koji su pisali biomedicinsku, obrazovnu i psihometrijsku literaturu doprineli su brzom razvoju metodologije za analizu ovih *„longitudinalnih“* podataka. U proteklih 30 godina došlo je do značajnog napretka u razvoju statističkih metoda za analizu longitudinalnih podataka.

2 Longitudinalni podaci

Longitudinalni podaci se odnose na podatke koji se prikupljaju tokom vremena od istih pojedinaca ili grupa. Ova vrsta podataka omogucava analizu promena i trendova u podacima tokom vremena, pružajuci razumevanje fenomena koji se proučava. Longitudinalni podaci se obično koriste u istraživanjima iz oblasti psihologije, sociologije, epidemiologije, ekonomije, itd.

Jedna od glavnih prednosti longitudinalnih podataka je ta što omogucava istraživačima da prate promene u pojedincima ili grupama tokom vremena. Ovo može pružiti vredan uvid u razvoj pojedinaca ili grupa, kao i uticaj različitih faktora na njihovo ponašanje ili ishode. Na primer, longitudinalno proučavanje kognitivnog razvoja dece može pratiti njihov napredak tokom vremena, pružajuci detaljno razumevanje faktora koji utiču na njihov razvoj.

Još jedna prednost longitudinalnih podataka je što omogucava istraživačima da kontrolišu individualne razlike, što može biti glavni izvor pristrasnosti u studijama preseka. Pošto se longitudinalni podaci prikupljaju od istih pojedinaca ili grupa tokom vremena, istraživači mogu da kontrolišu individualne razlike koje mogu uticati na ishod studije. Ovo može pomoci da se poveca validnost nalaza i osigura da su rezultati robusniji.

Longitudinalni podaci se mogu prikupiti korišcenjem različitih metoda, uključujuci ankete, intervjue i zapažanja. Ankete su uobičajeni metod prikupljanja longitudinalnih podataka, jer omogucavaju istraživačima da prikupe veliku količinu podataka od velikog uzorka pojedinaca ili grupa. Intervjui, s druge strane, omogucavaju dublje i detaljnije prikupljanje podataka iz manjeg uzorka. Posmatranja su takođe korisna metoda za prikupljanje longitudinalnih podataka, jer omogucavaju istraživačima da posmatraju pojedince ili grupe tokom vremena u njihovom prirodnom okruženju.

Longitudinalni podaci se mogu analizirati korišcenjem metoda kao što su modeliranje krive rasta, analiza istorije događaja i analiza latentne krive rasta. Ove metode omogucavaju istraživačima da ispitaju promene u podacima tokom vremena i identifikuju obrasce ili trendove. Oni takođe omogucavaju istraživačima da ispitaju kako različiti faktori utiču na promene ili trendove koji se posmatraju. Postoji nekoliko tipova longitudinalnih podataka:

- *Panel podaci* - Poznati kao podaci na nivou pojedinca, prikupljaju se od istih pojedinaca u više vremenskih tačaka. Ova vrsta podataka se često koristi u društvenim naukama i ekonomiji za proučavanje promena i ponašanja na nivou pojedinca tokom vremena.

- *Podaci o vremenskim serijama* - Prikupljaju se u redovnim intervalima tokom vremena i često se koriste za proučavanje trendova i obrazaca u varijablama kao što su cene akcija, vremenske prilike i drugi fenomeni koji se menjaju tokom vremena.

- *Kohortni podaci* - Prikupljaju od pojedinaca koji dele zajedničku karakteristiku, kao što je starost ili geografska lokacija, i često se koriste u epidemiološkim i javnozdravstvenim istraživanjima za proučavanje učestalosti i prevalencije bolesti tokom vremena.

- *Podaci istorije događaja* - Prikupljaju o nastanku događaja, kao što je početak bolesti i vreme nastanka. Ova vrsta podataka se često koristi u analizi preživljavanja i drugim vrstama analiza vremena do događaja.

- *Podaci o ponovljenim merama* - Prikupljaju od istih pojedinaca u više vremenskih tačaka, ali vremenski intervali između posmatranja možda nece biti jednaki. Ova vrsta podataka se često koristi u kliničkim ispitivanjima i drugim vrstama medicinskih istraživanja.

2.0.1 Primer longitudinalne baze podataka

Za ilustraciju baze longitudinalnih podataka posmatraćemo skup podataka *sleepstudy* koji je deo paketa *lme4* u *R*-u. Ovaj skup podataka sadrži merenja vremena reakcije za 18 učesnika koji su bili lišeni sna u različito vreme. Kako je ovo longitudinalna studija, učesnici su ispitivani u više vremenskih tačaka, te je cilj studije bio da se ispita kako es vreme reakcije menjalo tokom vremena. Za početak učitavamo gorepomenutu bazu:

```
library(lme4)
Baza <- lme4::sleepstudy
head (Baza)
```

Baza podataka *sleepstudy* sastoji se od sledećih promenljivih:

- Reaction - *Vreme koje je potrebno ispitanicima da izvše neku reakciju.*
- Days - *Broj dana koje su proveli bez sna (0-9).*
- ID - *Identifikacioni broj ispitanika.*

Prikazujemo kako izgleda baza sa longitudalnim podacima (tj. njenih prvih 6 rezultata.

```
## Loading required package: Matrix

##   Reaction Days Subject
## 1 249.5600    0     308
## 2 258.7047    1     308
## 3 250.8006    2     308
## 4 321.4398    3     308
## 5 356.8519    4     308
## 6 414.6901    5     308
```

2.1 Osnovne karakteristike longitudalnih podataka

Kao što je rečeno, longitudinalni podaci se odnose na podatke koji se prikupljaju tokom vremena od istih pojedinaca ili grupa. Njihove osnovne karakteristike su:

- *Vremenski redosled* - Longitudinalni podaci se često predstavljaju kao skup zapažanja jednog pojedinca ili grupe pojedinaca, gde je svako posmatranje skup varijabli merenih u različito vreme. Ovo omogućava proučavanje promena varijabli tokom vremena i ispitivanje trendova i obrazaca.

- *Intra-individualna varijabilnost* - Longitudinalni podaci omogućavaju proučavanje individualnih razlika tokom vremena. Matematički, ovo se može predstaviti varijacijom vrednosti varijabli merenih tokom vremena za svakog pojedinca. Varijabilnost se može proučavati korišćenjem statističkih tehnika kao što su varijansa i standardna devijacija.

- *Grupisanje (Clustering)* - Grupisanje u longitudinalnim podacima odnosi se na činjenicu da pojedinci unutar grupe ili populacije mogu biti sličniji jedni drugima nego pojedincima u drugim grupama ili populacijama. Za analizu longitudinalnih podataka često se koriste posebne tehnike kao što su modeli mešovitih efekata ili modeli na više nivoa, koji uzimaju u obzir korelaciju unutar klastera.

- *Nezavisnost* - Longitudinalni podaci mogu imati zavisnost između posmatranja unutar klastera, što može dovesti do kršenja pretpostavki nezavisnosti među posmatranjima. Ovo se može predstaviti korelacionom strukturom između posmatranja unutar klastera.

- *Nedostajući podaci* - Longitudinalni podaci često sadrže podatke koji nedostaju zato što pojedinci napuštaju studiju ili ne daju potpune podatke u svakoj vremenskoj tački.

- *Visoka dimenzionalnost* - Longitudinalni podaci često imaju mnogo varijabli merenih u više vremenskih tačaka, što dovodi do visoke dimenzionalnosti podataka. Ovo se može predstaviti velikim brojem varijabli i zapažanja u skupu podataka, što može biti izazovno za analizu i tumačenje.

Longitudinalni podaci mogu pružiti informacije o pravcu i veličini promene, uzročnosti istih i za donošenje zaključaka o osnovnim mehanizmima koji pokreću ove promene mogu se koristiti specijalne tehnike poput modela meštovitih efekata ili analize vremenskih serija, jer je potrebno uključiti korelaciju kroz vreme i grupisanja. Matematički ovo se predstavlja modelima koji uzimaju u obzir vremensku strukturu i strukturu klastera podataka.

2.2 Matematička notacija

Za razumevanje rada neophodno je upoznati se sa matematičkom notacijom i oznakama koje će s u njemu koristiti. Neka Y_{ij} označava promenljivu odgovora i - tog ispitanika ($i = 1, ..., n$) u j-om merenju ($j = 1, ..., m$). Prilikom uvođenja ove notacije pretpostavljamo da su merenja izvršena u jednakim vremenskim intervalima. Kao što je već poznato, velikim slovima ćemo označavati slučajne veličine, a malim realizovane vrednosti.

U sledećoj tabeli u vidu dvodimenzionalnog niza predstavljeno je n zapažanja, tj. realizovanih vrednosti slučajne veličine Y_i, na N pojedinaca. Redovi ovog niza predstavljaju pojedince, dok kolone predstavljaju odgovor i-tog pojedinca u j-tom merenju.

	Broj opservacija				
Pojedinac	1	2	3	...	m
1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1m}
2	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2m}
...	...	...	...	...	...
n	y_{n1}	y_{n2}	y_{n3}	$\cdots$	y_{nm}

Uzevši u ozbir da su merenja ponovljena n puta, promenljivu odgovora možemo predstaviti kao vektor dimenzija $n \times 1$:

$$Yi = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ ... \\ Y_{in} \end{pmatrix} = \left(Y_{i1}, Y_{i2}, ...Y_{in}\right)^T \tag{1}$$

Kada posmatramo longitudinalnu studiju, prilikom analize podataka najviše se bavimo analizom srednjeg odgovora (*meanresponse*), tačnije promenom vrednosti srednjeg odgovora tokom vremena, te zavisnosti tih promena od kovarijanata (npr. nova terapija kod bolesnika).

2.2.1 Srednja vrednost

Srednju vrednost, tj. očekivanje svakog odgovora Y_{ij} označavamo sa: $\mu_i = E(Y_{ij})$, gde E predstavlja dugoročno očekivanje za veliku populaciju u j-tom slučaju. Kao što je već poznato, očekivanje slučajne veličine Y_{ij} je ponderisani prosek svih mogućih vrednosti Y_{ij}, pri čemu su ponderi verovatnoće pojavljivanja svake moguće vrednosti. U velikom broju longitudalnih studija glavni cilj je povezivanje promene vrednosti srednjeg odgovora sa kovarijantima, te kako srednji odgovor i njegove promene variraju od pojedinca u zavisnosti od parametara, potrebno je uvesti proširenu notaciju. U okviru nje očekivanje označava dugoročni prosek za veliku potpopulaciju subjekata koji dele slične vrednosti kovarijata (npr. subjekti koji su dodeljeni grupi aktivnog tretmana, neeksponirani subjekti) prilikom j-tog merenja. Tada je μ_{ij} uslovni srednji odgovor prilikom j-tog merenja.Matematički, gorepomenuto očekivanje označavaćemo sa:

$$\mu_{ij} = E(Y_{ij}) \tag{2}$$

2.2.2 Nezavisnost i korelacija

Zavisnost ili korelacija između odgovora iste osobe je važna u statistici. Ako su dve varijable nezavisne, raspodela jedne ne zavisi od druge. Standardne statističke metode pretpostavljaju da su posmatranja studije nezavisne slučajne varijable, što je razumno kada se zapažanja uzimaju od nasumično odabranih pojedinaca ili kada odgovori različitih pojedinaca nisu povezani. Međutim, ako se više zapažanja uzme od iste osobe, nezavisnost se više ne može pretpostaviti jer će prošli odgovori verovatno predviđati buduce. Ova zavisnost se može meriti korelacijom. Korelacija između ponovljenih mera je pozitivan aspekt longitudinalnih podataka jer pruža preciznije procene promene ili efekta varijabli na promene. Modeli za longitudinalne podatke koriste ovu korelaciju za tačnu procenu promena tokom vremena.

Kao što nam je već poznato, uslovno očekivanje, tj. srednja vrednost slučajne veličine Y_{ij} definisano je sa $\mu_{ij} = E(Y_{ij})$ i predstavlja meru lokacije centra raspodele Y_{ij}. Varijansa slučajne veličine predstavlja meru širenja ili disperzije vrednosti Y_{ij} oko njihovog uslovnog očekivanja (srednje vrednosti μ_{ij}). Varijansa slučajne veličine Y_{ij} je:

$$\sigma_j^2 = E(Y_{ij} - E(Y_{ij}))^2 = E(Y_{ij} - \mu_{ij})^2 \tag{3}$$

Posmatramo sada zavisnost među odgovorima u longitudinalnoj studiji. Kovarijansa između odgovora prilikom dva različita merenja Y_{ij} i Y_{ik} predstavlja meru linearne zavisnosti između Y_{ij} i Y_{ik} uzevši u obzir kovarijante, i definisana je sa:

$$\sigma_{jk} = E((Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})) \tag{4}$$

Kovarijansa, tj. zavisnost između Y_{ij} i Y_{ik} može biti pozitivna, negativna ili jednaka nuli (ne postoji linearna zavisnost između j-tog i k-tog odgovora s obzirom na kovarijante) i zavisi od mernih jedinica i sistema u kome se promenljive nalaze. Kovarijansa promenljive sa samom sobom (Y_{ij} i Y_{ij}) je varijansa promenljive.

Kako na osnovu kovarijanse ne postoji uniformisani sistem preciznog određivanja zavisnosti promenljivih, tj. veličina kovarijanse ne daje uvek precizne rezultate, rešenje za gorepomenuti problem nalazi se u standardizaciji. Naime, definišemo koeficijent korelacije u kome σ_j i σ_k predsta- vljaju standardne devijacije (kvadratni koren varijanse σ_j^2, tj. σ_k^2) slučajnih veličina Y_{ij} i Y_{ik}. Korelacija je mera zavisnosti koja je bez jedinica ili bez skala merenja, što je čini adekvatnijom za posmatranje zavisnosti između promenljivih. Definisana je sa:

$$\rho_{jk} = \frac{E((Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik}))}{\sigma_j \sigma_k} \tag{5}$$

Korelacija mora imati vrednosti između -1 i 1, tj. uzima tačno te vrednosti kada postoji savršen pravolinijski odnos između promenljivih (pozitivno, tj. negativno korelisane), a kako tačke odstupaju od savršenog pravolinijskog odnosa, korelacija se približava nuli.

- $\rho_{jk} \in (-1, 0)$ - Promenljive negativno korelisane, ako vrednost j-te raste, vrednost k-te opada i obrnuto.
- $\rho_{jk} = 0$ - Ne postoji linearna zavisnost između j-te i k-te promenljive, rastom ili opadanjem jedne promenljive druga ostaje ista.
- $\rho_{jk} \in (0, 1)$ - Promenljive pozitivno korelisane, ako vrednost j-te raste, raste i vrednost k-te i obrnuto.

Kako korelacija meri samo linearnu zavisnost, varijable mogu biti zavisne, ali nekorelisane, dok obratno nije moguće. Tj. Statistički nezavisne varijable će nužno biti nekorelisane.

U okviru analize longitudinalnih podataka pretpostavlja se da će prilikom ponovljenih merenja na ispitanicima odgovori biti pozitivno korelisani.

Posmatramo jednostavne longitudinalne koji su uravnoteženi i potpuni, sa n ponovljenih merenja varijable odgovora na N pojedinaca. Predstavićemo podatke dobijene u n ponovljenih merenja vektorom $Y_i = \left(Y_{i1}, Y_{i2}, ...Y_{in}\right)^T$ i definisati kovarijacionu matricu:

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ ... \\ Y_{in} \end{pmatrix} = \begin{pmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) & ... & Cov(Y_{i1}, Y_{in}) \\ Cov(Y_{i2}, Y_{i1}) & Var(Y_{i2}) & ... & Cov(Y_{i2}, Y_{in}) \\ ... & ... & ... & ... \\ Cov(Y_{in}, Y_{i1}) & Cov(Y_{in}, Y_{i2}) & ... & Var(Y_{in}) \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & ... & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & ... & \sigma_{2n} \\ ... & ... & ... & ... \\ \sigma_{n1} & \sigma_{n2} & ... & \sigma_{nn} \end{pmatrix} \tag{6}$$

Gde su $Cov(Y_{ij}, Y_{ik}) = \sigma_{jk}$. Pretpostavili smo i da su varijanse i kovarijanse konstantne među pojedincima. Kovarijaciona matrica je simetrična u smislu da su $Cov(Y_{ij}, Y_{ik}) = \sigma_{jk} = \sigma_{kj} = Cov(Y_{ik}, Y_{ij})$ i $\sigma_{kk} = Cov(Y_{ik}, Y_{ik}) = \sigma_k^2$. Dakle, kovarijacionu matricu Y_i možemo predstaviti i na sledeći način:

$$Cov(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & ... & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & ... & \sigma_{2n} \\ ... & ... & ... & ... \\ \sigma_{n1} & \sigma_{n2} & ... & \sigma_n^2 \end{pmatrix} \tag{7}$$

Definišemo sada i korelacionu matricu. Ona je simetrična u smislu da su $Corr(Y_{ij}, Y_{ik}) = \rho_{jk} = \rho_{kj} = Corr(Y_{ik}, Y_{ij})$, dok se na njenoj dijagonali nalaze jedinice koje predstavljaju korelisanost varijable sa samom sobom. Korelaciona matrica slučajne veličine Y_i:

$$Corr(Y_i) = \begin{pmatrix} 1 & \rho_{12} & ... & \rho_{1n} \\ \rho_{21} & 1 & ... & \rho_{2n} \\ ... & ... & ... & ... \\ \rho_{n1} & \rho_{n2} & ... & 1 \end{pmatrix} \tag{8}$$

Prilikom analize longitudinalnih podataka, očekuje se da korelacije budu pozitivne, a sekve- ncijalna priroda longitudinalnih podataka implicira da može postojati obrazac u korelacijama. Na primer, očekuje se da će par ponovljenih mera koje su se međusobno približavale u vremenu imati veću korelaciju od para ponovljenih mera koje su dalje razdvojene u vremenu. Tj. prilikom analize ovih podataka, očekuje se da ce korelacija između ponovljenih mera opasti sa povećanjem vremenskog odvajanja.

2.3 Uzroci korelacije u longitudinalnim podacima

Unatoč činjenici da su longitudinalni podaci često povezani, važno je razmotriti zbog čega je to slučaj i zašto su obično u pozitivnoj korelaciji. Tri glavna potencijalna izvora varijabilnosti koji utiču na korelaciju ponovljenih merenja kod ispitanika su:

- *Heterogenost između pojedinaca*

  Jedan od izvora pozitivne korelacije između ponovljenih mera je heterogenost ili varijabilnost varijable odgovora između pojedinaca u populaciji. Različite osobe reaguju različito na različite faktore, što dovodi do heterogenosti između subjekata i varijabilnosti unutar subjekta u longitudinalnim studijama. Očekuje se da će par ponovljenih mera na istoj osobi biti sličniji od pojedinačnih zapažanja dobijenih od dve nasumično odabrane osobe.Postoji pozitivna korelacija među longitudinalnim odgovorima jer očekujemo da ponovljeni odgovori iste osobe budu sličniji od odgovora različitih pojedinaca. Prilikom longitudinalne studije, ponavljanjem merenja možemo pronaći i obrazac opadanja korelacije sa povećanjem vremenskog odvajanja.

  U statističkim modelima za longitudinalne podatke, varijabilnost između pojedinaca može se objasniti uvođenjem „slučajnih efekata“ specifičnih za pojedinca (npr. nasumično promenljivi preseci i nagibi). Tačnije, da bi se uzela u obzir heterogenost između pojedinaca u sklonosti ka reagovanju, pretpostavlja se da neki efekti ili koeficijenti regresije u statističkim modelima variraju nasumično.

- *Biološka varijabilnost unutar pojedinca*

  Inherentna biološka varijabilnost mnogih zdravstvenih ishoda je važan izvor varijabilnosti koja utiče na korelaciju između longitudinalnih odgovora. Ova varijabilnost može biti rezultat cirkadijalnih ritmova ili drugih faktora kao što su temperatura, svetlost, godišnje doba, ishrana ili infekcija. Većina varijabli koje se odnose na zdravlje nemaju predvidljive ciklične obrasce, ali umesto toga, ponovljena merenja na bilo kojoj individui će varirati oko neke homeostatske zadate tačke na nasumičan način, što se naziva inherentna biološka varijabilnost unutar pojedinca. Ova vrsta varijabilnosti je evidentna u skoro svim izmerenim biološkim parametrima.

  Postoji neki osnovni biološki proces (ili kombinacija procesa) koji se menja tokom vremena na relativno gladak i kontinuiran način. Slučajnost ili odstupanja od osnovnog puta odgovora pojedinca će verovatno biti sličnija kada se merenja vrše veoma blizu u vremenu. Kao rezultat toga, ne može se pretpostaviti da su uzastopna merenja nezavisna, a merenja na istoj individui će biti sličnija kada su vremenski bliže, a manje slična kada su udaljenija. Ovo uvodi serijsku korelaciju između ponovljenih mera, što rezultira matricom korelacije koja ima karakterističnu strukturu gde se korelacija smanjuje kako se vremensko razdvajanje između ponovljenih mera povecava.

  Druga konceptualizacija bioloških varijacija unutar pojedinca je u smislu neuspeha da se tačno odredi put odgovora svakog pojedinca tokom vremena. Ako svaki pojedinac ima malo drugačiji put odgovora tokom vremena, onda će svaka pogrešna specifikacija ovih puteva odgovora izazvati korelaciju između ponovljenih mera.

- *Greška u merenju*

  Slučajna greška merenja može biti značajan izvor varijabilnosti u longitudinalnim podacima i uticati na tačnost procene efekata tretmana ili odnosa među varijablama. Ukoliko pretposta- vimo da je moguće izvršiti dva merenja u isto vreme na jednom ispitanku, ne možemo govoriti o biološkoj varijabilnosti, te, ukoliko se vrednosti ne poklapaju, možemo sumnjati da je u pitanju neprezicnost postupka merenja. Greška merenja prisutna je kod skoro svih studija, te je potrebno proceniti relativnu veličinu grešaka koje uzrokuje određena procedura merenja. Nju izražavamo putem koeficijenta pouzdanosti.

  Pouzdanost je, iz statističke perspective, stepen u kome su ponovljena merenja, sprovedena pod istim uslovima, slična. Kada bismo mogli, hipotetički, da dobijemo mnogo repliciranih merenja na pojedincu pod što je moguce bliže ujednačenim uslovima, rezultat bez greške bi se definisao kao prosek svih (hipotetičkih) ponovljenih merenja. Pouzdanost je tada proporcija ukupne varijabilnosti koja je posledica individualne varijabilnosti u pravim rezultatima. U drugim situacijama pouzdanost može zavisiti od heterogenosti pravih rezultata u populaciji (anketa na temu kvaliteta života), tako da možemo zaključiti da pouzdanost nije fiksna karakteristika merenja. Zbog toga je poželjno da se preciznost merenja izrazi direktno u smislu varijanse mernih grešaka ili alternativno njenog kvadratnog korena (standardna greška merenja).

  Efekat nepouzdanosti je da „ublaži“ ili smanji korelaciju između ponovljenih mera bliže nuli. Tj. što je veća varijansa grešaka merenja, to je veće slabljenje korelacije između

ponovljenih merenja u longitudinalnoj studiji.Stoga će upotreba manje pouzdane merne procedure ili instrumenta rezultirati ponovljenim merenjima sa manjim korelacijama nego da je korišćena pouzdanija merna procedura ili instrument.

Iskustvo sa mnogim longitudinalnim studijama u biološkim i zdravstvenim naukama je dovelo do nekoliko zapažanja o prirodi korelacije između ponovljenih merenja:

1. Korelacije su obično pozitivne.

   Direktna posledica heterogenosti između pojedinaca i bioloških varijacija u odgovoru tokom vremena. Ova dva izvora varijabilnosti deluju zajedno da indukuju pozitivnu korelaciju između ponovljenih mera.

2. Korelacije se smanjuju s povećanjem vremenskog razmaka između merenja.

   Direktna posledica inherentne biološke varijacije u odgovoru tokom vremena i/ili heterogenosti putanja odgovora između pojedinaca tokom vremena.

3. Korelacije između ponovljenih merenja retko dostižu nulu, čak i ako su rađena u razmaku od mnogo godina.

   Direktna posledica heterogenosti između pojedinaca u osnovnoj sklonosti da se odgovori, tj. sklonosti pojedinca da reaguje traje tokom svih ponovljenih merenja na toj osobi, bez obzira na to koliko su merenja udaljena u vremenu.

4. Korelacija između merenja koja su izvršena vrlo blizu jedna drugoj u vremenu retko dostiže jedan.

   Direktna posledica greške merenja. Korelacija između bilo kog para ponovljenih merenja, bez obzira na to koliko su bliske prilike merenja, ograničena je pouzdanošcu postupka merenja.

Iako ova tri izvora varijabilnosti postoje unutar longitudinalnih istraživanja, često se dešava da nema dovoljno podataka za njihovu individualnu procenu. Zato se oni često kombinuju u jednu komponentu varijanse unutar subjekta.

Ignorisanjem korelacije može doći do netačnih procena varijabilnosti uzorka (niska procena preciznosti, velike standardne greške, preširoki intervali poverenja), koje vode do pogrešnih zaključaka.

2.4 Longitudinalni podaci i podaci preseka

Longitudinalni podaci se odnose na podatke koji se prikupljaju tokom vremena od istih pojedina- ca ili grupa, dok se podaci preseka odnose na podatke koji se prikupljaju od grupa ili grupa pojedinaca. Obe vrste podataka mogu se koristiti za proučavanje promena i trendova tokom vremena, ali imaju neke ključne razlike.

Longitudinalni podaci se obično prikupljaju od istih pojedinaca ili grupa tokom vremena, dok se podaci preseka prikupljaju od različitih grupa ili grupa pojedinaca u isto vreme.U studiji preseka, gde se odgovor meri u jednoj prilici, može se dobiti samo procena razlika između pojedinačnih odgovora, tj. studija preseka može dozvoliti poređenja među podpopula-

cijama koje se razlikuju po godinama, ali ne pruža nikakve informacije o tome kako se pojedinci menjaju tokom odgovarajuceg perioda.

Jedna od glavnih razlika između longitudinalnih i podataka preseka je nivo analize. Longitudinalni podaci se obično analiziraju na individualnom nivou, dok se podaci preseka analiziraju na nivou grupe. To znači da longitudinalni podaci mogu pružiti detaljnije informacije o promenama i trendovima kod pojedinaca, dok podaci preseka mogu pružiti opštije informacije o promenama i trendovima u grupama.

Kao što smo već naveli, longitudinalni podaci se obično koriste u oblastima kao što su psihologija, sociologija, epidemiologija i ekonomija, dok se podaci preseka obično koriste u oblastima kao što su javno zdravstvo, obrazovanje i političke nauke.

3 Linearni modeli na neprekidnim longitudinalnim podacima

Posmatraćemo linearne modele za longitudinalne podatke sa varijablama odgovora koje su kontinuirane i imaju približno simetrične distribucije, bez preterano dugih repova (ili iskrivljenosti) ili odstupanja. Ovi modeli pružaju osnovu za opštije modele za longitudinalne podatke kada je varijabla odgovora diskretna ili brojčana. Uvodimo neke vektorske i matrične zapise i predstavljamo opšti model linearne regresije za longitudinalne podatke. Specifičnost ovog modela je da je srednji odgovor linearan u parametrima regresije.

Valja naglasiti da se statističke metode koje koristimo služe pretpostavkom da logitudinalni odgovori imaju multivarijacionu normalnu raspodelu, ali to ne zahtevaju.

3.1 Matematička notacija

Prilikom definisanja osobina longitudinalnih podataka pretpostavili smo da se uzorak od N ispitanika meri više puta tokom vremena tako da Y_{ij} predstavlja promenljivu odgovora za i-tu osobu prilikom j-tog merenja, te da ispitanici ne moraju imati isti broj ponovljenih merenja, pa se ne mogu meriti u istom skupu trenutaka.

Kako bi obe karakteristike bile ispunjene, pretpostavimo da postoji n_i ponovljenih merenja odgovora i-tog ispitanika i da je svaki odgovor Y_{ij} zabeležen u trenutku t_{ij}. Intuitivno, $n_i \leq n$, jer n_i predstavlja broj merenja na kojima je i-ti ispitanik učestvovao od ukupnog broja merenja n. Gorepomenuta notacija uvodi se zbog problema nedostajućih podataka koji je česta pojava u longitudinalnim studijama i koji za posledicu ima različit broj ponovljenih merenja, te utiče na validnost metoda analize podataka.

Takođe, može doći do pogrešnih merenja, tj. do situacije gde merenja nisu prikupljena u planiranim trenucima n, već nešto kasnije ili nešto pre predviđenih trenutaka merenja. Zato zaključujemo da broj i vreme ponovljenih merenja ne moraju biti zajednički za sve subjekte.

Pogodno je, kao u prethodnom delu, grupisati n_i ponovljenih mera varijable odgovora za i-tog ispitanika u $n_i \times 1$ vektor:

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ ... \\ Y_{in_i} \end{pmatrix}, i = 1, .., N \tag{9}$$

Zaključujemo da je vektor odgovora Y_i vremenski uređena kolekcija n_i varijabli odgovora i-tog ispitanika. Pretpostavlja se da su vektori odgovora Y_i za N ispitanika međusobno nezavisni, dok se, kao što znamo, za rezultate ponovljenih merenja Y_{ij} i-tog ispitanika nezavisnost ne pretpostavlja.

U vezi sa svakim odgovorom Y_{ij} postoji vektor prediktora X_{ij} dimenzije $p \times 1$, gde p predstavlja broj prediktora, i definisan je na sledeći način:

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ ... \\ X_{ijp} \end{pmatrix}, i = 1, .., N, j = 1, ..., n_i \tag{10}$$

Za svaku od n_i ponovljenih mera na i-tom ispitaniku postoji odgovarajući vektor prediktora, tj. elementi vektora X_{i1} dimenzija $p \times 1$ su kovarijatne vrednosti koje odgovaraju varijabli odgovora i-tog ispitanika prilikom prvog merenja, itd.

Vektor X_{ij} mogu činiti dva tipa prediktora:

- Prediktori čije se vrednosti ne menjaju tokom trajanja studije

  Primeri ovog tipa prediktora su pol, fiksni eksperimentalni tretmani. U ovom slučaju iste vrednosti prediktora se repliciraju u odgovarajuće redove vektora X_{ij} za $j = 1, ..., n_i$.

- Prediktori čije se vrednosti menjaju tokom vremena

  Primeri ovog tipa prediktora su vreme od početne vrednosti, trenutni status pušenja, izloženost životnoj sredini. U ovom slučaju vrednosti koje uzimaju prediktori mogu da variraju tokom vremena (za najmanje neke pojedince) i vrednosti u odgovarajućim redovima X_{ij} mogu biti različite u svakoj prilici merenja.

Ukoliko grupišemo vektore prediktora X_{ij} u matricu prediktora dimenzija $n_i \times p$, tada će matrica X_i biti uređena kolekcija vrednosti p prediktora i-tog ispitanika u svakom od n_i merenja. Tada redovi matrice X_i odgovaraju prediktorima povezanim sa odgovorima u n_i različitim merenjima, a kolone matrice X_i odgovaraju p različitih prediktora.

$$X_i = \begin{pmatrix} X_{i1}^T \\ X_{i2}^T \\ ... \\ X_{in_i}^T \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & .... & X_{i1p} \\ X_{i21} & X_{i22} & .... & X_{i2p} \\ ... & ... & ... & ... \\ X_{in_i1} & X_{in_i2} & .... & X_{in_ip} \end{pmatrix} \tag{11}$$

Dakle, do sada smo pretpostavili da svaki ispitanik u okviru longitudinalne studije ima svoj vektor ponovljenih odgovora Y_i koji je povezan sa svakom ponovljenom merom i vektor p prediktora koji mogu biti predstavljeni matricom X_i.

3.1.1 Linearna regresija i promena srednjeg odgovora

Model linearne regresije za promene srednjeg odgovora Y_{ij} tokom vremena i povezivanje tih promena sa prediktorima definisan je jednačinom:

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + ... + \beta_p X_{ijp} + e_{ij}, j = 1, .., n_i \tag{12}$$

Gde su $\beta_1, \beta_2, ..., \beta_p$ nepoznati regresioni koeficijenti koji povezuju srednji odgovor Y_{ij} sa odgova- rajućim prediktorima.

Ovaj regresioni model opsiuje nam način na koji su odgovori u svakom trenutku povezani sa prediktorima, tj. postoji n_i različitih regresionih jednačina za varijablu odgovora Y_{ij} u svakom od n_i različitih merenja.

$$\begin{array}{rcrcl} Y_{i1} & = & \beta_1 X_{i11} + \beta_2 X_{i12} + ... + \beta_p X_{i1p} + e_{i1} & = & X_{i1}^T \beta + e_{i1} \\ Y_{i2} & = & \beta_1 X_{i21} + \beta_2 X_{i22} + ... + \beta_p X_{i2p} + e_{i2} & = & X_{i2}^T \beta + e_{i2} \\ ... & & & & ... \\ Y_{in_i} & = & \beta_1 X_{in_i1} + \beta_2 X_{in_i2} + ... + \beta_p X_{in_ip} + e_{in_i} & = & X_{in_i}^T \beta + e_{in_i} \end{array}$$

Nepoznati regresioni koeficijenti $\beta_1, \beta_2, ..., \beta_p$ su predstavljeni vektorom $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$. Slučajne greške e_{ij} sa očekivanjem 0 predstavljaju odstupanja odgovora od njihovih odgovarajućih očekivanih srednjih vrednosti koje su date jednačinom:

$$E(Y_{ij}|X_{ij}) = \beta_1 X_{i11} + \beta_2 X_{i12} + ... + \beta_p X_{i1p} \tag{13}$$

U najvećem broju slučajeva je vrednost $X_{ij1} = 1$ za svako i i j, pa β_1 intercept (konstanta), tj. koeficijent $beta_1$ predstavlja očekivanu srednju vrednost Y_{ij} kada je svako $X_{ijp} = 0$.

Regresioni model Y_{ij} može se kraće predstaviti pomoću matrične jednačine:

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ ... \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & .... & X_{i1p} \\ X_{i21} & X_{i22} & .... & X_{i2p} \\ ... & ... & ... & ... \\ X_{in_{i1}} & X_{in_{i2}} & .... & X_{in_ip} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ ... \\ e_{in_i} \end{pmatrix} = X_i\beta + e_i, \tag{14}$$

Dakle, dodatna pretpostavka koju smo uveli u okviru ove notacije jeste da je srednja vrednost vektora odgovora Y_i povezana sa prediktorima preko modela linearne regresije.

3.2 Pretpostavke o raspodeli modela

Jedina pretpostavka koja je do sada navedena odnosi se na obrasce promene srednjeg odgovora tokom vremena i odnose tih promena sa kovarijatima. Kako se pretpostavlja da vektor slučajnih grešaka e_i ima srednju vrednost 0, regresioni model $Y_i = \beta X_i + e_i$ implicira da je veza vektora srednjih odgovora i kovarijata data sledećom jednačinom:

$$E(Y_i|X_i) = \mu_i = X_i\beta \tag{15}$$

Gde je $\mu_i = (\mu_{i1}, ...\mu_{in_i)^T}$. Prethodna jednačina je vektor dimenzija $n_i \times 1$ i predstavlja uslovne srednje vrednosti za i-tog ispitanika, gde je $\mu_{ij} = E(Y_{ij}|X_i) = E(Y_{ij}|X_{ij})$.

3.2.1 Pretpostavke o raspodeli vektora slučajnih grešaka e_i

U okviru regresionog modela $Y_i = \beta X_i + e_i$, za vektor odgovora Y_i se pretpostavlja da se sastoji od dve komponente - sistemske komponente βX_i, gde srednji odgovor može biti izražen kao ponderisana suma fiksiranih nepoznatih regresionih koeficijenata β i slučajne komponente e_i koja je uzrok slučajne varijabilnosti regresionog modela Y_i.Zato se pretpostavke o obliku raspodele slučajnih grešaka mogu posmatrati kao pretpostavke o obliku uslovne raspodele $Y_i|X_i$, jer se njihove raspodele jedino razlikuju u smislu pomeranja lokacije.

Slučajne greške imaju raspodelu čija je srednja vrednost (očekivanje) centrirana u nuli, dok je srednja vrednost uslovne raspodele $Y_i|X_i$ centrirana u $X_i\beta$.

U nastavku ćemo se često pozivati na raspodele Y_i i e_i, tj. njihovu kovarijacionu matricu.

3.2.2 Pretpostavke o raspodeli vektora neprekidnih odgovora Y_i

Za vektor Y_i neprekidnih odgovora pretpostavlja se da je njegova uslovna raspodela $Y_i|X_i$ multivarijaciona normalna raspodela, sa vektorom srednjeg odgovora $E(Y_i|X_i) = \mu_i = X_i\beta$ i kovarijacionom matricom $\sum_i = Cov(Y_i|X_i)$.

Multivarijaciona normalna raspodela u potpunosti je određena vektorom srednjeg odgovora μ_i i kovarijacionom matricom $\sum_i$. Svaka od komponenata Y_{ij} vektora Y_i imaće normalnu raspodelu sa uslovnom srednjom vrednošću μ_i i uslovnom disperzijom σ^2.

Statističke metode na koje ćemo se kasnije fokusirati koriste pretpostavku da longitudinalni odgovori Y_i imaju približnu multivarijantnu normalnu distribuciju za izvođenje procena i statističkih testova. Ova pretpostavka se ne zahteva u slučaju da su podaci potpuni ili se podaci koje analiziramo mogu posmatrati kao slučajan uzorak kompletnih podataka.

U okviru longitudinalnih studija, n_i ponovljenih mera na i-tom ispitaniku predstavljamo vektorom odgovora i razmatramo zajedničku raspodelu verovatnoće tih vektora. Multivarijantna normalna raspodela je prirodno proširenje univarijantne normalne raspodele za jedan odgovor na vektor odgovora. Funkcija gustine verovatnoće zajedničke multivarijacione normalne raspodele za $Y_i|X_i$ predstavljena je sledećom funkcijom:

$$f(y_i) = f(y_{i1}, y_{i2}, ...y_{in_i}) = \frac{1}{\sqrt{(2\pi)^{n_i}|\Sigma_i|}} e^{-\frac{(y_i-\mu_i)^T \Sigma_i^{-1}(y_i-\mu_i)}{2}}, y_{ij} \in (-\infty, +\infty) \tag{16}$$

Gde je, kao što je poznato, $j = 1, ..., n_i$, $\mu_i = E(Y_i|X_i) = (\mu_{i1}, ..., \mu_{in_i})$, $\Sigma_i = Cov(Y_i|X_i)$, a $|\Sigma_i|$ je determinanta od Σ_i.

Početna pretpostavka da su X_i nezavisni,a da X_{ij} nisu se može videti u $_i$. Jasno je da je ovo kovarijaciona matrica X_i pa to možemo primetiti. Ako za sve ispitanike važi da su svaki put u jednakim vremenskim intervalima ponavljali istraživanje, kovarijacionu matricu možemo jednostavno označiti sa $\sum$. Kao što je rečeno na početku, to često neće biti slučaj tako da će kovarijaciona matrica zavisti i od vremenskog intervala u kom je ispitanik opet uradio istraživanje kao i broja ponavljanja tog istraživanja.

Pretpostavku o multivarijacionoj normalnoj raspodeli je teško potvrditi. Najkorisnija procena validnosti pretpostavke multivarijantne normalnosti korišćenjem grafičkih prikaza poput histograma i boxplotova reziduala za otkrivanje grubih odstupanja e_{ij} od univarijantne normalnosti. Ovi grafički prikazi se takođe mogu koristiti za određivanje odgovarajuće transformacije varijable odgovora tako da marginalne distribucije e_{ij} bliže normalne distribucije. Međutim, iako multivarijantna normalna raspodela za e_i ukazuje na to da svaki od e_{ij} ima univarijantnu normalnu raspodelu, obrnuto ne važi, tj. pretpostavka multivarijantne normalnosti se ne može formalno potvrditi ispitivanjem svake od komponentnih varijabli e_{ij} posebno. Sa druge strane, gruba odstupanja e_{ij} od univarijantne normalnosti ukazuju da raspodela e_i nije multivarijantna normalna raspodela.

Svojstvo multivarijacione normalne raspodele $Y|X_i$ je da je veza između bilo kog para odgovora linearna. Samim tim, ukoliko je uslovna raspodela Y multivarijaciona normalna, dijagrami rasejanja reziduala u svim mogućim parovima ne bi trebalo da daju dokaze o uočljivim odstupanjima od linearnog trenda među parovima varijabli.

Pretpostavka univarijantne normalnosti e_{ij} nije toliko kritična kao pretpostavke o nezavisnosti grešaka i homogenosti varijanse grešaka. U longitudinalnim podešavanjima podataka postoje veoma slični rezultati, što sugeriše da su pretpostavke o zavisnosti između grešaka i pretpostavke o varija- nsama i kovarijansama koje imaju najveći uticaj na statističko zaključivanje.

Kao što već znamo, u longitudinalnim studijama ponovljena merenja na istoj osobi su inherentno zavisna ili korelisana. Kada se bavimo longitudinalnim podacima koji su neprekidni, pretpostavljamo da je njihova zajednička raspodela multivarijantna normalna u cilju izvođe- nja ocena i statističkih testova, ali u praksi se ne očekuje da longitudinalni podaci imaju zajedničku distribuciju koja je tačno multivarijantna normalna. Ona je usvojena kao aproksi- macija koja ima veliki broj pogodnih statističkih svojstava.

3.3 Jednostavne opisne metode analize longitudinalnih podataka

Pošto smo se dobro upoznali sa novim tipom podataka, trebalo bi još opisati neke od načina kako analiziramo i popravljamo te podatke.

Opisne metode analize longitudinalnih podataka koriste se kako bi se imao uvid u strukturu podataka i tok promena kroz vreme. Neki od najčešće korišćenih opisnih metoda su:

- Grafici promena

  Grafici promena (eng. "change plots") su korisni za vizualizaciju promena u vrednostima varijable tokom vremena. Ovi grafici prikazuju trendove i fluktuacije u vrednostima varijable i mogu pomoći u identifikaciji potencijalnih autlajera. Primeri grafika promena uključuju linije vremenskih nizova (eng. "time series plots"), boxplotove, violinske plotove, te grafike gustine (eng. "density plots").

- Statistički pokazatelji

  Statistički pokazatelji (eng. "summary statistics") su kvantitativni opisi distribucije vrednosti varijable u različitim vremenskim tačkama. Ovi pokazatelji uključuju srednju vrednost, standardnu devijaciju, kvartile, medijanu i druge mere centralne tendencije i disperzije. Korisno je napraviti statistički opis za svaku vremensku tačku kako bi se razumela distribucija vrednosti i kako bi se utvrdili trendovi tokom vremena.

- Tablice frekvencija

  Tablice frekvencija se koriste za opisivanje raspodele kategoričkih varijabli tokom vremena. Ove tablice prikazuju broj i postotak ispitanika koji pripadaju svakoj kategoriji varijable za svaku vremensku tačku. Korisne za identifikaciju trendova i promena u raspodeli kategorija varijable.

- Korelaciona matrica

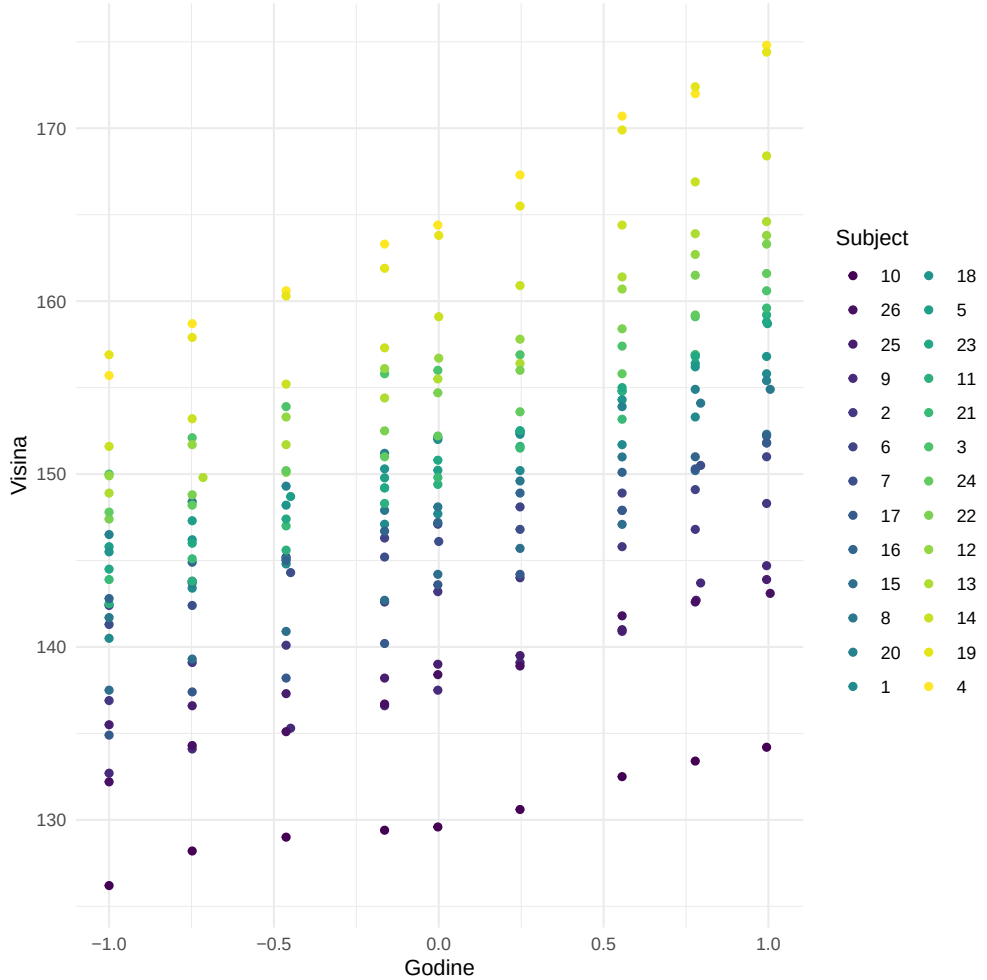
  Korelacione matrice se koriste kako bi se dobili uvidi u povezanost između varijabli tokom vremena. Ove matrice prikazuju korelacije između svake kombinacije varijabli za svaku vremensku tačku. Korisne su za identifikaciju povezanosti između varijabli i za utvrđivanje kojim se varijablama treba posvetiti posebna pažnja prilikom dalje analize.

U svakom statističkom istraživanju treba izvršiti pregled podataka i upoznati se sa podacima. Najednostavni način za to su grafičke reprezentacije podataka. Pošto su logitudialni podaci veoma specifična vrsta podatka, postoje razni načini za njihovu vizualizaciju i za popravljanje podataka u cilju što boljeg predstavljanja.

3.3.1 Grafička reprezentacija podataka

Razmotrimo neke jednostavne grafičke alate za vizuelizaciju longitudinalnih podataka. Priro- dan način za prikaz longitudinalnih podataka je *time plot.*To je time *scatterplot* sa odgovorima na vertikalnoj osi i vremenima kada su odgovori zabeleženi na horizontalnoj osi. Kao rezultat *time plot* nam daje mnogo podataka koji se preklapaju. Dodatni problemi nastaju ukoliko ukoliko su naši podaci binarni. U tom slučaju nećemo moći da uočimo nikakve trendove rasta zbog preklapanja podataka. Takođe *time plot* nam ne pokazuje koje tačke predstavljaju ponovljeno merenje za kog ispitanika. Kako bismo što bolje razumeli sve vrste grafičke analize, koristićemo bazu *Oxboy* koja se nalazi u okviru pakata *nlme*. U njoj se nalaze podaci o 26 dečaka kojima je praćena visina tokom 8 godina.

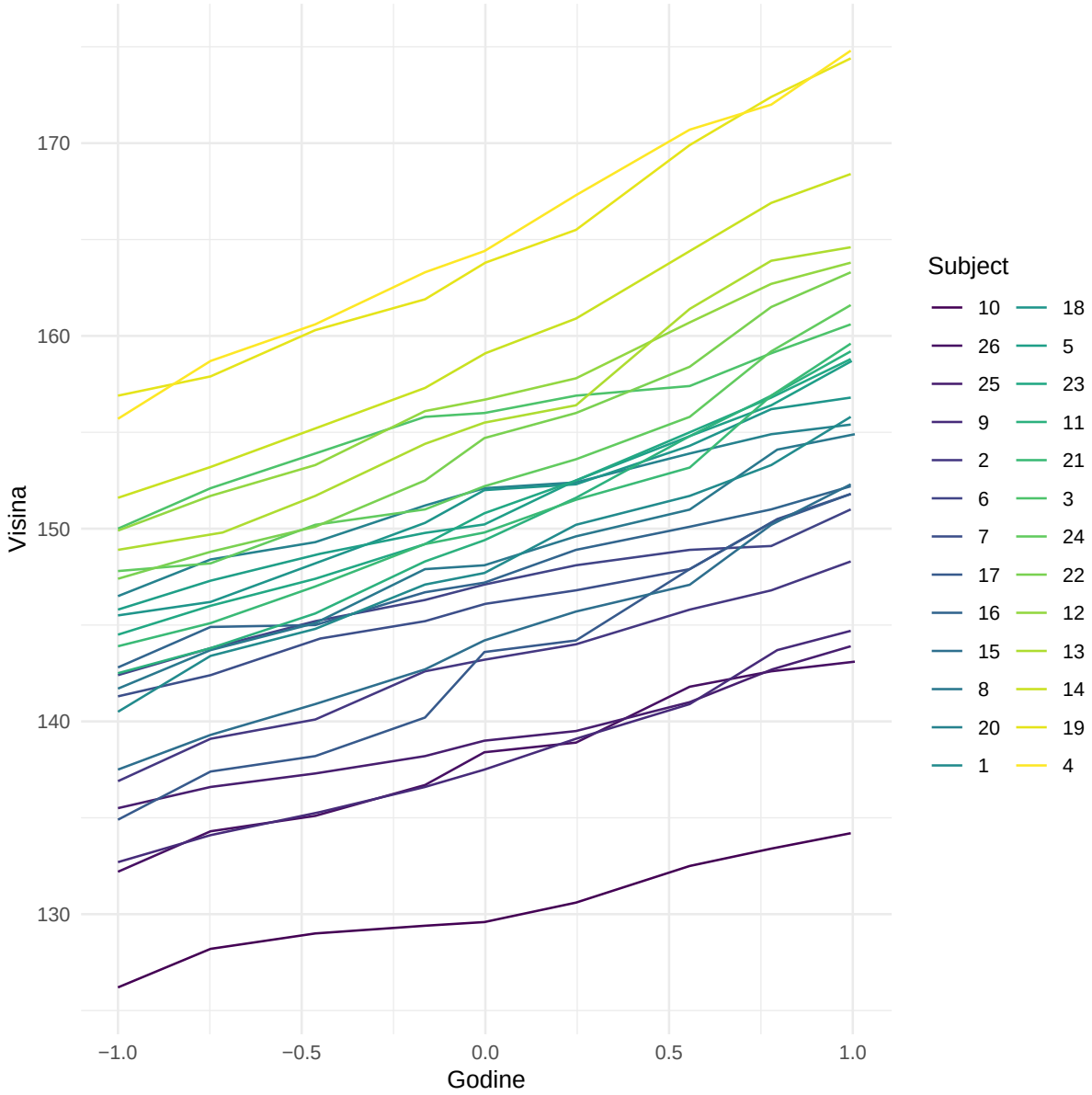
```
##
## Attaching package: 'nlme'
## The following object is masked from 'package:lme4':
##
##     lmList
```



Primećujemo da je grafik težak za čitanje, te da potencijano možemo pretpostaviti

autlajere, ali ne možemo pratiti trend promena visine. Na ovom grafiku godine standardizovane i uzimaju vrednosti između $[-1, 1]$.

Iako je *time plot* možda najintuitivniji alat za rad, vidimo da nam ne pomaže previše. Takođe, grafik koji možemo koristiti za analizu podataka je *"spaghetti"plot*.

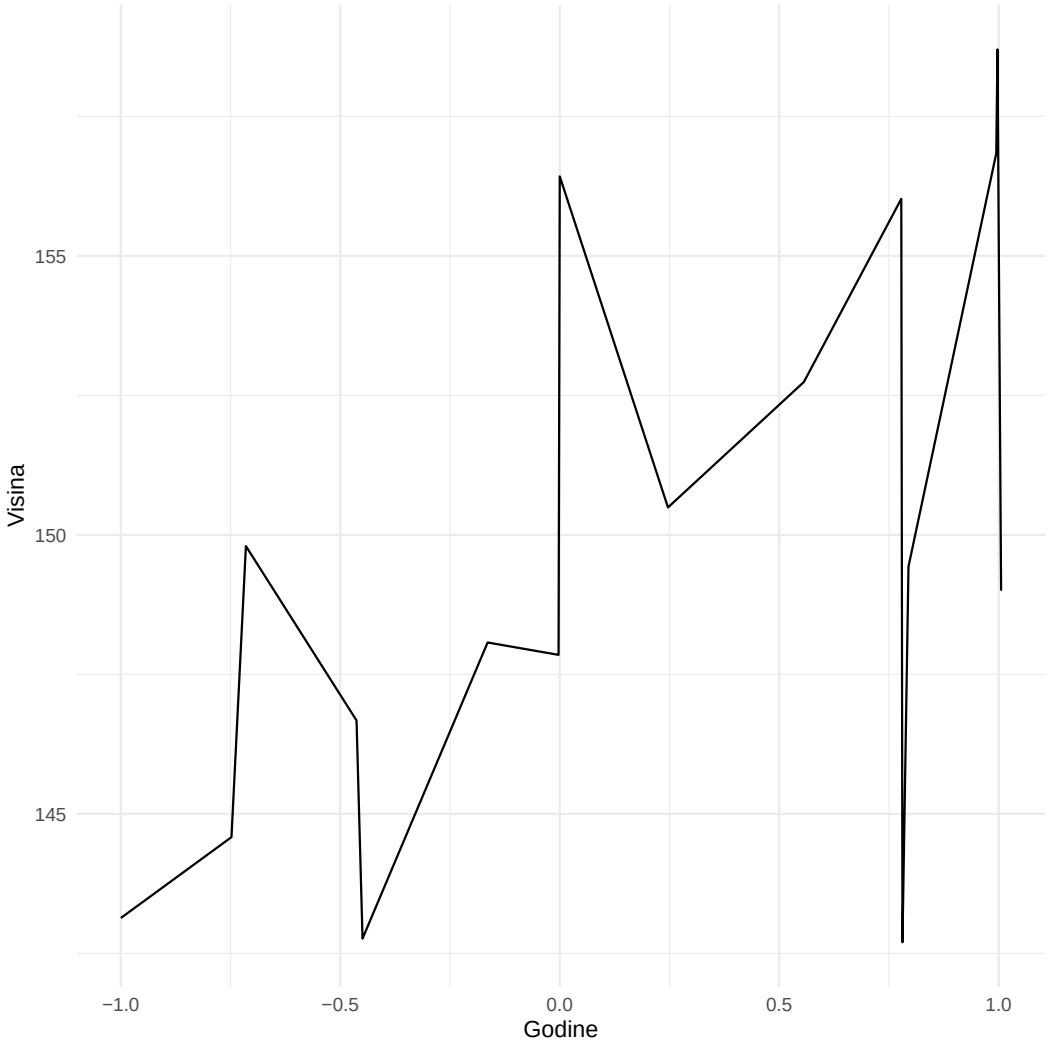


"Spaghetti"plot nam omogućava da pratimo promene za odgovarajuće subjekete tokom istraživanja. Ovaj vid prikaza ima svoje mane, ali nas može jasnije informisati o mogućim autlajerima.

Možemo zaključiti da *time plot* longitudinalnih podataka nije najbolji način za prikaz naših podataka. Ovo je posebno uočljivo ako se podaci tokom vremena uravnoteže i tada postaje veoma teško uočiti razlike, nove trendove ili autlajere u istraživanju.

Kako bismo ovo izbegli bolje je da podatke predstavljamo objedinjeno, kao *time plot*

proseka ili *time plot očekivane vrednosti*. Tada dobijemoo grafik koji ima samo jednu liniju i dosta je lakši za analiziranje.



Ovakav način prikazivanja je puno informativniji za celokupnu studiju. Grafički prikaz podataka nam može pomoći da lakše izaberemo tip modela koji je najadekvatnije koristiti prilikom obrade podataka.

Time plot srednjih vrednosti je malo teže napraviti ako radimo sa podacima koji su merljivi (npr.visina), te se preporučuje pravljenje kategorija u cilju boljeg sagledavanja podataka. Kategorije mogu biti "niži od vršnjaka", "iste visine kao vršnjaci", "viši od vršnjaka". Ovo će nam omogućiti da srednju vrednost gledamo po kategorijama, što je za prikaz i rad sa podacima dosa lakše.

Treba još samo odgovoriti na pitanje, kako najbolje definisati ovakve kategorije? Bilo bi idealno da budu odabrane dve ili tri kategorije tako da istaživač može jasno da uoči suštinski značaj prelaska sa jednog nivoa na drugi. Ukoliko bi naše podatke grupisali u tri grupe mogli bismo uzeti da u prvu grupu spada prvi kvantil, u drugu drugi i treći kvantil, a četvrti kvantil spada u treću grupu. Na taj način bismo jasno razdvojili naše podatke. Naravno ovo nije jedini način na koji se podaci mogu grupisati. Ponekad su prelasci iz jedne u drugu grupu veoma očigledni i nije ih potrebno teorijski razmatrati kao u ovom slučaju.

3.3.2 Smooting techniques

Do sada smo pretpostavljali da su podaci koje smo koristili uzimani u istom vremenskom intervalu, ali kada to nije slučaj konstruisanje dijagrama postaje teško jer nam nedostaju podaci zabeleženi u tačnim trenucima.

U ovom slučaju, pomogla bi konstrukcija lepšeg grafika srednjih odgovora tokom vremena i to se postiže korišćenjem tehnika koje se jednim imenom zovu *"smooting techniques"*. Veliki broj ovih tehnika razmatra odgovore u određenom trenutku istraživanja uzimajući u obzir i odgovore susednih trenutaka ispitivanja, te procenjenu srednju vrednost u traženom trenutku zasnivaju na odgovorima koji su se desili pre i posle traženog trenutka.

Jedna od najpoznatijih tehnika koja se koristi za rešavanje ovog problema je tzv. *"pomerajući prosek"*. Za izbalansirane i potpune longitudinalne podatke pomerajući prosek u trenutku t označava se sa S_t.

$$S_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=-k}^{k}\omega_j y_{i,t+j}, t = k+1, ..., n-k, k \in \mathbb{N} \tag{17}$$

Broj $2k+1$ predstavlja red pomerajućeg proseka. On određuje simetričnost bliskih vrednosti koje su korišćene u proceni srednjeg odgovora u trenutku t. Što je red pomerajućeg proseka veći, veća je glatkoća procenjenih vrednosti i obrnuto. Analogno tome, veći red pomerajućeg proseka insinuira na bolji grafik, dok manji red predviđa hrapav, tj. nazubljen grafik.

Važno je napomenuti da se kod veoma neizbalansiranih ili/i nepotpunih podataka može definisati slična jednačina.

Definišimo sada ω_j kao *skup težina* čije je jedino ograničenje da $\sum_{j=-k}^{k}\omega_j = 1$. Poznato je da su ω_j pozitivne vrednosti, a kada to nije slučaj, ω_j se biraju tako da se simetrično smenjuju oko maksimalne vrednosti ω_j, tj. $\omega_j = \omega_{-j}$ i $\omega_0 > \omega_1 .... > \omega_k$, tako da opservacije koje su zabeležene u najbližoj okolini vremena t imaju najveći uticaj na težine.

Ovakva procena težina postaje problematična kada se nalazimo na početku ili kraju istraživanja jer tada nema dovoljno podataka kako bi se simetričnost jasno izrazila. Do prevazilaženja ovog problema dolazi sumiranjem težina od $j = max(-k, 1-t)$ do $j = min(k, n-t)$ i deljenjem dobijenog broja brojem težina koje učestvuju u zbiru.

Pomerajući prosek je je najprilagođeniji za rad sa podacima koji su glatki i koji se događaju u sličnim vremenskim intervalima, tj. ovaj izbor metodologije je adekvatan u slučaju da podatke koje imamo treba malo popraviti. U situaciji kada su podaci potpuno neregularni, ovaj metod nije primenljiv.

U slučaju kada su podaci neregularni, tj. *raštrkani* u vremenu, za njihovo popravljanje koristi se neparametarski regresioni pristup koji procenjuje srednju vrednost tokom vremena.

Ovom prilikom izdvojili bismo metod pod nazivom *lokalna težinska regresija*, tj. *lowess* koji se nalazi u većini statističkih paketa za obradu podataka.

Neparametarska regresija pokušava da uđe u trag istaknutim karakteristikama srednjeg odgovora kao funkcije od vremena i na taj način pravi minimalne pretpostavke o istom. U cilju razumevanja *lowess* procene, potrebno je da pretpostavimo da postoji *"prozor"* koji je centriran u trenutku t. Ova procena, pomoću *"robusne regresije"* koja daje veću težinu onim tačkama koje se nalaze bliže centru prozora, generiše pravu koja najbolje odgovara podacima gorenavedenog prozora, te se lakše detektuju autlajeri. Dakle, procena srednje vrednosti u

trenutku t je vrednost koja zavisi od regresione linije koju dobijamo tako što prozor fiksne širine pomeramo od prvog do poslednjeg merenja i ovaj proces ponavljamo za svako merenje.

Sve *smooting tehnike* zahtevaju postojanje *bandwith parametra* koji služi za ocenjivanje glatkoće grafika. U slučaju *lowess* tehnike, širina prozora je parametar koji određuje glatkoću grafika.

Prekomerno zaglađivanje dovodi do smanjenja disperzije, ali povećava mogućnost pristrasnosti, pa je u okviru *smooting tehnika* cilj naći kompromis između ova dva. Neparametarski *smooting modeli* ignorišu korelaciju između ponovljenih merenja kod istog ispitanika. Iako korelacija ne bi preterano uticala na srednju vrednost, potrebno je imati je na umu.

Ova tehnika može biti korisna za otkrivanje trendova u srednjoj vrednosti tokom vremena, uz upozorenje da ne treba zanemariti opseg pouzdanosti koji koriste standardni statistički paketi. U slučaju da su odstupanja velika (veća od 5) ili ispitanici napuštaju studiju, kriva može biti značajno iskrivljena na krajevima.

3.4 Modelovanje srednje vrednosti

Kao što je već rečeno, analiza longitudinalnih podataka zasniva se na promenama srednjeg odgovora tokom vremena i na odnos ovih promena i prediktora. Za modelovanje srednje vrednosti vektora odgovora pri analizi longitudinalnih podataka najčešće se koriste:

- *Analiza profila odgovora*

  Ne pretpostavlja se određeni vremenski trend već se vremena merenja smatraju nivoima diskretnog faktora.Metod dozvoljava proizvoljan obrazac srednjih odgovora tokom vremena. Ovaj pristup analizi longitudinalnih podataka je primenljiv samo kada se svi pojedinci mere u istom skupu prilika i kada je broj prilika obično mali.

- *Parametarske ili polumarametarske krive*

  Pretpostavlja parametarsku krivu (npr. linearni ili kvadratni trend) za srednji odgovor tokom vremena. Ovaj pristup može smanjiti broj parametara modela, jer parametarske krive pružaju što optimalniji opis ponašanja srednjeg odgovora i efekta prediktora na njega tokom vremena. Parametarske krive opisuju srednji odgovor kao eksplicitnu funkciju vremena, pa se ne zahteva da svi ispitanici imaju isti skup vremena ili isti broj ponovoljenih merenja. Parametarske krive nameću eksplicitnu sktrukturu srednjih odgovora.

U regresionim modelima za longitudalne podatke nepoznati parametri regresije β povezuju promene srednjeg odgovora sa prediktorima i smatraju se parametrima od primarnog, tj. suštinskog interesa. Njih možemo definisati tako da sumiraju važne aspekte istraživačkih pitanja, te se zato parametri β nazivaju *suštinski parametri.*

U većini slučajeva se parametri koji sumiraju aspekte kovarijanse ili korelacije između ponovljenih mera, tj. kovarijacioni i korelacioni parametri posmatraju kao sekundarni interes u odnosu na srednjih odgovor Y_i tokom vremena i nazivamo ih *parametri smetnji.* Ipak, potrebno je primetiti kovarijansu i korelaciju u cilju odabira adekvatnog metoda analize.

U okviru longitudinalnih podataka, suštinski parametri biće oni koji daju nepristrasne procene srednjeg odgovora tokom vremena. Iako je najčešće suštinski parametar β, određiva-

nje suštinskih i smetajućih parametra treba sprovesti pre svake obrade podataka, jer podaci mogu biti povezani (porodične studije bolesti), te su u tim situacijama korelacije između odgovora suštinski parametri.

3.5 Modelovanje kovarijanse

Karakteristika longitudinalnih podataka je da se odgovori ponavljaju na istim osobama tokom vremena, a rezultirajući odgovori su u korelaciji. Iako korelacija nije uvek od suštinskog interesa, ne može se jednostavno zanemariti jer utiče na tačnost i preciznost procena regresionih parametara. Korelacija među ponovljenim merama važan je deo specifikacije modela regresije za longitudinalne podatke jer povećava preciznost i efikasnost sa kojom se regresioni parametri mogu oceniti, te u slučaju nedostajućih podataka je često uslov za dobijanje validnih procena parametara, pa je potrebno odabrati adekvatan model za kovarijansu. Ukoliko je napravljen adekvatan izbor, dobijaju se validne standardne greške i može se ispravno zaključiti o ponašanju i osobinama regresionih parametara.

Postoji metod koji zanemaruje korelaciju između ponovljenih mera u okviru procene regresionih parametara, ali za potrebe zaključivanja vrši određeno prilagođavanje standardnim greškama.

Razlikujemo tri pristupa modeliranju kovarijanse među ponovljenim merama: nestrukturirana kovarijansa, modeli kovarijansnog uzorka i strukture kovarijanse sa slučajnim efektima.

3.5.1 Nestruktuirana kovarijansa

Ovaj pristup dozvoljava bilo koji proizvoljan obrazac kovarijanse među ponovljenim merama, tj. ne pretpostavlja se bilo kakva eksplicitna struktura za nju osim homogenosti kovarijanse među različitim ispitanicima, te se kovarijansa naziva *nestruktuirana*.

Kada postoji n ponovljenih merenja, postoji n varijansi (disperzija) koje se procenjuju u svakom od n merenja, pa se procenjuje $\frac{n(n-1)}{2}$ parova kovarijansi (ili korelacija).

Nestruktuirana matrica kovarijanse je bila model izbora za kovarijansu u *Analizi profila odgovora*, tj. ovaj metod pretpostavlja proizvoljne obrasce za srednji odgovor tokom vremena (i njihov odnos sa prediktorima), varijanse i kovarijanse. Ovaj pristup može se koristiti i u slučaju modelovanja srednjeg odgovora parametarskim ili semiparametarskim krivama.

Nedostaci nestruktuirane matrice kovarijanse:

1. Broj parametara kovarijanse može biti veliki.

   Ako postoji n slučajnih merenja, onda matrica kovarijanse dimenzija $n \times n$ ima $\frac{n(n-1)}{2}$ jedinstvenih parametara. U slučaju da se desilo $n = 10$ ponovljenih merenja, nestruktuirana matrica kovarijanse imaće 55 parametara, tj. 10 varijansi i 45 kovarijansi. Kada je broj parametara kovarijanse za procenu veliki u odnosu na veličinu uzorka, procene će biti nestabilne.

2. Primenljiv samo kada se svi pojedinci mere u istom skupu trenutaka.

   Pristup ne može prihvatiti pogrešna merenja ili merenja sa nepravilnim vremenskim rasporedom.

3.5.2 Modeli kovarijansnog uzorka - *CPM*

U ovom pristupu pretpostavlja se struktura za matricu kovarijanse i on se služi idejom iz analize vremenskih serija – gde su podaci, za razliku od longitudinalnih, sa malim brojem replikacija (često $N = 1$) ili pojedinaca i velikim brojem ponovljenih mera (n), te za njih možemo reći da se sastoje od malog broja dugih nizova ponovljenih merenja. U longitudalnoj studiji, broj replikacija ili pojedinaca (N) je veliki u odnosu na broj ponovljenih mera (n), pa je situacija obrnuta, tj. longitudalni podaci se sastoje od velikog broja relativno kratkih nizova ponovljenih merenja. Zajednička osobina ova dva tipa podataka jeste da su ponovljene mere u korelaciji. Kao što je već poznato, očekuje se da će ponovljene mere koje su bliže u vremenu biti u većoj korelaciji od onih koje su udaljenije u vremenu, tj. korelacije opadaju kako se vremensko razdvajanje povećava. Ova korelacija može se izraziti kao eksplicitna funkcija vremenskog odvajanja i modeli se mogu koristiti sa nejednako raspoređenim posmatranjima. Mnogi modeli za varijansu pretpostavljaju stacionarnost, odnosno da se varijansa ne menja kao funkcija vremena.

Pristup se zasniva na pretpostavci da postoje latentne promenljive koje uzrokuju ponovljena merenja i koje su međusobno korelisane. Cilj modela je procena parametara koji opisuju ove latentne promenljive i njihovu korelaciju, što omogućava da se dobiju tačniji proceni regresionih parametara.

U CPM, kovarijanse između ponovljenih mera se modeliraju putem kovarijansne matrice, koja prikazuje međusobne kovarijanse između različitih merenja. Postoje različite vrste CPM-a, kao što su unutarosobni (unutar-subject) modeli, međusobni (between-subject) modeli i hibridni modeli koji kombinuju ove dve vrste.

Unutarosobni modeli se koriste kada se interesuje za promene tokom vremena na nivou pojedinca, dok se međusobni modeli koriste kada se interesuje za razlike između grupa ispitanika. Hibridni modeli su korisni kada se interesuje za razlike između grupa u promenama tokom vremena.

3.5.3 Strukture kovarijanse sa slučajnim efektima

U ovom pristupu pretpostavlja se struktura za matricu kovarijanse uvođenjem slučajnih efekata. Prilikom analize podataka o ponovljenim merenjima, modeli slučajnih efekata bili su jedan od najranijih pristupa.

Naime, u *ANOVA* modelu univarijantnih ponovljenih merenja, korelacija između ponovljenih merenja se objašnjava uključivanjem jednog specifičnog slučajnog efekata koji se sastoji od svih neopaženih ili neizmerenih faktora koji utiču na osobine ispitanika. Kao posledica dodavanja ovog efekta svakom merenju bilo kog ispitanika nameće se informacija da će rezultirajuća ponovljena merenja biti u pozitivnoj korelaciji, te možemo zaključiti da uključivanje slučajnih efeketa nameće strukturu kovarijansi.

Pored gorenavedenog, karakteristike ovog modela jesu da je rezultujuća pozitivna korelacija konstantna i ne varira kao funkcija vremena između bilo kog para ponovljenih merenja, te da je varijansa konstantna tokom vremena.

Iako ove karakteristike nisu adekvatne za longitudinalne podatke, prilagođavamo model uvođenjem više od jednog slučajnog efekta, tj. pretpostavljamo da podskup parametara regresije varira nasumično među pojedincima. Uvođenje novih slučajnih efekata može dovesti

do manje restriktivnih obrazaca korelacije, te do glatkih promena varijanse tokom vremena, pa se često koriste za obradu vremenski neregularnih longitudinalnih podataka.

Izbor strukture kovarijanse slučajnih efekata zavisi od specifičnog istraživačkog pitanja i prirode podataka koji se analiziraju.

3.6 Ocenjivanje regresionih koeficijenata i matrice kovarijanse greške

Procena parametara u modelima longitudinalne regresije podataka uključuje procenu fiksnih efekata, slučajnih efekata i matrice kovarijanse greške.

Fiksni efekti predstavljaju srednji odgovor svih pojedinaca, dok nasumični efekti obuhvataju individualne specifične varijacije u odgovoru tokom vremena.

3.6.1 Matrica kovarijanse greške

Matrica kovarijanse greške (ECM) u modelu longitudinalne regresije podataka predstavlja obrazac korelacije između reziduala ili grešaka u modelu. Ostaci ili greške predstavljaju varijaciju koja nije objašnjena fiksnim ili slučajnim efektima u modelu.

Matrica kovarijanse greške je kvadratna matrica sa istim brojem redova i kolona kao i broj posmatranja u skupu podataka. Dijagonalni elementi matrice kovarijanse predstavljaju varijansu svakog ostatka ili greške, dok vandijagonalni elementi predstavljaju kovarijansu između svakog para reziduala ili grešaka.

Matrica kovarijanse greške je važna iz nekoliko razloga:

- Koristi se za procenu standardnih grešaka procena parametara u modelu.

  Standardne greške se koriste za konstruisanje intervala poverenja i testova hipoteza za procene parametara, koji se koriste za procenu statističke značajnosti efekata prediktorskih varijabli.

- Koristi se za procenu kvaliteta fittovanja modela.

  Model koji se dobro uklapa treba da ima matricu kovarijanse koja tačno obuhvata obrazac korelacija između reziduala ili grešaka u podacima.

- Koristi se za predviđanje novih zapažanja.

  Matrica kovarijanse se koristi za izračunavanje intervala predviđanja, koji predstavlja opseg vrednosti u koji ce novo posmatranje verovatno pasti, s obzirom na nesigurnost u modelu.

Matrica kovarijanse greške u modelu longitudinalne regresije podataka je važna komponenta modela, za njeno ocenjivanje koriste se iste metode kao za ocenjivanje regresionih parametara, a njeno tumačenje i procena su od ključne važnosti za precizno statističko zaključivanje i predviđanje.

3.6.2 Metod maksimalne verodostojnosti - *MLE*

Jedna uobičajena metoda za procenu parametara u modelima longitudinalne regresije podataka je procena maksimalne verodostojnosti (*MLE*). U *MLE*, funkcija verovatnoće je maksimizirana u odnosu na nepoznate parametre modela. Ovo uključuje pronalaženje vrednosti parametara koje maksimiziraju verovatnocu posmatranja podataka datih modelu.

Proces pronalaženja procene maksimalne verodostojnosti uključuje uzimanje parcijalnih izvoda funkcije verovatnoće u odnosu na svaki parametar i njihovo postavljanje jednakim nuli. Ovo daje sistem jednačina koji se može rešiti da bi se dobile procene maksimalne verodostojnosti parametara.

Ovaj metod ima nekoliko poželjnih osobina poput konzistentnosti (kako se veličina uzorka povećava, ocene dobijene ovim metodom konvergiraju ka pravim vrednostima parametara), efikasnosti (ocene imaju najmanju disperziju od svih ostalih nepristrasnih ocena) i asimptotske normalnosti.

Metod maksimalne verodostojnosti se koristi kada je cilj da se procene parametri modela i komponente varijanse povezane sa slučajnim efektima. *MLE* obezbeđuje procene fiksnih efekata i komponenti varijanse koje maksimiziraju verovatnoću podataka, i pogodan je za modele sa izbalansiranim ili neuravnoteženim podacima. U slučaju da je cilj da se procene komponente varijanse povezane sa slučajnim efektima ili da se uporede različiti modeli, onda *MLE* može biti prikladan metod jer obezbeđuje procene i fiksnih efekata i komponenti varijanse.

MLE se široko koristi u statističkom zaključivanju i standardni je metod za procenu parametara u mnogim tipovima statističkih modela. Međutim, *MLE* zahteva da se funkcija verovatnoce može specificirati i maksimizirati, što možda nije uvek moguće.

3.6.3 Metod ograničene maksimalne verodostojnosti - *REML*

Ograničena maksimalna verodostojnost (*REML*) je metoda za procenu parametara linearnog modela mešovitih efekata. Metoda je razvijena kao alternativa proceni maksimalne verodostojnosti (*MLE*) za linearne mešovite modele i bavi se nekim ograničenjima *MLE*, posebno u slučaju neuravnoteženih ili nekompletnih podataka. Metod objašnjava činjenicu da se slučajni efekti procenjuju na osnovu podataka, a ne da su poznati. Ovo može rezultirati preciznijim procenama fiksnih efekata.

Osnovna ideja iza *REML*-a je procena fiksnih efekata modela uklanjanjem slučajnih efekata iz funkcije verovatnoće. Drugim rečima, komponente varijanse povezane sa slučajnim efektima se procenjuju odvojeno, a zatim se izračunava funkcija verovatnoće za ostatke nakon uklanjanja slučajnih efekata.

Funkcija verovatnoće u *REML*-u je definisana kao zajednička funkcija gustine verovatnoće posmatranih podataka, uslovljena fiksnim efektima i komponentama varijanse povezanim sa slučajnim efektima. Funkcija verovatnoće se može izraziti kao proizvod uslovnih gustina varijable odgovora, s obzirom na fiksne efekte i slučajne efekte. Funkcija preostale verovatnoće se zatim definiše kao funkcija verovatnoće reziduala, koje su razlike između posmatrane varijable odgovora i predviđenih vrednosti varijable odgovora, na osnovu fiksnih i slučajnih efekata.

Prednost ovog pristupa je što obezbeđuje efikasnije i manje pristrasne procene fiksnih

efekata, posebno kada je broj slučajnih efekata veliki ili kada su podaci neuravnoteženi ili nekompletni. To je zato što *REML* pristup koristi samo deo funkcije verovatnoće koji je relevantan za procenu fiksnih efekata, dok *MLE* pristup uključuje komponente varijanse povezane sa slučajnim efektima, što može dodati šum i pristrasnost procenama fiksnih efekata. Koraci pri upotrebi ovog metoda su sledeći:

1. Procena komponenti varijanse modela

   Prvi korak je procena komponenti varijanse povezanih sa slučajnim efektima modela. Ovo se može uraditi korišcenjem različitih metoda, kao što je rezidualna maksimalna verovatnoca (REML) ili metoda trenutaka.

2. Računanje reziduala

   Drugi korak je izračunavanje reziduala modela oduzimanjem prilagođenih vrednosti slučajnih efekata od posmatranih vrednosti promenljive odgovora.

3. Računanje funkcije verovatnoće reziduala

   Treci korak je izračunavanje funkcije verovatnoće reziduala koristeći standardnu funkciju verovatnoće, kao što je normalna funkcija verovatnoće za neprekidne podatke ili Puasonova funkcija verovatnoće za podatke o brojanju.

4. Procena fiksnih efekata

   *Četvrti korak je pro*cena *fiksnih efekata modela korišćenjem procene maksimalne verodostojnosti parametara dobijenih iz funkcije verovatnoće.*

5. Izračunavanje standardnih grešaka

   Poslednji korak je izračunavanje standardnih grešaka fiksnih efekata korišćenjem inverzne Fišerove informacione matrice.

Valja napomenuti da je *REML* pristup računarski zahtevniji i nije uvek izvodljiv za velike skupove podataka ili složene modele.

3.6.4 Ostali pristupi ocenjivanja parametara

Pored *MLE* i *REML*, Bajesove metode se takođe mogu koristiti za ocenjivanje parametara u modelima longitudinalne regresije podataka. Bajesove metode uključuju specificiranje prethodnih distribucija za nepoznate parametre, a zatim korišćenje Bajesove teoreme za ažuriranje ovih distribucija na osnovu posmatranih podataka.

Bez obzira na metod procene koji se koristi, važno je pažljivo razmotriti izbor strukture kovarijanse za slučajne efekte i greške. Izbor strukture kovarijanse može imati značajan uticaj na procenu parametara i interpretaciju rezultata.

4 Regresioni modeli za analizu longitudinalnih podataka

Regresioni modeli za longitudinalne podatke su statistički modeli koji se koriste za analizu podataka prikupljenih tokom vremena. Ovi modeli se koriste za praćenje promena vrednosti zavisne promenjive na osnovu podataka koje smo prikupili u ponovljenim istraživanjima.

U nastavku ćemo predstaviti neke od linearnih modela koji se koriste u analizi lognitudalnih podataka. Valja napomenuti da svaki od njih ima i generalizovani oblik, tzv. *GLM - Generalizovani linearni model*, koji, za razliku od linearnog modela koji pretpostavlja da promenljiva odgovora ima normalnu raspodelu, dozvoljava da raspodela promeljive odgovora nije normalna i zahteva specifikaciju funkcije veze i distribucije iz eksponencijalne porodice, kao što je logit ili Poissonova distribucija.Koji od ova dva tipa modela ćemo korisiti zavisi od prirode varijable odgovora.

Sada ćemo predstaviti linearne regresione modele, tj. pretpostavićemo da promenljiva odgovora modela koje izučavamo ima normalnu raspodelu.

4.1 Modeli mešovitih efekata

Modeli mešovitih efekata, takođe poznati kao modeli na više nivoa ili hijerarhijski modeli tip modela linearne regresije koji uključuje i fiksne i slučajne efekte koji se zajednički nazivaju *„mešoviti efekti“*.

- *Fiksni efekti* - Efekti prediktora na odgovor koji su fiksni i konstantni kod svih pojedinaca u populaciji. Oni predstavljaju prosečan efekat prediktora na varijablu ishoda. U modelu mešovitih efekata, fiksni efekti se procenjuju na isti način kao u standardnom modelu linearne regresije.
- *Slučajni efekti* - Efekti prediktora na odgovor koji su specifični za svakog pojedinca u populaciji. Oni predstavljaju varijaciju u efektu prediktora na varijablu ishoda među pojedincima. U modelu mešovitih efekata, pretpostavlja se da su slučajni efekti normalno raspoređeni sa srednjom vrednošću nula i varijansom koja se procenjuje na osnovu podataka.

Kod ovih modela vršimo ispitivanje više puta tokom vremena, a merenja za svakog ispitanika su nezavisna. Modeli mešovitih efekata omogućavaju procenu uticaja efekata na pojedinca, i uticaj izmedju pojedinaca, što može pomoći da se detektuju faktori koji utiču na zavisnu promenjivu.

Osnovna ideja modela mešovitih efekata je modeliranje varijable odgovora kao funkcije i fiksnih i slučajnih efekata.U ovom modelu varijabla odgovora je modelirana kao:

$$Y = \beta X + \gamma Z + \epsilon \tag{18}$$

Gde Y predstavlja varijablu odgovora, X matricu fiksnih efekata, β vektor koeficijenata fiksnih efekata, Z matricu slučajnih efekata, γ vektor koeficijenata slučajnih efekata i ϵ vektor grešaka reziduala. Pretpostavlja se da slučajni efekti imaju multivarijacionu normalnu raspodelu sa očekivanjem 0 i kovarijacionom matricom.

Modeli mešovitih efekata imaju nekoliko prednosti u odnosu na druge modele za analizu longitudinalnih podataka. Oni mogu da obrađuju neuravnotežene podatke, gde svi subjekti

nemaju isti broj merenja, i mogu da rukuju podacima koji nedostaju. Oni takođe mogu uzeti u obzir korelaciju između ponovljenih mera i prilagoditi varijabilnosti unutar subjekta. Pored toga, modeli mešovitih efekata mogu proceniti komponente varijanse slučajnih efekata, pružajuci uvid u izvore varijabilnosti u podacima.

Modeli mešovitih efekata su tipično linearni modeli, jer su zasnovani na linearnim kombinacijama fiksnih i slučajnih efekata za modelovanje srednje vrednosti promenljive odgovora. Fiksni efekti predstavljaju prosečan odnos između varijable odgovora i prediktora, dok slučajni efekti obuhvataju varijaciju odnosa na nivou pojedinca ili klastera. Linearna kombinacija fiksnih i nasumičnih efekata pretpostavlja linearnu vezu između varijable odgovora i prediktora, iako model može uključiti nelinearne termine, kao što su kvadratni ili kubni, da bi se uhvatili složeniji odnosi. Pored toga, modeli sa mešovitim efektima se mogu koristiti za modelovanje nelinearnih odnosa između promenljive odgovora i prediktora transformacijom prediktora, kao što je korišcenje logaritamskih ili eksponencijalnih transformacija.

Međutim, ako je odnos između varijable odgovora i prediktora veoma nelinearan, model sa mešovitim efektima možda nije najbolji izbor. U takvim slučajevima, fleksibilniji pristup modeliranju, kao što su generalizovani aditivni modeli (GAM) ili modeli mašinskog učenja kao što su neuronske mreže, može biti prikladniji.

4.1.1 Ograničenja modela

Iako modeli mešovitih efekata imaju mnoge prednosti za analizu longitudinalnih podataka, oni takođe imaju neka ograničenja i potencijalne nedostatke. Neke od uobičajenih kritika modela mešovitih efekata su:

- *Pretpostavke* - Oslanjaju se na nekoliko pretpostavki, uključujući normalnost slučajnih efekata, linearnost fiksnih efekata i homoskedastičnost grešaka. Kršenje ovih pretpostavki može dovesti do pristrasnih ili neefikasnih procena parametara modela.

- *Složenost modela* - Mogu biti računarski intenzivni, posebno za velike skupove podataka ili složene modele sa mnogo slučajnih efekata. Složenost modela takođe može otežati tumačenje rezultata.

- *Pogrešna specifikacija modela* - Izbor odgovarajuće strukture slučajnih efekata i kovarijacione matrice može biti težak, a u slučaju greške može doći do pristrasnih procena i netačnih zaključaka.

- *Osetljivost na autlajere* - Modeli mešovitih efekata mogu biti osetljivi na odstupanja, jer oni mogu imati veliki uticaj na procene slučajnih efekata.

- *Pretpostavka linearnosti* - U opštem slučaju pretpostavlja se da su odnosi između prediktora i promenljive odgovora linearni, što, kao što smo gore naveli, nije nužno slučaj.

Uprkos ovim ograničenjima, modeli mešovitih efekata se i dalje široko koriste i korisni su za analizu longitudinalnih podataka, posebno kada podaci imaju hijerarhijsku strukturu ili kada postoji više nivoa varijacija koje treba uzeti u obzir.

4.1.2 Korišćenje u R-u

Nakon što smo opisali model, hajde da se osvrnemo na njegovu primenu na realnim podacima. Za modele mešovitih efekata koristimo paket *lme4*, a longitudinalne podatke osnovu kojih ćemo praviti model uzećemo iz baze *sleepstudy* koju smo ranije pominjali.

Kako bismo napravili regresioni model zavisnosti reakcije od ostalih prediktora koristićemo funkciju *lmer* u okviru gorepomenutog paketa.

```
model <- lmer(Reaction ~ Days + ( Days | Subject), data =
   sleepstudy)
```

Model ima za cilj da objasni odnos između vremena reakcije (tj. varijable ishoda) i broja dana deprivacije sna (tj. varijable prediktora), uzimajući u obzir činjenicu da različiti učesnici mogu imati različite obrasce odgovora tokom vremena.

Prilikom pravljenja modela, primećujemo da postoji oznaka *Days | Subject* koja označava da nagib odnosa između prediktora i zavisne promenljive može varirati u zavisnosti od ispitanika. Ovo je termin slučajnih efekata jer on uzima u obzir varijacije u efektu *Days*“ među ispitanicima, umesto da se pretpostavlja da je ovaj efekat konstantan kod svih učesnika.

Uključivanjem termina slučajnih efekata u model, možemo da procenimo varijaciju unutar subjekta, kao i varijaciju između subjekata, što može dovesti do preciznijih i pouzdanijih procena parametara. U ovom slučaju, termin slučajnih efekata *Days | Subject* pomaže da se uzme u obzir činjenica da različiti učesnici mogu imati različita osnovna vremena reakcije i mogu različito reagovati na nedostatak sna, što može uticati na nagib odnosa.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9536 -0.4634  0.0231  0.4634  5.1793
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 612.10   24.741
##           Days         35.07    5.922   0.07
##  Residual             654.94   25.592
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  251.405      6.825  36.838
## Days          10.467      1.546   6.771
##
```

```
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138
```

Pozivanjem funkcije *summary*, dobijamo pregled napravljenog modela. Kao što je gore navedeno, u okviru Modela meštovitih efekata za procenu parametara koristi se *Metod ograničene maksimalne verodostojnosti - REML* kao mera uklapanja modela u modele linearnih mešovitih efekata i njegovi rezultati ukazuju da je naš model konvergirao sa datom vrednošću pri konvergenciji.

Uzevši u obzir da je srednja vrednost reziduala blizu nule $(0,0231)$, sa opsegom vrednosti između $-3,9536$ i $5,1793$, možemo zaključiti da je većina reziduala blizu nule i da je pretpostavka o normalnoj raspodeli i srednjoj vrednosti reziduala adekvatna. Ipak, relativno velika maksimalna vrednost reziudala u poređenju sa interkvartilnim opsegom sugeriše da u podacima mogu postojati autlajeri ili uticajne tačke.

Na osnovu analize slučajnih efekata, možemo primetiti ocenjenu disperziju slučajnih efekata, standardnu devijaciju za dva faktora grupisanja u modelu i ocenjenu korelaciju između dva slučajna efekta. Sve gorepomenuto ukazuje da postoje tri izvora varijabilnosti u podacima:

Konkretno, izlaz ukazuje da postoje tri izvora varijabilnosti u podacima: varijabilnost između pojedinaca ima varijansu $612,10$ i standardnu devijaciju $24,741$, dok varijabilnosti u efektu dana na reakciju između pojedinaca (nasumični nagibi) odgovara drugi red rezultata. Preostala varijabilnost unutar pojedinca odgovara trećem redu rezultata. Ove procene pružaju informacije o količini varijabilnosti u podacima koja je objašnjena svakim od izvora varijacije. Procene slučajnih efekata sugerišu da postoji značajna varijabilnost u prekidima i nagibima između pojedinaca, što može odražavati razlike u početnim vremenima reakcije i/ili razlike u načinu na koji pojedinci reaguju na deprivaciju sna.

Vrednosti koje se dobijaju prilikom procene fiksnih efekata pokazuju da se procenjena vrednost zavisne promenljive, kojoj odgovara prvi red tabele, značajno razlikuje od nule, što se može zaključiti i relativno velike t-vrednosti i relativno male standardne greške. Za efekat dana može se zaključiti slično. Korelacija između dana i zavisne promenljive je negativna, što znači da će reakcije biti manje sa povećanjem broja dana i obrnuto.

Generalno gledano, rezultati sugerišu da se model dobro uklapa u podatke.

4.2 Model krive rasta

Model krive rasta je model koji se koristi za analizu podataka gde se varijable menjaju tokom vremena. U ovom modelu putanju promene srednjeg odgovora Y u zavisnosti od vremena modelujemo običnom polinomnom, tj. eksponencijalnom funkcijom.

Najčešći model krive rasta je linearni model krive rasta koji je dat jednačinom:

$$Y_{ij} = \beta_{0i} + \beta_{11} \times vreme_{ij} + \epsilon \tag{19}$$

Gde je Y_{ij} promenjiva ishoda, koja predstavlja promenjivu ishoda pojedinca i u trenutku j, β_{0i} i β_{10} su koeficijenti modela za i-tog ispitanika, a $vreme_{ij}$ predstavlja vremensku promenjivu i-tog ispitanika u j-tom trenutku.

Vremenska promenjiva je veoma bitna za ovaj model jer predstavlja promenu Y tokom vremena, te nam uključivanje ove promenjive u model omogućava da procenimo promenu Y tokom vremena, početno stanje promenjive Y i predvidimo buduće vrednosti ove promenjive.

Model krive rasta se može proširiti i na složenije oblike funkcija poput polinoma višeg stepena, eksponencijalne funkcije i slično.

4.2.1 Model slučajnih koeficijenata

Model slučajnih koeficijenata je vrsta modela krive. Za razliku od klasičnog modela krive, ovaj model omogućava varijacije u parametrima rasta (nagib i presek) među pojedincima u uzorku. Ovo nam omogućava individualne razlike u promeni zavisne promenjive. To znači da svaki pojedinac ima svoj jedinstven skup podataka koji ga opisuje i on se menja tokom vremena. Model procenjuje varijabilnost parametara rasta među pojedincima kao i prosečne vrendosti parametara rasta u populaciji.

Parametri rasta su skup parametara koji opisuje promenu putanje varijable tokom vremena. U modelima krive rasta to su *nagib* i *presek*. Presek se definiše kao početno vreme ili prvi trenutak merenja i može smatrati početnim nivom zavisne promenjive koja se očekuje za sve nivoe ishoda. Nagib predstavlja stopu promene zavisne promenjive od vremena. Nagib može biti pozitivan i negativan što ukazuje na pozitivne i negativne promene zavisne promenjive.

Model slučajnih koeficijenata je koristan kada postoji heterogenost u stopi i obrascima rasta zavisne promenjive između pojedinaca. Neki ispitanici mogu pokazati brzi porast zavisne promenjive, dok kod drugih vrednost može sporije da raste. Dopuštajući individualne promene u parametrima rasta dobijamo puno bolji model.

4.2.2 Ograničenja modela

Iako modeli krive rasta imaju mnoge prednosti za analizu longitudinalnih podataka, oni takođe imaju neka ograničenja i potencijalne nedostatke. Neke od uobičajenih kritika krive rasta su:

- *Generalizacija* - Model se obično koristi kako bi opisao jednu populaciju pod posebnim uslovima u specifičnom vremenskom periodu, pa da generalizacija na neke druge populacije obično nije moguća.

- *Vreme* - Model krive rasta se oslanja na ispitivanja koja se dogadjaju u dugom vremenskom periodu jednakim vremenskim intervalima, što može biti veoma zahtevno za istraživače.
- *Kompleksnost* - Model krive rasta može biti veoma kompleksan ako se uvede i klasterovanje podataka, i to bitno otežava tumaćenje podataka.
- *Pretpostavke modela* - Model krive rasta traži određene predpostavke o raspodeli, te ukoliko se te pretpostavke naruše mogu se dobiti loši rezultati pri čemu se greška teško pronalazi.
- *Interpretacije parametara* - U opštem slučaju interpretacija parametara ovog modela može biti teška i zbog toga se mora napraviti veoma precizna analiza prediktora.

4.2.3 Korišćenje u R-u

Nakon što smo opisali model, hajde da se osvrnemo na njegovu primenu na realnim podacima. Za modele krive rasta koristimo paket *nlme*, a longitudinalne podatke osnovu kojih ćemo praviti model uzećemo iz baze *Oxboys* o kojoj smo ranije pričali. Ona sadrži podatke o 26 dečaka kojima je praćena visina tokom 8 godina.

Kako bismo napravili regresioni model zavisnosti visine od ostalih prediktora koristi cemo funkciju *lme* u okviru gorepomenutog paketa.

```
data(Oxboys)

Oxboys$age <- as.numeric(Oxboys$age)
model <- lme(height ~ age, random = ~ age | Subject, data =
   Oxboys)

ggplot(Oxboys, aes(x = age, y = height)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE)
```

Pošto funkcija *lme* zahteva numeričke promenljive kao ulaz, potrebno je konvertovati promenljivu *age* u numeričku.

Pravimo model krive rasta koristeći funkciju *lmer*. Varijabla visine je promenljiva odgovora, a starost ispitanika je prediktor.

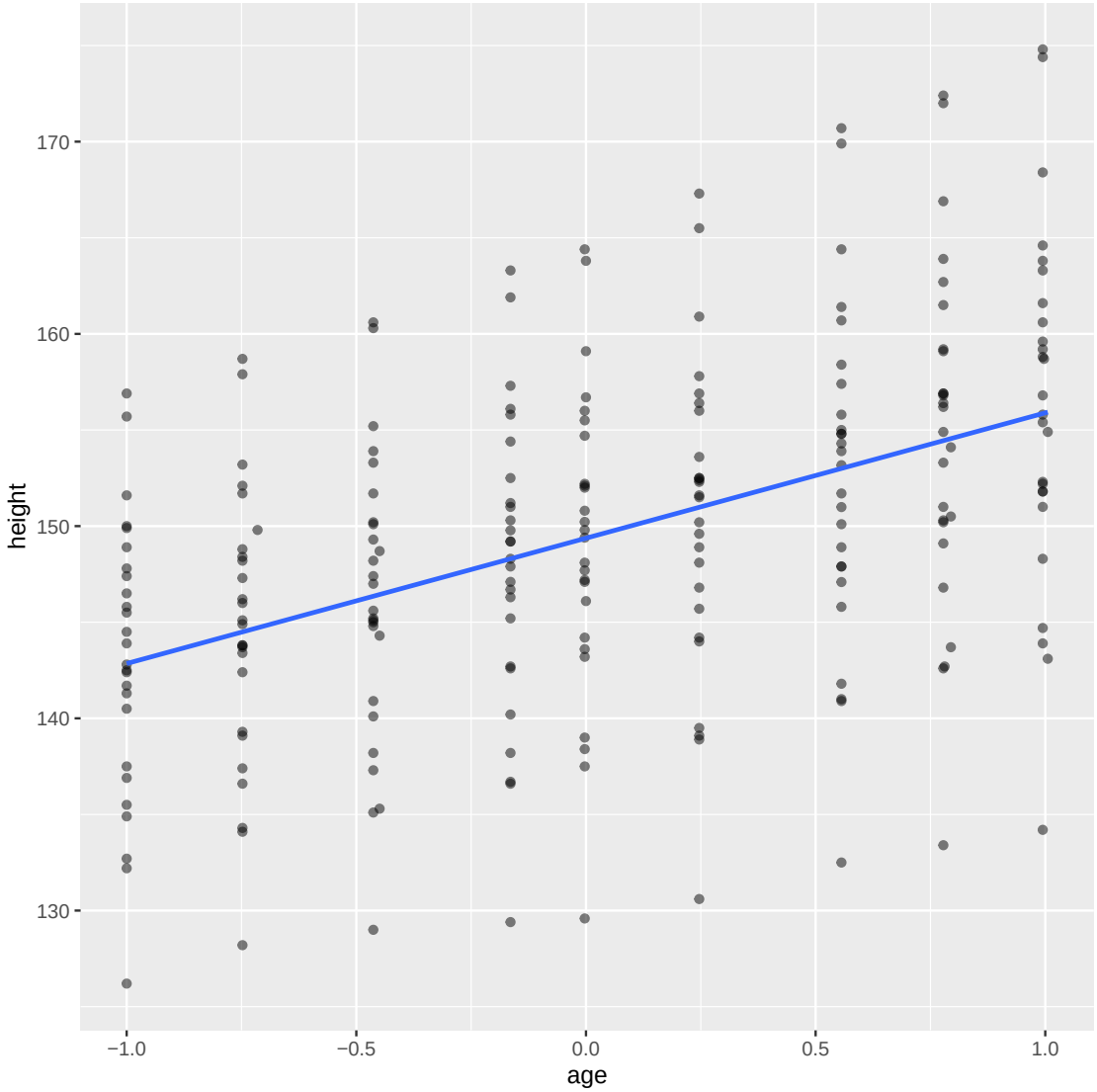
U kontekstu modela krive rasta *random = age | Subject* specificira strukturu slučajnih efekata modela. To znači da postoje nasumični preseci i nagibi za efekat starosti za svakog pojedinačnog subjekta u skupu podataka. Tačnije, termin *age | Subject* ukazuje da su slučajni efekti specifični za svaki subjekt i da se vrednost starosti koristi kao prediktor u ovim slučajnim efektima. To znači da svaki subjekt ima svoj presek i nagib, što omogučava individualne varijacije u putanji rasta tokom vremena.

U ovom modelu krive rasta parametri rasta uključuju fiksni efekat starosti i slučajni efekat starosti za svakog subjekta. Fiksni efekat starosti obuhvata ukupni odnos između starosti i visine, dok slučajni efekat starosti za svakog subjekta obuhvata individualnu varijabilnost u

odnosu između starosti i visine. Ovi slučajni efekti predstavljaju odstupanje putanje rasta svakog subjekta od ukupne prosečne trajektorije rasta.

Pomoću funkcije *ggplot* vizuelno predstavljamo ponašanje podataka, a kako podaci pokazuju obrazac povećanja ili smanjenja visine sa godinama, ova linija se može smatrati grubom aproksimacijom krive rasta.

```
## `geom_smooth()` using formula 'y ~ x'
```



Nagib linije označava prosečnu brzinu promene visine u odnosu na uzrast, a presek predstavlja visinu dečaka u određenom uzrastu (u ovom slučaju 2 godine). Dakle, linija obuhvata ukupni trend rasta dečaka u visinu tokom vremena, čineci je vrstom krive rasta.

Ovaj model se može smatrati tipom modela krive rasta jer opisuje odnos između visine i starosti, uzimajući u obzir individualnu varijabilnost među subjektima u odnosu na njihove putanje rasta tokom vremena.

4.3 Metode za popravljanje modela

Kako bismo napravili što adekvatniji model koji odgovara našim podacima potrebno je da poznajemo metode za popravku modela. Kod logitudialnih studija postoje dva osnovna problema: nedostajući podaci i različito vreme uzorkovanja ispitanika. Postoji puno načina da se nosimo sa ovim problemima, a neke od njih ćemo u nastavku obrazložiti.

4.3.1 Nedostajući podaci

Kao što je već navedeno, čest problem koji se može javiti u okviru longitudinalnih studija je da nam nedostaju određeni podaci, te treba odabrati adekvatan pristup rešavanju ovog problema.

Najlakši način za rešavanje ovog problema je izbacivanje nedostajućih podataka iz modela. Ovakav pristup nije pogodan jer može dovesti do loših rezultata u daljoj analizi i korelacija među podacima koji inicijalno uopšte nisu povezani. Međutim, ovaj pristup je koristan u situaciji kada je skup podataka koji nedostaju dovoljno mali da njihovo izbacivanje neće narušiti naš model.

U *R*-u se izbacivanje podataka može izvršiti korišćenjem funkcije *na.omit()*.Demonstriraćemo ovaj postupak na sledećoj bazi:

```
df1<- data.frame(x= 1:10, y = c(1, 2, 3, NA, 5, NA, 0, 8, NA,
   3))
d1NA<-na.omit(df1)
d1NA
```

Iako su u okviru baze *df1* postojali nedostajući podaci, primetimo da su podaci koji sadrže *NA* vrednosti izbačeni iz novonastale baze.

```
##     x y
## 1   1 1
## 2   2 2
## 3   3 3
## 5   5 5
## 7   7 0
## 8   8 8
## 10 10 3
```

Drugi način za nošenje sa nedostajućim podacima je *imputacija*. Imputacija podrazumeva postupak *pravljenja* podataka koji nedostaju od već postojećih podataka i time popunimo *"rupu"* u podacima koje obrađujemo.

U okviru *Smooting techniques* dela istakli smo neke od teorijskih pristupa pomoću kojih možemo popraviti podatke koje nam nedostaju. Sada ćemo kratko objasniti metod *LOCF (Last observation carried forward)* kao jedan od potencijalnih načina rešavanja gorepomenutog problema. Naime, on funkcioniše tako što uzima poslednji zabeleženi podatak i prenosi ga na novu opservaciju dok novi podatak ne bude zabeležen, pa prediktor ostaje konstantan tokom vremena. Valja pomenuti da je to i mana ovog pristupa, jer pretpostavka da će prediktor ostati konstantan tokom vremena može imati velike posledice na celu studiju.

U *R*-u se rad sa *LOCF* može izvršiti korišćenjem funkcije *na.locf* koja se nalazi u okviru *zoo* paketa. Demonstriraćemo ovaj postupak na sledećoj bazi:

```
library(zoo)
df<- data.frame(x= 1:10, y = c(1, 2, 3, NA, 5, NA, 7, 8, NA,
   10))
df$locf <- na.locf(df$y)
cbind(df[, c("x", "y")], df$locf)
```

Nedostajuće podatke u okviru baze *df* u ovom slučaju zamenile su vrednosti dobijene prethodnom opservacijom.

```
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##     x  y df$locf
## 1   1  1       1
## 2   2  2       2
## 3   3  3       3
## 4   4 NA       3
## 5   5  5       5
## 6   6 NA       5
## 7   7  7       7
## 8   8  8       8
## 9   9 NA       8
## 10 10 10      10
```

Treći način na koji možemo rešiti problem nedostajućih podataka je *dodavanje težina*. Tačnije, težine se dodaju onim opservacijama za koje je veća verovatnoća da će biti zabeležene. Cilj nam je da dobijemo podatke koji nisu pristrasni i koji ne narušavaju samu strukturu modela. Jedan od uobičajenih načina za dodavanje težine je *IPW - Inverse probability weighting*. Težine koje se dodeljuju podacima se izračunavaju tako što prvo iskoristimo model koji na osnovu prethodnih zapažanja predviđa verovatnoću da neki podatak nedostaje,a zatim svakoj opservaciji dodelimo težinu koja je jednaka izverznoj vrednosti te verovatnoće. Na ovaj način veća težina se daje onim opservacijama čija je verovatnoća događanja manja.

Kako bismo primenili *IPW* u *R*-u koristićemo paket *ipw* i funkciju *ipwpoint* na podacima koje smo generisali u cilju dobijanja težine. Demonstriraćemo ovaj postupak na sledećoj bazi:

```
n <- 1000
simdat <- data.frame(l = rnorm(n, 10, 5))
a.lin <- simdat$l - 10
pa <- exp(a.lin)/(1 + exp(a.lin))
simdat$a <- rbinom(n, 1, prob = pa)
simdat$y <- 10*simdat$a + 0.5*simdat$l + rnorm(n, -10, 5)
```

```
temp <- ipwpoint(exposure = a,
  family = "binomial",
  link = "logit",
  numerator = ~ 1,
  denominator = ~ l,
  data = podaci)
  temp$ipw.weights
```

Dakle, *ipw.weights* je vektor koji sadrži inverzne verovatnoće za svako merenje, a težine se vraćaju redom kojim su uzete iz opservacije.

4.3.2 Uzorkovanje u različitim vremenskim intervalima

Analiza longitudinalnih podataka za koje važi da ispitanici nisu odgovarali u istom trenutku na pitanja može biti veoma problematična. Problemi koji mogu nastati su da su se odgovori nekih ispitanika promenili u razlici vremana u kojoj smo izvršili ponovno istraživanje u odnosu na ostale ispitanike.

Ovo se najviše odnosi na varijable koje se u studiji menjaju drastično tokom vremena, i one imaju veliki uticaj na sam ishod studije.

Postoji dosta metoda koje se mogu koristiti u cilju kontrolisanja ovih podataka. Da bismo smo kontrolisali njihov uticaj možemo uvesti novi prediktor koji opisuje ponašanje te varijable u svakom trenutku našeg istraživanja na osnovu postojećih rezulatata. Kako je prethodnu ideju teško realizovati, ovaj metod se retko koristi.

Drugi način za rešavanje ovog problema je procena klauzalnih efekata.Kauzalni efekti su efekti koji pokušavaju da objasne vezu uzroka i posledice. Ovi modeli nam omogućavaju da procenimo uticaj varijable na sam ishod studija, ali kako spadaju u veoma napredne algebarske i statističke metode, ovde navidmo samo mogućnost za njihovo korišćenje.

5 Reziduali

Analiza longitudinalnih podataka nije kompletna bez analize reziduala modela. Reziduali se mogu koristiti u cilju ocenjivanja modela, ali i detekcije autlajera.

Kod longitudinalnih podataka definišemo rezidualni vektor za svakog ispitanika:

$$r_i = Y_i - X_i\hat{\beta} \tag{20}$$

Valja napomenuti da vektor reziduala ima očekivanje nula. Takođe, reziduali se mogu koristi za proveru sistemskih odstupanja i očekivanih vrednosti.

Grafik reziduala u odnosu na grafik očekivanih vrednosti može nam pokazati određene trednove i pravilnosti u modelu. U dobro napravljenom modelu dijagram treba da prikazuje podatke koji su raštrkani oko nule. Dodatno ovaj dijagram nam može pokazati da li je neophodno da popravimo naš model dodavanjem kvadratnog člana ili težina. Grafički prikaz reziduala se koristi za detekciju neslaganja u modelu, i prikaz nekih podataka koji su udaljeni i koje bi možda trebalo dodatno razmotriti.

Postoje dve veoma bitne osobine longitudinalnih podataka koje trebamo imati na umu. Prva je da su komponente vektora reziduala korelisane i da ne moraju nužno imati istu disperziju. Podsetimo se da je očekivanje reziduala u modelu nula, što na neki način oponaša očekivanje kod grešaka modela, međutim njihove kovarijaciione matrice neće biti iste. U cilju olakšavanja rada sa rezidualima, pretpostavićemo da jesu jednake, odnosno:

$$Cov(r_1) \approx Cov(e_i) = \sum_i \tag{21}$$

Pošto za reziduale pretpostavljamo da približno imaju istu kovarijacionu matricu kao greške, to će takođe imati veliki uticaj na ispitivanje reziduala. Za početak, pošto disperzija nije konstantna, dijagram reziduala u odnosu na predviđanje neće nužno imati konstantan opseg.

Druga bitna osobina na koju treba obratiti pažnju je korelisanost prediktora i reziduala. Kod linearne regresije reziduali i prediktori nisu korelisani, ali kada radimo sa longitudinalnim podacima to neće biti slučaj tako da na dijagramu možemo uočiti trendove u odnosu na izabrani prediktor.

5.0.1 Transformisanje reziduala

Iako postoji mnogo načina da se transformišu rezudali, ono što želimo da postignemo ovim transformacijama je da reziduali longitudinalnih podataka oponašaju osobine reziduala kod lineranih modela.Zato reziduale transformišemo tako da im očekivanje bude nula, a disperzija konstantna, što se može postići korišćenjem Čoleski dekompozicije.

Pošto smo procenili da je približnu vrednost kovarijacione matrice $\hat{\sum}_i$, Čoleski dekompoziciju možemo iskoristi da napravimo donje trougaonu matricu L_i za koju važi:

$$\hat{\sum_i} = L_i L_i^t \tag{22}$$

Nakon ove transformacije koristimo matricu L^{-1}, kako bismo transformisali reziduale tako da gorenavedene osobine.

$$r_i^* = L_i^{-1} r_i = L_i^{-1}(Y_i - X_i\hat{\beta}) \tag{23}$$

Ovako transformisani reziduali imaju iste osobine kao oni u linearnoj regresiji. Prvi element r_i će biti popravljeni rezidual prvog merenja (polazno stanje sistema), a transformisani reziduali od drugog do poslednjeg elementa su odstupanja od procenje vrednosti u odnosu na sva prethodna zapažanja. Tako da je procenjena vrednost k-tog člana u rezidualnom vektoru r_i:

$$\frac{Y_{ik} - E(Y_{ik}|Y_{i1}....Y_{ik-1})}{\sqrt{Var(Y_{ik}|Y_{i1}....Y_{ik-1})}} \tag{24}$$

Ova procena kod transformisanih reziduala nije najočiglednija i zahteva malo dublju analizu pomoću linearne algebre.

Kada smo popravili rezidual, za dijagnostiku možemo koristi sve tehnike koje se koriste u dijagnostici u standardnim linearnim modelima. Na primer, možemo konstruisati grafik r_{ij}^* u zavisnosti od μ_{ij}^*, gde je $\mu_i^* = L^{-1}\mu_i$. Ukoliko je popravka reziduala bila dobra, oni bi sada trebalo da budu centrirani oko nule. Možemo da konstruišemo grafik popravljenih reziduala u zavisnosti od vremena, te pomoću njega da uočimo da li je model koji koristimo adekvantan, kao i primetimo neke promene koje se događaju tokom vremena. Popravljeni reziduali su nam adekvatniji u slučaju da želimo da detektujemo autlajere. Kao i u dijagnostici standardnih linearnih modela, možemo koristiti *qqplot* kako bismo proverili normalnu raspodelu reziduala.

Kao što je već rečeno, reziduali se mogu koristiti za detekciju autlajera, ali još jedna osobina koju valja istaći jeste da se uz pomoć reziduala može primetiti koji ispitanik odudara od istraživanja. Naime, za svakog ispitanika možemo izračunati udaljenost između njegove uočene i prilagođene mere, a za računanje tog rastojanja koristi se *Mahalanobisova udaljenost*:

$$d_i = r_i^{*t} r_i \tag{25}$$

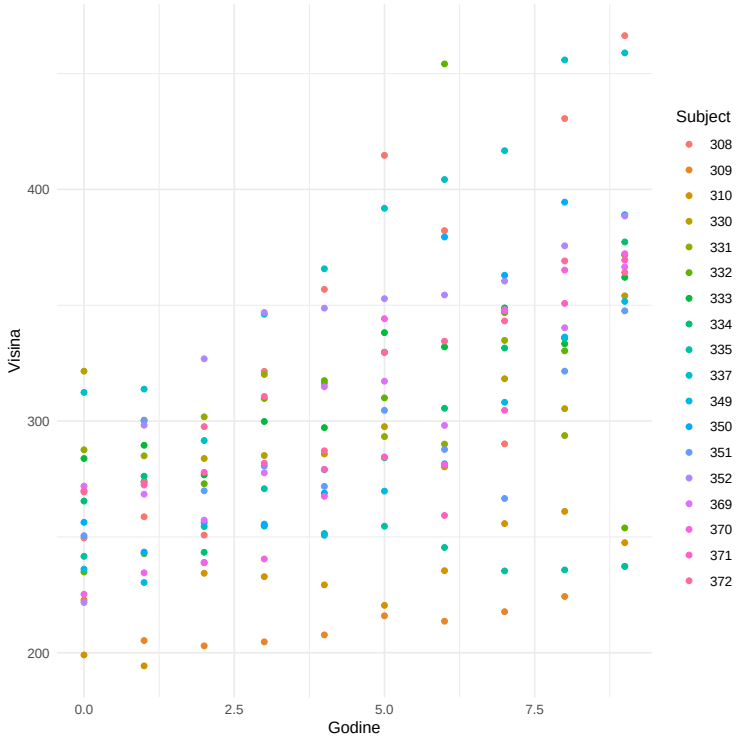
Ako je model dobro definisan d_i ima χ^2 raspodelu, sa onoliko stepeni slobode kolika je dimenzija r_i^*. Ispitanici koji nisu pogodni za naše istraživanje će imati male p-vrednosti, te ćemo moći da ih detektujemo. Ova metoda nije u potpunosti precizna, tako da bi odbacivanje subjekta trebalo dodatno ispitati u slučaju da je p-vrednost na granici.

Kao što je gore navedno, reziduali se mogu ispitivati pomoću grafika zavisnosti vremena ili srednjeg odgovara. U dobro definisanom modelu grafik zavisnosti r_{ij}^* bi trebali da budu konstantan u zavisnosti od μ_{ij}^*. Najčešće se za grafičku reprezentaciju reziduala koristi njihova apsolutna vrednost u zavisnosti od vremena ili srednjeg odgovora.

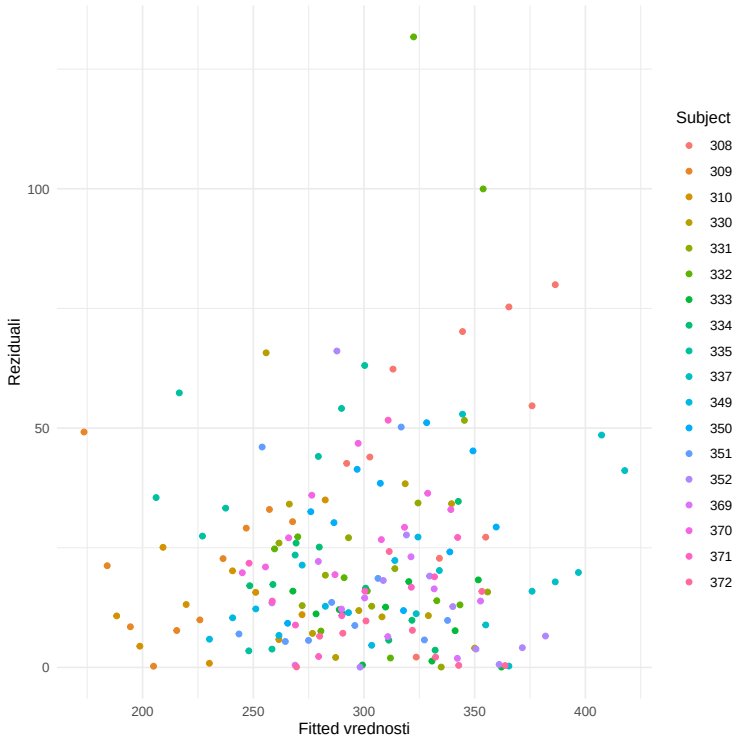
5.1 Ispitivanje reziduala u R-u

Da bismo popravili naš model ispitaćemo da li su reziduali heteroskedastični i normalno raspoređeni i pogledaćemo da li postoje autlajeri.

Kako bismo demonstrirali postupak, koristićemo model koji smo prethodno napravili za demonstraciju modela mešovitih efekata. Prvo ćemo pogledati *time plot*, kako bismo potencijalno uočili autlajere i primetićemo da ne postoje tačke koje posebno odudaraju od ostalih podataka.

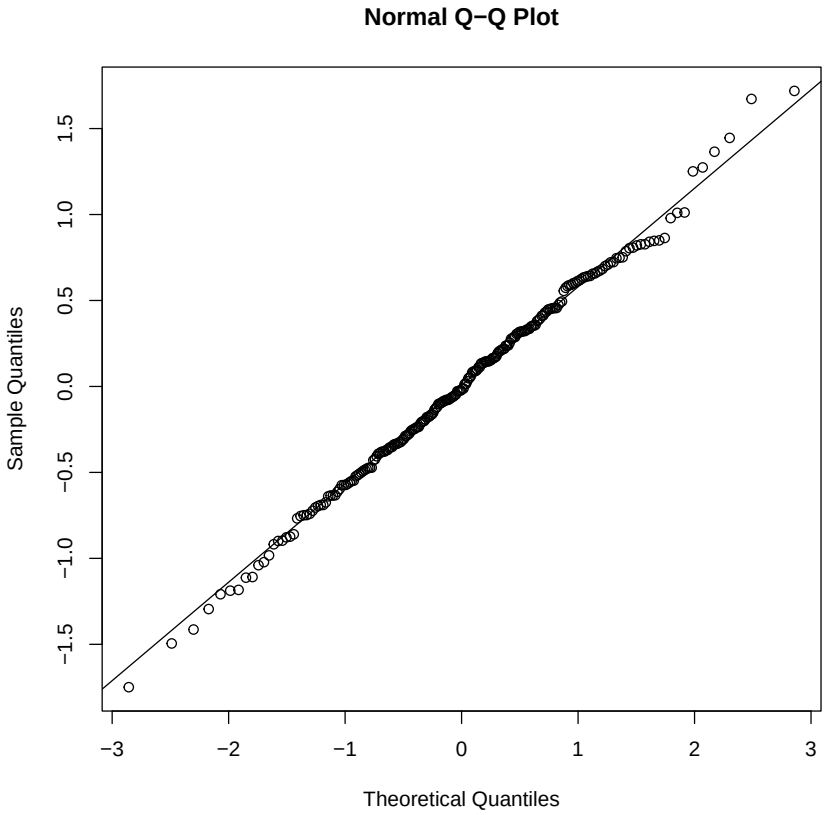


Prva pretpostavka je da greške imaju jednaku disperziju, tj. *heteroskedastičnost*. Nju proveravamo grafički, sa grafika zavisnosti reziduala od predviđenih vrednosti. Umesto reziduala možemo gledati kvadrate reziduala ili njihove norme.



Primećujemo da će disperzija biti na nekom sličnom nivou, tj. ne uočavamo veliki porast.

Kako bismo proverili normalnu rapodelu za reziduale koristimo *Q-Q plot*.



Primećujemo da reziduali uglavnom prate pravu liniju, tako da je pretpostavka o normalnoj raspodeli rezidaula opravdana.

6 Primer obrade longitudinalnih podataka u *R*-u

Za ilustraciju longitudinalnog istraživanja posmatraćemo bazu podataka *Orthodont* iz paketa *nlme*. Ova baza sadrži ponovljena merenja udaljenosti između gornjih i donjih zuba kod dece koja su podvrgnuta ortodontskom lečenju. Za početak, učitavamo bazu podataka:

```
data(Orthodont)
head(Orthodont)
```

Baza podataka *Orthodont* sastoji se od sledećih promenljivih:

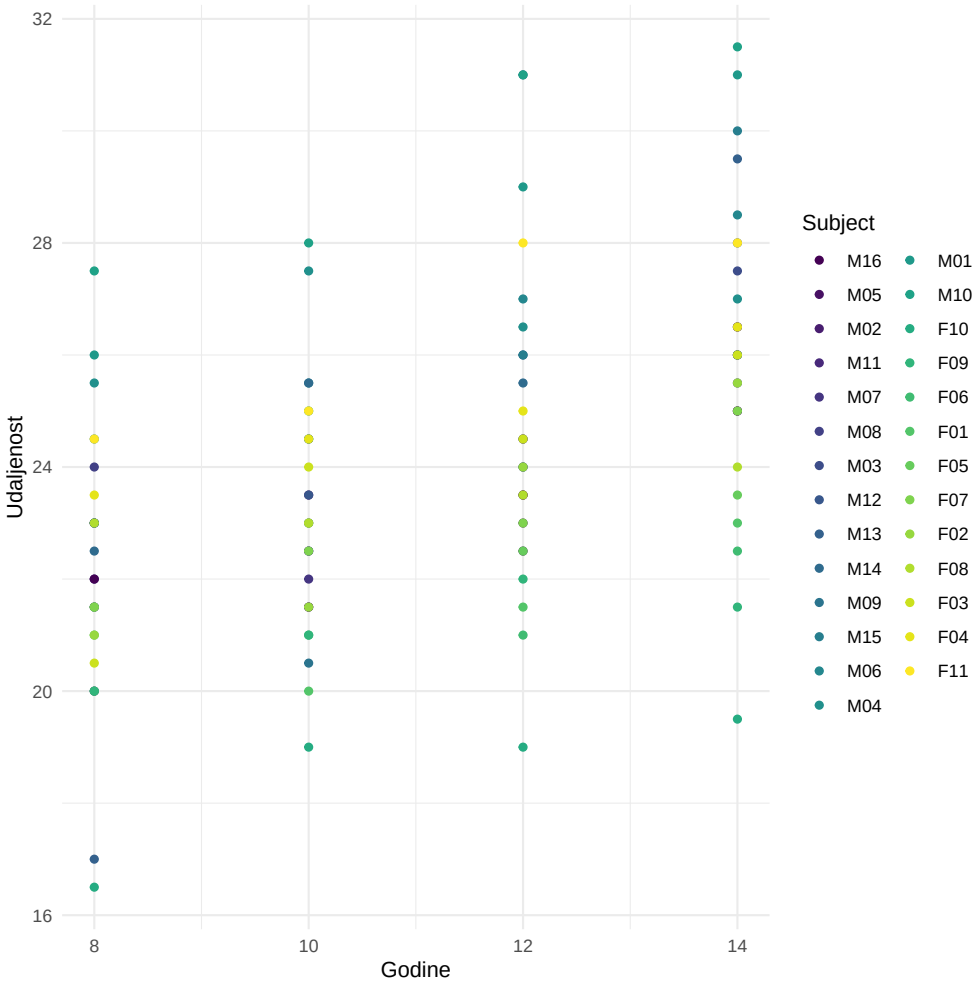
- Distance - *Rastojanje između hipofize i pterigomaksilarne fisure, mereno u milimetrima.*
- Age - *Godine ispitanika u trenutku ispitivanja.*
- Subject - *Identifikaciona šifra ispitanika.*
- Sex -*Pol ispitanika*

```
## Grouped Data: distance ~ age | Subject
##   distance age Subject  Sex
## 1     26.0   8     M01 Male
## 2     25.0  10     M01 Male
## 3     29.0  12     M01 Male
## 4     31.0  14     M01 Male
## 5     21.5   8     M02 Male
## 6     22.5  10     M02 Male
```

Istraživanje je uključivalo 27 dece, tj. 11 devojčica i 16 dečaka. Skup podataka ima 108 redova i 3 kolone. Svaki red predstavlja jedno posmatranje udaljenosti i starosti za određeno dete u određenoj vremenskoj tački. Skup podataka ima grupisanu strukturu, sa više zapažanja po subjektu. Pozivanjem funkcije *summary* uočavamo glavne karakteristike podataka.

```
##     distance          age          Subject       Sex
##  Min.   :16.50   Min.   : 8.0   M16    : 4   Male  :64
##  1st Qu.:22.00   1st Qu.: 9.5   M05    : 4   Female:44
##  Median :23.75   Median :11.0   M02    : 4
##  Mean   :24.02   Mean   :11.0   M11    : 4
##  3rd Qu.:26.00   3rd Qu.:12.5   M07    : 4
##  Max.   :31.50   Max.   :14.0   M08    : 4
##                                 (Other):84
```

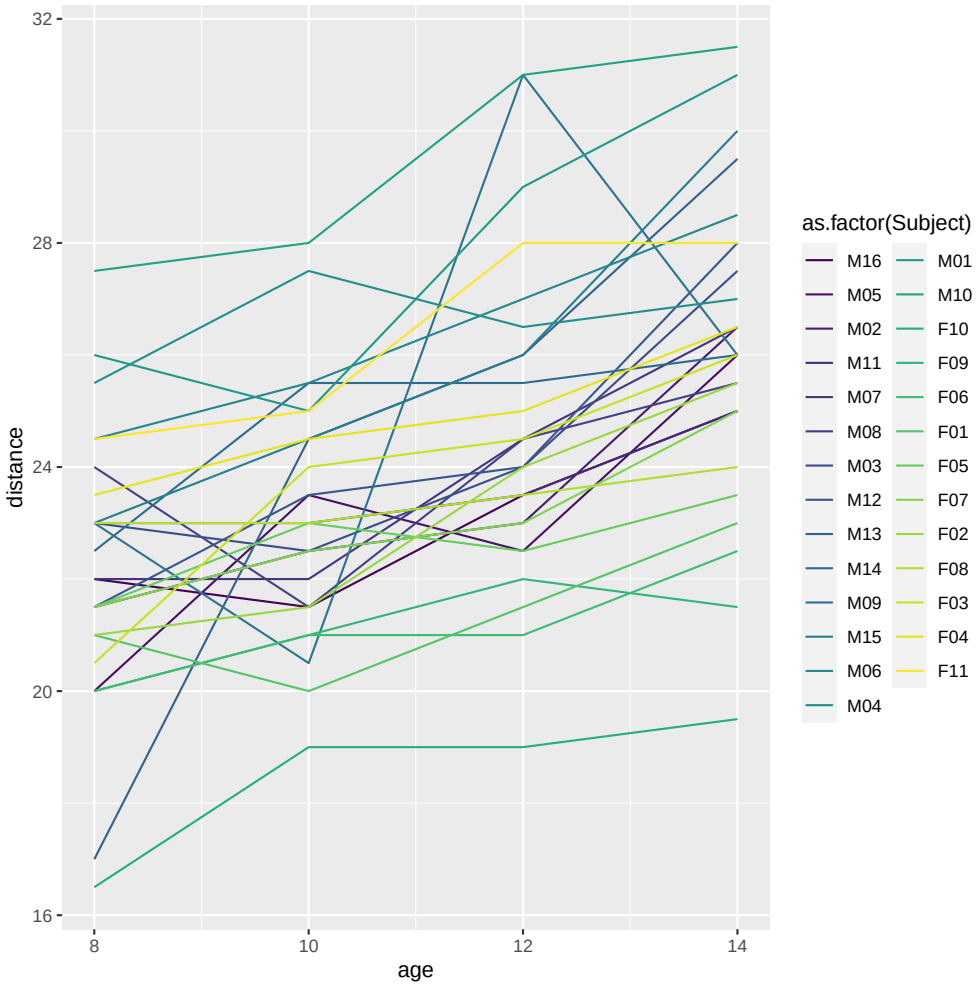
Sada ćemo grafički predstaviti podatke pomoću *time plot*-a koji smo ranije koristili.



Ispitivanjem *Time plot*-a datog skupa podataka, možemo videti da postoji opšti trend povećanja udaljenosti između dve tačke kako deca rastu, na šta ukazuje nagib linije koja povezuje tačke naviše. Takođe postoji značajna varijabilnost u podacima, pri čemu neka deca pokazuju veće povećanje udaljenosti od druge, a neka pokazuju smanjenje udaljenosti tokom vremena.

Sada posmatramo *Spaghetti plot* podataka. On pokazuje da postoji značajna individualna varijabilnost u udaljenosti između hipofize i pterigomaksilarne fisure tokom vremena. Neka deca pokazuju brzo povećanje udaljenosti, dok druga pokazuju sporije povecanje ili čak smanjenje udaljenosti u određenim vremenskim tačkama. Ova individualna varijabilnost nije tako očigledna u*Time plot*-u , koji pokazuje prosečnu promenu udaljenosti tokom vremena za sve subjekte zajedno.

Ova vrsta reprezentacije podataka se takođe može koristiti za identifikaciju autlajera ili uticajnih zapažanja. U našem skupu podataka postoji jedan ispitanik koji izgleda da ima neuobičajeno visoko merenje udaljenosti u vremenskoj tački 3. Ovo može biti izuzetak koji zahteva dalje istraživanje.



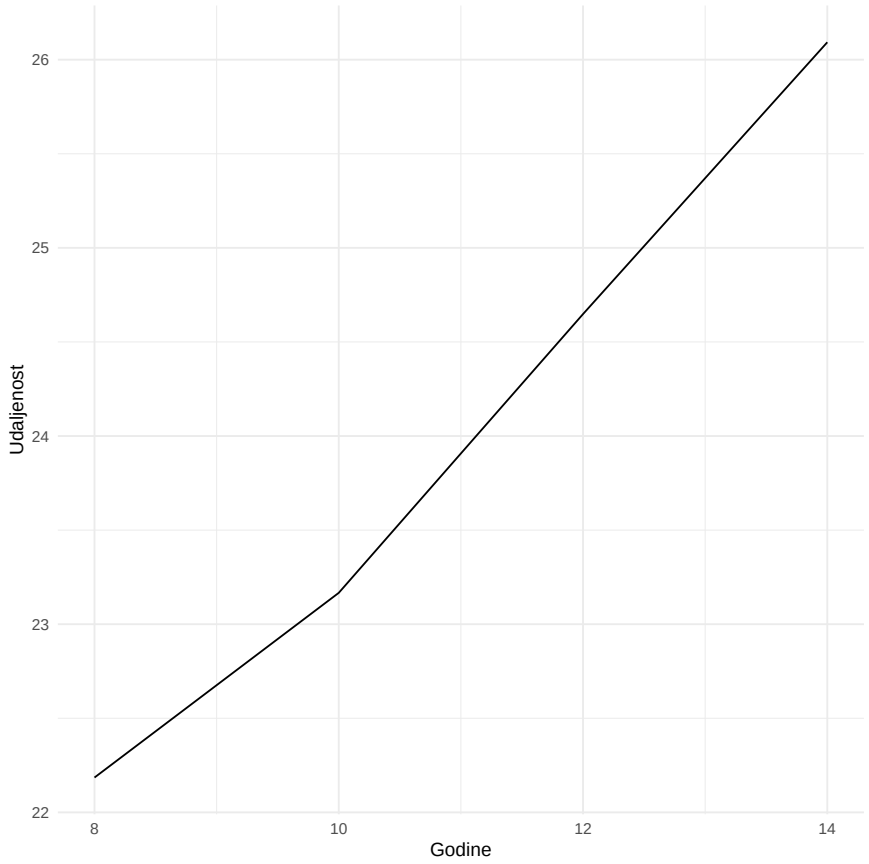

```
## NULL
```

Postoji određena varijabilnost u prosečnoj udaljenosti između zuba za svaku starosnu grupu, što bi moglo sugerisati da drugi faktori (kao što su genetika ili higijena zuba) takođe igraju ulogu u određivanju razmaka između zuba.

Postoji dosta preklapanja u prosečnoj udaljenosti između zuba za susedne starosne grupe, što bi moglo da sugeriše da su promene u razmaku između zuba kako osoba stari relativno postepene i da nije uvek lako uočljive.

Rastojanja između zuba se vremenom povećavaju što se može objasniti odrastanjem deteta ili specifičnim ortodontskim tretmanom koji su ispitanici primili. Ortodontski tretman obično uključuje upotrebu aparata poput proteza koji vrše pritisak na zube kako bi ih pomerili u željeni položaj. Ovaj pritisak može prouzrokovati da se zubi razdvoje, što može dovesti do povecanja izmerenih rastojanja tokom vremena. Međutim, takođe je moguće da su se neke udaljenosti u ortodontskom skupu podataka smanjile kao rezultat ortodontskog tretmana, u zavisnosti od specifičnih ciljeva i korišćenih tehnika.

Sada pravimo grafik srednje vrednosti naših podataka. Na osnovu njega možemo pretpostaviti da je zavisnost između podataka skoro pa linearna, jer srednja vrednost naših podataka odgovara pravoj koja opisuje linearnu zavisnost prediktora x od zavisne promenljive y.



Hajde sada da modeliramo naše podatke. Koristićemo metod mešovitih efekata, tj. *LME* i implementirati ga na sledeći način:

```
data(Orthodont)
model_primer <-  lmer(distance ~ age + Sex + (1 | Subject),
   data = Orthodont)
summary(model_primer)
```

Pozivanjem funkcije *summary*, dobijamo osobine našeg modela. *REML* kriterijum pri konvergenciji je $437,5$, što ukazuje da se model dobro uklapa u podatke, jer nije visoka vrednost kriterijuma.

Na osnovu posmatranja analize reziduala, statistike sugerišu da su reziduali u modelu približno normalno raspoređeni sa srednjom vrednošću oko nule i standardnom devijacijom od približno 1. Činjenica da su minimalne i maksimalne vrednosti relativno daleko od nule sugeriše da mogu postojati neki autlajeri ili ekstremne vrednosti u podacima.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: distance ~ age + Sex + (1 | Subject)
##    Data: Orthodont
##
## REML criterion at convergence: 437.5
##
## Scaled residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.7489 -0.5503 -0.0252  0.4534  3.6575
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 3.267    1.807
##  Residual             2.049    1.432
## Number of obs: 108, groups:  Subject, 27
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 17.70671    0.83392  21.233
## age          0.66019    0.06161  10.716
## SexFemale   -2.32102    0.76142  -3.048
##
## Correlation of Fixed Effects:
##           (Intr) age
## age       -0.813
## SexFemale -0.372  0.000
```

Kada posmatramo analizu slučajnih efekata, uočavamo da postoji samo jedan. Procenjena varijansa ovog slučajnog efekta je $3,267$, a procenjena standardna devijacija je $1,807$ što sugeriše da postoje značajne varijacije u *Incercept*-ima među pojedinačnim ispitanicima.

U modelu se takođe procenjuju rezidualna varijansa i standardna devijacija. Preostala varijansa je $2,049$, a rezidualna standardna devijacija je $1,432$ i one predstavljaju količinu neobjašnjive varijabilnosti u varijabli ishoda koju model ne uzima u obzir. Rezultat sugeriše da slučajni efekat obuhvata značajnu količinu varijabilnosti u podacima, pošto je procenjena varijansa presretanja subjekta veca od preostale varijanse.

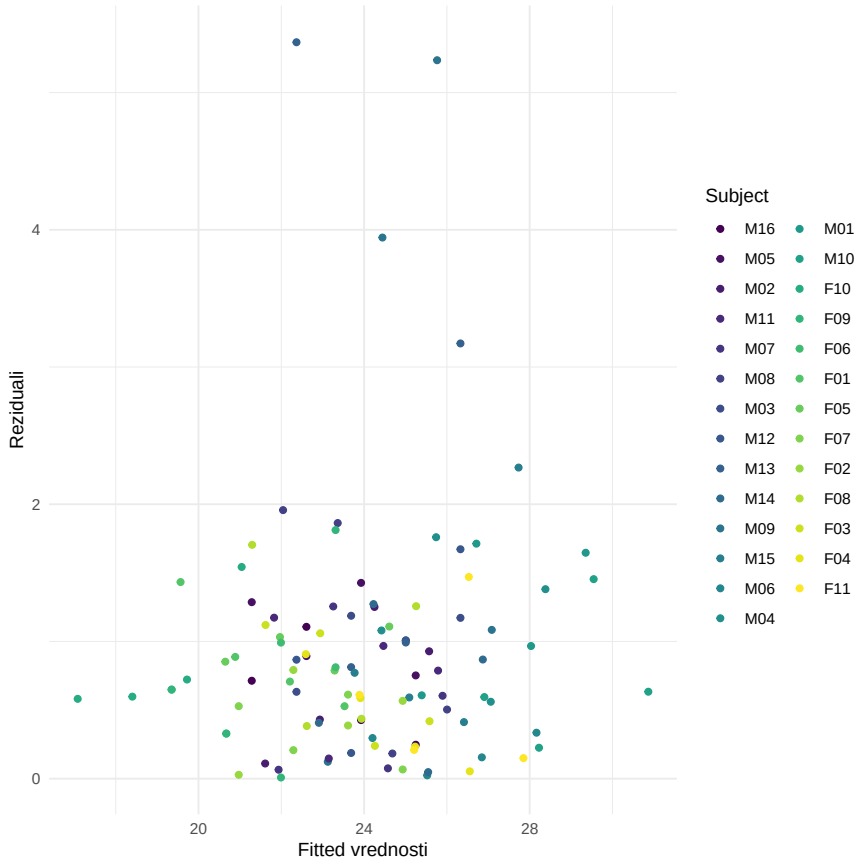
Na osnovu procene fiksnih efekata u modelu, vidimo da postoje 3 fiksna efekta. *Incercept* je $17,70671$, što predstavlja očekivanu vrednost varijable ishoda kada su svi ostali prediktori jednaki nuli. Standardna greška procene *Incercept*-a je $0,83392$, a t-vrednost za testiranje nulte hipoteze da je *Incercept* jednak nuli je $21,233$. Ovo ukazuje da se *Incercept* značajno razlikuje od nule ($p < 0,001$).

Koeficijent za starost je procenjen na $0,66019$, što znači da se za svako povećanje starosti za jednu jedinicu, očekivana vrednost varijable ishoda povecava za $0,66019$ jedinica, držeći sve ostale prediktore konstantnim.

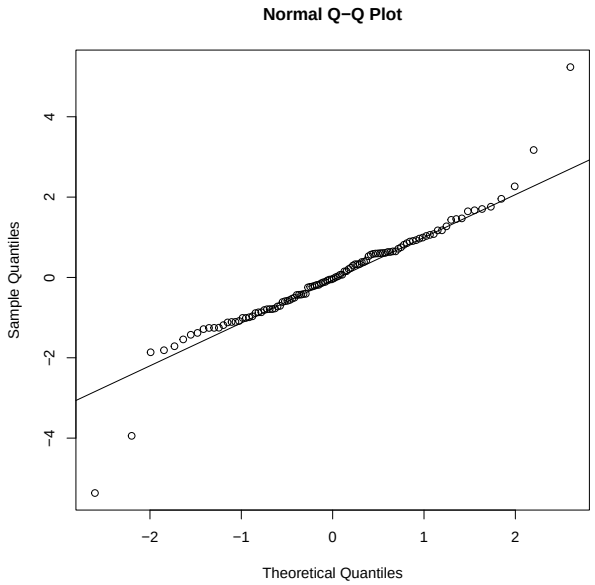
Koeficijent za ženski pol je procenjen na $-2,32102$, što znači da u proseku žene imaju vrednost varijable ishoda koja je za $2,32102$ jedinice niža od muškaraca, držeći sve ostale prediktore konstantnim.

Procenjena korelacija između dva fiksna efekta je $-0,813$, te kako starost raste, očekivana vrednost varijable ishoda ima tendenciju povećanja, dok očekivana vrednost varijable ishoda ima tendenciju da se smanji za žene u poređenju sa muškarcima.

Heteroskedastičnost ispitujemo pomoću grafika na kome vidimo da postoje neka odstupanja, ali je većina vrednosti koncentrisana u istom opsegu te možemo zaključiti da im je disperzija slična.



Prepostavku o normalnoj raspodeli proveravamo Q-Q plotom:



Vidimo da reziduali prate qqline, pa možemo prepostaviti da bi potencijalno mogla da bude normalna raspodela, ali zbog loše pozicije qqline bi možda trebalo sprovesti još neke analize.

7 Zaključak

Regresioni modeli za longitudinalne podatke nude moćan alat za analizu promena tokom vremena i postali su standardni alat za mnoge istraživače u širokom spektru oblasti. Pažljivim odabirom statističkih metoda i promišljenom analizom podataka, istraživači mogu otkriti važne odnose između prediktora i ishoda i pratiti promene u tim odnosima tokom vremena. U ovom radu raspravljali smo o longitudinalnim podacima, njihovim osobinama, te glavnim karakteristikama regresionih modela za longitudinalne podatke, uključujući linearne i generalizovane linearne modele, kao i naprednije metode kao što su modeli mešovitih efekata i modeli latentne krive rasta.

Takođe smo istakli neke od izazova i ograničenja regresionih modela za longitudinalne podatke, kao što su podaci koji nedostaju, greška merenja i pitanja vezana za uzročno zaključivanje. Uprkos ovim izazovima, regresioni modeli za longitudinalne podatke ostaju suštinski alat za razumevanje složenih procesa promene i razvoja u mnogim različitim oblastima, uključujuci psihologiju, medicinu i društvene nauke.

Kako polje regresionih modela za longitudinalne podatke nastavlja da se razvija, biće važno da istraživači ostanu pažljivi povodom pitanja specifikacije modela, validacije modela i odgovarajuće interpretacije rezultata. Štaviše, istraživači bi trebalo da nastave da istražuju nove metode i pristupe rukovanju složenim podacima, uključujući integraciju tehnika mašinskog učenja i naprednih metoda simulacije.

Verujemo da će regresioni modeli za longitudinalne podatke i dalje biti kritično sredstvo za unapređenje našeg razumevanja složenih fenomena koji oblikuju naše živote, kao i za informisanje politike i prakse u različitim domenima.

Литература

[1] Fitzmaurice, G. Laird, N. Ware, J. (2012).*Applied Longitudinal Analysis*. Wiley-Interscience, Hoboken, N.J., 2nd ed.

[2] Hedeker, D., Gibbons, R. D. (2006).*Longitudinal Data Analysis*. Wiley., 1st ed.

[3] Milošević, B. (2019). *Linearni statistički modeli.*, 1st ed.

[4] Long, J. D. (2012). *Longitudinal Data Analysis for the Behavioral Sciences Using R* . Sage Publications., 1st ed.

.