BILJANA Č. POPOVIĆ
PREDRAG M. POPOVIĆ

STATISTIČKO MODELIRANJE

Prvo izdanje

Serija: udžbenici

Univerzitet u Nišu, Prirodno–matematički fakultet
NIŠ, 2018

STATISTIČKO MODELIRANJE

Autori:
Dr Biljana Č. Popović, redovni profesor
Prirodno–matematičkog fakulteta Univerziteta u Nišu
Dr Predrag M. Popović, docent
Građevinsko–arhitektonskog fakulteta Univerziteta u Nišu
Recenzenti:
Dr Zagorka Lozanov–Crvenković, redovni profesor
Prirodno–matematichkog fakulteta Univerziteta u Novom Sadu
Dr Mila Stojaković, redovni profesor
Fakulteta tehnichkih nauka Univerziteta u Novom Sadu
Dr Miodrag Đorđević, docent
Prirodno–matematičkog fakulteta Univerziteta u Nišu
Tehničko uređivanje: Autori

Odlukom Nastavno–naučnog veća Prirodno–matematičkog fakulteta u Nišu, broj 1329/1–01 od 12.12.2018. godine, rukopis je odobren za štampu kao univerzitetski udžbenik.

Izdavač: Prirodno–matematički fakultet, Niš
Štampa: UNIGRAF–X–COPY, Niš
Tiraž: 120 primeraka

CIP-Каталогизација у публикацији
Народна библиотека Србије

519.2(075.8)

ПОПОВИЋ, Биљана Ч., 1954-

Statističko modeliranje / Biljana Č. Popović, Predrag M. Popović. - 1. izd. - Niš : Univerzitet, Prirodno-matematički fakultet, 2018 (Niš : Unigraf-X-Copy). - III, 171 str. : graf. prikazi, tabele ; 25 cm. - (Serija Udžbenici / [Prirodno-matematički fakultet, Niš])
Na nasl. str.: Univerzitet u Nišu. - Tiraž 120. - Napomene uz tekst. - Bibliografija: str. 165-167. - Registar.
ISBN 978-86-6275-085-3
1. Поповић, Предраг М., 1982-[аутор]
a) Математичка статистика
COBISS.SR–ID

Sadržaj

Predgovor

Knjiga je napisana sa namerom da predstavlja osnovnu literaturu, odnosno udžbenik, i kao takva je prevashodno namenjena studentima Osnovnih akademskih studija matematike Prirodno–matematičkog fakulteta Univerziteta u Nišu za predmet sa istoimenim nazivom, Statističko modeliranje. Međutim, knjigu mogu da koriste i svi zainteresovani za početne informacije iz oblasti statističkog modeliranja, kao i za početne informacije u vezi sa programskim jezikom *R*.

Da bi se ova knjiga uspešno koristila neophodno je predznanje koje obuhvata osnovne principe statističkog zaključivanja, ocenjivanje parametara i testiranje hipoteza, kao i osnove teorije uzoraka. Konkretno, studenti Osnovnih akademskih studija matematike Prirodno–matematičkog fakulteta Univerziteta u Nišu ova znanja stiču u okviru predmeta Matematička statistika koji na Fakultetu prethodi predmetu Statističko modeliranje.

Svi brojčani primeri i neke statistike su urađeni ili izračunati u programskom jeziku `R`. Međutim, iako su u Dodatku date osnove ovog programskog jezika, njegovo poznavanje nije neophodno, naprotiv, za korišćenje ove knjige.

Autori se unapred zahvaljuju svim čitaocima ove knjige koji im ukažu na greške i nedostatke.

Niš, oktobar 2018. godine Autori

Glava 1

Uvod

Kao i u drugim revolucijama vezanim za ljudsku misao, teško je odrediti tačan momenat kada je ideja statističkog modela postala deo nauke. Neki primeri statističkog modeliranja se mogu naći u radovima naučnika sa početka devetnaestog veka. Statističke tehnike i modeli se koriste danas u svim oblastima ljudskog zaključivanja. Na primer, u oblasti prodaje i potrošnje treba analizirati potrebe, ukus, platežnu sposobnost i sl. potrošača sa najčešćom namerom da se predvidi količina, kvalitet pa i cena proizvoda koji se nude na tržištu. U medicini, između ostalog, ispituju se efekti raznih lekova i kontrolišu uslovi okoline na ljude, sa namerom da se na pravi način tretiraju pojedine bolesti. Inženjeri uzorkuju karakteristike kvaliteta proizvoda i različite promenljive veličine koje se u posmatranom procesu mogu da kontrolišu sa namerom da otkriju ključne promenljive od kojih zavisi kvalitet proizvoda. Ekonomisti posmatraju različite indekse koji mogu da ilustruju ekonomsko zdravlje društva u određenom vremenskom periodu sa ciljem da predvide stanje ekonomije u budućnosti itd. Istina je u jednom, a to je da su ulagana i ulažu se ozbiljna novčana sredstva da bi se prikupljali statistički podaci i da bi se iz njih izvlačili odgovarajući zaključci. U tu svrhu se koriste ne samo osnovni principi statističkog zaključivanja, već se izgrađuju statistički modeli na osnovu kojih se vrši zaključivanje. Kako je zadatak statistike da izvrši zaključivanje o čitavoj populaciji na osnovu informacija koje su sadržane u uzorku uzetom iz te populacije, to se i poseže za različitim statističkim modelima koji će na osnovu određenih statističkih kriterijuma dobro da opisuju

stvarnost. Naravno, bitno je da se uoči da postoji razlika između modela, odnosno teorije, i stvarnosti. Teorija je samo ideja koja se predlaže za opisivanje realnog sveta i, kao takva je samo aproksimacija stvarnosti. No, to ne umanjuje vrednost ove teorije.

Nadalje ćemo koristiti termin *eksperiment* u najširem smislu, tj. njime ćemo označavati podatke dobijene posmatranjem, odnosno merenjem, u potpuno nekontrolisanim (slučajnim) uslovima, kao i podatke dobijene u laboratorijskim potpuno kontrolisanim uslovima. U istom smislu ćemo koristiti i termin *plan eksperimenta*. Da bismo opisali plan eksperimenta, koristićemo još termine kao što su *tretman*, *faktor*, *nivo faktora*, *kvantitativna merenja*, *kvalitativna merenja*, a koji će nadalje biti i objašnjeni.

Mnogi statistički podaci su zavisni od vremena, tj. moraju se uzimati i analizirati hronološki. U tom smislu se pojavljuje termin *slučajni proces*, a kao njegov specijalan slučaj i *slučajni niz*, ili poznatiji kao *vremenski niz* ili *vremenska serija*. Slučajni proces sa neprekidnim vremenom (odsečkom realne prave) i vremenski niz (čije je vreme ceo skup celih brojeva ili njegov pravi podskup) se analiziraju različitim matematičkim tehnikama, te se, kao takvi, moraju da razlikuju. Za ovakve podatke se takođe prave statistički modeli i neki od njih će biti razmatrani u poslednjoj glavi.

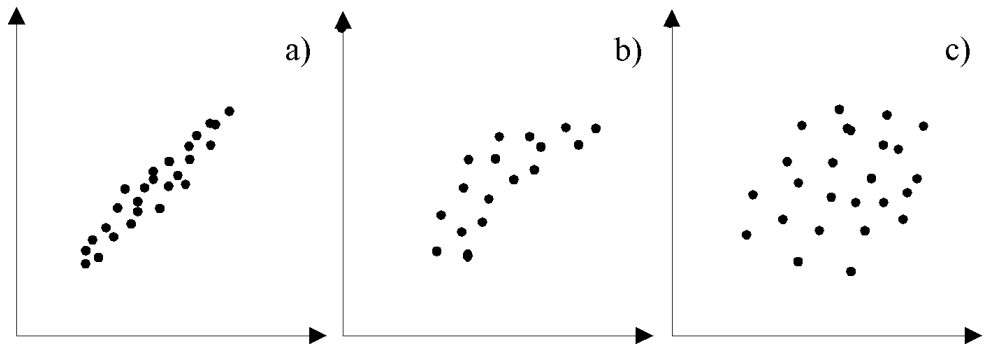
Glava 2

Regresija

Regresiona analiza je jedan od najstarijih statističkih postupaka. U velikom broju istraživanja ili eksperimenata uočava se veza između dve ili više promenljivih veličina. Od istraživača se u tom slučaju očekuje da utvrdi da li postoji i kakva je direktna funkcionalna zavisnost među tim veličinama. Na primeru dva svojstva X i Y koja se istražuju na nekom uzorku obima n, kao rezultat posmatranja dobija se n uređenih parova realizacija (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Oni se mogu predstaviti u Dekartovoj ravni (slika 2.1), a grafička reprezentacija koja tom prilikom nastaje naziva se **dijagram rasturanja**, odnosno, **dijagram rasipanja**.

Ukoliko se posmatra k obeležja na nekom uzorku obima n: X_1, X_2, ..., X_k, kao rezultat istraživanja javlja se n uređenih k-torki $(x_1^1, x_2^1, \ldots, x_k^1)$, $(x_1^2, x_2^2, \ldots, x_k^2)$, ..., $(x_1^n, x_2^n, \ldots, x_k^n)$, a dijagram rasturanja je skup od n tačaka k-dimenzionalnog euklidskog prostora E_k. Na osnovu tih podataka (tačaka) pokušava se da se otkrije funkcionalna veza među svojstvima, ako postoji. Opšti problem nalaženja funkcije koja dobro aproksimira dobijeni skup podataka, u statističkom žargonu se naziva "**fitovanje**[1] **krive**". Za određivanje odgovarajućeg tipa funkcionalne zavisnosti, u praksi se koristi upravo dijagram rasturanja. Radi ilustracije, to znači da za slučaj na slici 2.1 a) treba proveriti linearnu vezu $y = ax + b$, na istoj slici pod b) logaritamsku zavisnost tipa $y = a \ln(x + b)$, dok dijagram pod c) ne ukazuje ni na kakvu funkcionalnu zavisnost.

[1] od engleskog glagola *to fit* – upasovati, prilagoditi, podesiti



Slika 2.1: Dijagrami rasipanja tačaka.

Konstatujmo da se problem zavisnosti može posmatrati u dva pravca, koja će nadalje biti razjašnjena.

Posmatra se uticaj slučajnih veličina (obeležja) X_1, X_2, ..., X_p na slučajnu veličinu Y, tj. uticaj slučajnog vektora $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ na slučajnu promenljivu Y. Pri tome svaka vrednost slučajnog vektora $\mathbf{X}$ proizvodi odgovarajuću vrednost slučajne promenljive (obeležja) Y. U tom slučaju postoji očigledna potreba da se utvrdi, što preciznije, oblik zavisnosti ovih slučajnih veličina. Kako svaka realizovana vrednost $\mathbf{x}$ slučajnog vektora $\mathbf{X}$ proizvodi realizovanu vrednost y slučajne promenljive Y, zadatak se sastoji u nalaženju funkcije $\psi(\mathbf{x})$ koja "dobro" aproksimira vrednosti slučajne promenljive Y pri svakoj realizovanoj vrednosti slučajnog vektora $\mathbf{X}$. Funkcija $\psi(\mathbf{X})$ koja zadovoljava uslov da je

$$E(Y - \psi(\mathbf{X}))^2$$

minimalno, uzima se kao dobro predviđanje za Y na osnovu vektora $\mathbf{X}$. Najbolje predviđanje za Y po $\mathbf{X}$, u smislu definisanog srednjekvadratnog odstupanja, zove se **funkcija regresije** Y **na X.** Ovaj tip regresije je **regresija prve vrste.** Dakle, radi se o "slučajnom ulazu" i "slučajnom izlazu", tj. na slučajni ishod eksperimenta Y utiču samo slučajni faktori opisani vektorom $\mathbf{X}$.

Regresija druge vrste je drugi tip zavisnosti koji je takođe predmet statističkog proučavanja.

U većini eksperimentalnih istraživanja u laboratorijskim uslovima koncep-

cija je da se varira određeni broj neslučajnih veličina i posmatra njihov uticaj na ishod eksperimenta koji je slučajan. Neslučajne veličine o kojima je reč se nazivaju **kontrolisani faktori**. Ishod eksperimenta jeste slučajan, jer na posmatrano obeležje, osim kontrolisanih faktora, utiču po pravilu i slučajni faktori koji se ne mogu kontrolisati (na primer greške merenja), kao i neslučajni faktori koji su objektivno prisutni u eksperimentu, ali se njihov uticaj ne može sagledati ili pretpostaviti.

Slučajni ishod eksperimenta je obeležje Y kojim se eksperiment opisuje. Neslučajna ulazna promenljiva se označava odgovarajućim malim slovom, recimo x, pri čemu se različiti posmatrani nivoi ovog faktora označavaju donjim ili gornjim indeksima, tj. $x_1, x_2, \ldots, x_n$ ili $x^1, x^2, \ldots, x^n$. Ovde ćemo koristiti drugu oznaku, a donjim indeksom ćemo označavati postojanje više neslučajnih faktora čije se dejstvo na obeležje Y ispituje. Osim toga, gornji indeks ćemo stavljati u zagrade, da bi se vizuelno lakše razlikovao od stepena promenljive. Dakle, uticaj faktora x_1 na obeležje Y ispituje se na n nivoa: $x_1^{(1)}, x_1^{(2)}, \ldots, x_1^{(n)}$.

Slučajni uticaj na ishod eksperimenta koji se pri eksperimentu ne može da kontroliše, predstavićemo slučajnom promenljivom ε. Ovaj koncept dovodi do konstrukcije modela **regresije druge vrste**. Kada se ispituje uticaj samo jednog faktora, radi se o **jednostrukoj regresiji**, a kod ispitivanja istovremenog uticaja više faktora, reč je o **višestrukoj regresiji**. Kod ovog tipa regresije se pretpostavlja da se celokupan uticaj neslučajnih faktora sagledava kroz srednju vrednost obeležja Y, tj. da je slučajna komponenta aditivna i da joj je očekivanje nula.

Oba modela zavisnosti, regresija prve i regresija druge vrste, biće nadalje detaljnije razmatrana.

Naravno da u praksi ishod eksperimenta može da zavisi i od slučajnih i od kontrolisanih (neslučajnih) ulaza istovremeno i da su gore opisani modeli regresije samo idealne varijante jednog problema.

U daljoj analizi mi ćemo posmatrati samo uprošćene modele, tj. modele regresije prve i regresije druge vrste.

Primetimo da, koliko god funkcija regresije $\psi(\mathbf{X})$, odnosno $\psi(\mathbf{x})$, dobro aproksimira naše podatke, uvek postoji razlika između podatka koji smo izmerili i njegove aproksimacije funkcijom ψ. Ta razlika nosi naziv **rezidual** ili

ostatak.

Ako označimo sa ψ_i ocenu vrednosti y_i, $i = 1, \ldots, n$, na osnovu dobijenog modela regresije, kao mera valjanosti dobijene ocene često se koristi **koeficijent determinacije** ili preciznije, **uzorački koeficijent determinacije** koji se još naziva i **udeo objašnjene varijanse**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\psi_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{2.1}$$

gde je $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Koeficijent determinacije ima vrednosti $0 \leq R^2 \leq 1$ i što mu je vrednost bliža 1, to je bolja aproksimacija (fit). Međutim, nephodan je oprez prilikom korišćenja ove veličine za određivanje kvaliteta aproksimacije, jer ako u modelima linearne regresije o kojima će nadalje biti reči slobodan član (odsečak) bude jednak nuli, onda izraz u imeniocu ove veličine postaje nula pa je R^2 neopravdano veliki.

Kao ocena kvaliteta fita koristi se i disperzija, odnosno standardna devijacija ocene, o čemu će još biti reči.

2.1 Linearna regresija druge vrste

Posebno mesto među modelima regresije druge vrste imaju modeli linearne regresije kojima ćemo se nadalje baviti.

U opštem slučaju problem linearne regresije druge vrste polazi od pretpostavke da su matematička očekivanja opservacija Y_i , $i = 1, 2, \ldots, n$ linearne funkcije $\varphi_i(\boldsymbol{\beta})$ nepoznatih parametara $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$, koeficijenata regresije. Otuda i naziv linearna regresija. Slučajnu veličinu Y_i treba shvatiti kao ishod eksperimenta pri i–tom nivou posmatranih kontrolisanih faktora uticaja. Takođe se može da uvede pretpostavka o tome da posmatrani faktori utiču na ishod eksperimenta posredno, preko svojih funkcija $z_j = z_j(\mathbf{x})$, $j = 1, 2, \ldots, k$, $\mathbf{x} = (x_{j_1}, x_{j_2}, \ldots, x_{j_p})$, $j_l \in \{1, 2, \ldots, k\}$, $p \leq k$. Drugim rečima, u modelu linearne regresije koji je predmet proučavanja u ovom poglavlju, promenljive $z_1, \ldots, z_k$ mogu biti funkcionalno zavisne. Tako, sve promenljive z_j mogu biti funkcije samo od jednog kontrolisanog faktora x. Na primer, $z_j = x^j$, $j > 1$ daje model koji je poznat pod imenom paraboličkа linearna regresija. Ovaj i još neke primere ćemo kasnije detaljnije obrazložiti.

Nivoi faktora x_j će se tom prilikom ispoljiti kao "i"–ti nivo funkcije z_j i u modelu ćemo ga označavati sa $z_j^{(i)}$. Nadalje ćemo razmatrati linearne funkcije od $\boldsymbol{\beta}$ oblika $\mathbf{z}^{(i)'}\boldsymbol{\beta}$, $\mathbf{z}^{(i)} = (z_1^{(i)}, \ldots, z_k^{(i)})'$, $i = 1, \ldots, n$.

Pretpostavka koja se uvodi u model je, kao što smo rekli, da se celokupan uticaj neslučajnih faktora ostvaruje preko matematičkog očekivanja obeležja Y, tj. da u modelu

$$Y_i = \mathbf{z}^{(i)'}\boldsymbol{\beta} + \varepsilon_i\,, \quad i = 1, 2, \ldots, n\,, \tag{2.2}$$

za očekivane vrednosti zavisno promenljive i ostatka, $EY_i = \mathbf{z}^{(i)'}\boldsymbol{\beta}$ i $E\varepsilon_i = 0$, za svako i, i da raspodele ostataka ("grešaka") ε_i ne zavise od $\boldsymbol{\beta}$.

Planiranje eksperimenta u ovom slučaju podrazumeva uvođenje matrice plana $\mathbf{Z} = ||\mathbf{z}^{(1)} \ldots \mathbf{z}^{(n)}||$ dimenzije $k \times n$, $k < n$ i vektora grešaka $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ pa (2.2) dobija oblik

$$\mathbf{Y} = \mathbf{Z}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad , \quad E(\boldsymbol{\varepsilon}) = \mathbf{0}\,, \tag{2.3}$$

gde je $\mathbf{Y}$ vektor kolone slučajnih ishoda eksperimenta, $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.

U model se obično uvodi i pretpostavka da su komponente vektora $\boldsymbol{\varepsilon}$ nekorelirane među sobom i da imaju iste disperzije, što znači

$$D(Y_i) = D(\varepsilon_i) = \sigma^2 \quad , \quad i = 1, \ldots, n \quad \text{i} \quad Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \text{za} \quad i \neq j\,.$$

U tom slučaju je kovarijansna (autokovarijansna) matrica vektora $\mathbf{Y}$

$$\mathbf{D}(\mathbf{Y}) = \mathbf{D}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n\,, \tag{2.4}$$

gde je $\mathbf{I}_n$ jedinična matrica reda n. Ukoliko bi izostao uslov nekoreliranosti, autokovarijansna matrica vektora $\boldsymbol{\varepsilon}$ bi bila oblika

$$\mathbf{D}(\mathbf{Y}) = \mathbf{D}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{G}\,,$$

gde je matrica $\mathbf{G}$ simetrična pozitivno semidefinitna. Međutim, slučaj nesingularne matrice $\mathbf{G}$ bi omogućio transformaciju vektora $\mathbf{Y}$ u vektor $\mathbf{W} = \mathbf{G}^{-1/2}\mathbf{Y}$ čija bi autokovarijansna matrica bila ponovo dijagonalna (2.4). Slučaj singularne matrice $\mathbf{G}$ je, takođe moguć, ali ga ovde nećemo razmatrati.

U modelu linearne regresije druge vrste važnu ulogu ima matrica

$$\mathbf{A} = \mathbf{Z}\mathbf{Z}'. \tag{2.5}$$

Ova matrica je kvadratna simetrična matrica reda k i za nju važi sledeće tvrđenje.

Teorema 2.1. *Matrica* $\mathbf{A}$ *je pozitivno semidefinitna, a uslov* $rang\mathbf{Z} = k$ *je potreban i dovoljan da ona bude pozitivno definitna.*

Dokaz. Neka je $\mathbf{t} = (t_1, \ldots, t_k)'$ proizvoljan nenula vektor iz istog polja brojeva iz koga su elementi matrice $\mathbf{Z}$. Tada je

$$\mathbf{t}'\mathbf{A}\mathbf{t} = (\mathbf{Z}'\mathbf{t})'(\mathbf{Z}'\mathbf{t}) \geq 0\,,$$

kao kvadratna forma, pri čemu važi jednakost ako i samo ako je $\mathbf{Z}'\mathbf{t} = 0$, odnosno, ako i samo ako je $\sum_{j=1}^{k} t_j z_j^{(i)} = 0$ za svako $i = 1, \ldots, n$. Poslednji uslov je ekvivalentan sa uslovom

$$\sum_{j=1}^{k} t_j \mathbf{z}_j = \mathbf{0}\,, \tag{2.6}$$

gde je $\mathbf{z}_j$ vektor vrste matrice $\mathbf{Z}$. Poslednji iskaz izjednačava linearnu kombinaciju vektora vrsta matrice $\mathbf{Z}$ sa nula vektorom, što je svojstvo linearno zavisnih vrsta. Uslov (2.6) ekvivalentan je sa $rang\mathbf{Z} < k$. □

Osnovni zadatak statističkog postupka modela linearne regresije je ocenjivanje parametara regresionog modela. Pri tome se osim koeficijenata regresije, elemenata vektora $\boldsymbol{\beta}$, kao nepoznat parametar često javlja i disperzija σ^2. Za rešavanje ovog zadatka koristi se metod najmanjih kvadrata. Ovaj metod uveo je Gaus još 1809. godine.

2.1.1 Metod najmanjih kvadrata za ocenjivanje parametara modela linearne regresije

Primenićemo najpre metod najmanjih kvadrata na ocenjivanje vektora $\boldsymbol{\beta}$.

DEFINICIJA 2.1. Ocena nepoznatog vektora $\boldsymbol{\beta}$ dobijena metodom najmanjih kvadrata je vektor statistika $\widehat{\boldsymbol{\beta}}$ koje minimalizuju kvadratnu formu

$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta})\,, \tag{2.7}$$

dakle, vektor $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_k)'$ koji zadovoljava relaciju

$$S(\widehat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})\,.\diamondsuit$$

Formula (2.7) predstavlja sumu kvadrata razlika slučajnih rezultata eksperimenta i njihovih matematičkih očekivanja.

Uobičajenim postupkom za određivanje minimuma funkcije koja je diferencijabilna:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_j} = 0\,, \quad j = 1, \ldots, k\,, \tag{2.8}$$

dobijamo sledeći sistem linearnih jednačina po nepoznatim parametrima β_j, $j = 1, \ldots, k$

$$\sum_{i=1}^{n} z_j^{(i)} \left(\sum_{l=1}^{k} z_l^{(i)} \beta_l - Y_i \right) = 0 \quad , \quad j = 1, \ldots, k\,.$$

Koristeći matrični zapis i oznaku $\mathbf{ZY} = \mathbf{V}$, poslednji sistem se može zapisati kao

$$\mathbf{A}\boldsymbol{\beta} = \mathbf{V}\,. \tag{2.9}$$

Jednačina (2.9) nosi naziv **normalna jednačina metoda najmanjih kvadrata**, odnosno sistem linearnih jednačina koji se njome definiše nosi naziv **normalni sistem jednačina metoda najmanjih kvadrata**. O važnosti ovog sistema govori sledeća teorema.

Teorema 2.2. *Neka je $\boldsymbol{\beta}^*$ proizvoljno rešenje normalne jednačine (2.9). Tada je*

$$S(\widehat{\boldsymbol{\beta}}) = S(\boldsymbol{\beta}^*) = \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$$

i minimum je isti za proizvoljno (svako) rešenje sistema (2.9).

Ako je $det\mathbf{A} \neq 0$*, tada je ocena najmanjih kvadrata jedinstvena i jednaka*

$$\widehat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{V}\,.$$

Dokaz. Neka je $\boldsymbol{\beta}^*$ proizvoljno fiksirano rešenje jednačine (2.9). Tada je

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{Y}-\mathbf{Z}'\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{Z}'\boldsymbol{\beta}) = \\ &= (\mathbf{Y}-\mathbf{Z}'\boldsymbol{\beta}^*+\mathbf{Z}'\boldsymbol{\beta}^*-\mathbf{Z}'\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{Z}'\boldsymbol{\beta}^*+\mathbf{Z}'\boldsymbol{\beta}^*-\mathbf{Z}'\boldsymbol{\beta}) = \\ &= S(\boldsymbol{\beta}^*)+(\mathbf{V}-\mathbf{A}\boldsymbol{\beta}^*)'(\boldsymbol{\beta}^*-\boldsymbol{\beta})+(\boldsymbol{\beta}^*-\boldsymbol{\beta})'(\mathbf{V}-\mathbf{A}\boldsymbol{\beta}^*) \\ &+ (\boldsymbol{\beta}^*-\boldsymbol{\beta})'\mathbf{A}(\boldsymbol{\beta}^*-\boldsymbol{\beta}) = \\ &= S(\boldsymbol{\beta}^*)+(\boldsymbol{\beta}^*-\boldsymbol{\beta})'\mathbf{A}(\boldsymbol{\beta}^*-\boldsymbol{\beta}) \geq S(\boldsymbol{\beta}^*)\,, \end{aligned}$$

što je i trebalo dokazati. Rezultat sledi jer je matrica $\mathbf{A}$ pozitivno semidefinitna.

Za nesingularnu matricu $\mathbf{A}$ rešenje normalne jednačine je jedinstveno, pa je i ocena najmanjih kvadrata jedinstvena i ima oblik

$$\widehat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{V}\,. \quad \square$$

U mnogim praktičnim problemima primene od interesa su ne direktno koeficijenti regresije, već njihove linearne kombinacije. Reč je o statistikama za različite namene, a koje su funkcije od ocena koeficijenata regresije. Iz tog razloga ćemo nadalje razmatrati vektor linearnih kombinacija koeficijenata regresije $\mathbf{t} = \mathbf{T}\boldsymbol{\beta}$, gde je $\mathbf{T}$ zadata matrica reda $m \times k\,,\ m \leq k$. Ovo bi, na primer, bio slučaj kada se parametarski prostor R^k za parametre koji predstavljaju koeficijente linearne regresije sužava nekim linearnim ograničenjima, tj. sistemom od m linearnih ograničenja, sadržanih u vektoru $\mathbf{t}$ za neko $\mathbf{t} = \mathbf{t}_0$. U tom slučaju, kao ocenu najmanjih kvadrata vektora $\mathbf{t}$, u oznaci $\widehat{\mathbf{t}}$, imaćemo vektor statistika

$$\widehat{\mathbf{t}} = \mathbf{T}\widehat{\boldsymbol{\beta}}\,,$$

gde je, kao i do sada, $\widehat{\boldsymbol{\beta}}$ proizvoljno rešenje normalne jednačine. Jasno, ako je $det\mathbf{A} \neq 0$, $\widehat{\mathbf{t}}$ je jednoznačno određen, tj.

$$\widehat{\mathbf{t}} = \mathbf{T}\mathbf{A}^{-1}\mathbf{V}\,. \tag{2.10}$$

Nadalje ćemo razmatrati svojstva ocena dobijenih metodom najmanjih kvadrata. Razmatraćemo zadatak ocenjivanja vektora $\mathbf{t}$ samo u klasi linearnih ocena, tj. razmatraćemo ocene oblika

$$\mathbf{l} = \mathbf{LY}$$

koje su statistike od posmatranja (vektora uzorka) $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.

Teorema 2.3. *Neka je matrica* $\mathbf{A}$ *nesingularna. Tada, za proizvoljan vektor* $\mathbf{t} = \mathbf{T}\boldsymbol{\beta}$*, ocena najmanjih kvadrata definisana relacijom (2.10) je nepristrasna ocena sa minimalnom disperzijom u odnosu na sve ostale linearne nepristrasne ocene za* $\mathbf{t}$*. Pri tome je kovarijansna matrica vektora* $\widehat{\mathbf{t}}$ *data sa*

$$\mathbf{D}(\widehat{\mathbf{t}}) = \sigma^2 \mathbf{T}\mathbf{A}^{-1}\mathbf{T}',$$

pod uslovom da je disperzija ostataka regresionog modela, σ^2*, poznata.*

Dokaz. Nepristrasnost ocene $\widehat{\mathbf{t}}$ sledi direktno iz njene definicije i osobina matematičkog očekivanja.

Uporedimo sada proizvoljnu drugu linearnu nepristrasnu ocenu $\mathbf{l} = \mathbf{LY}$ vektora $\mathbf{t}$ sa ocenom $\widehat{\mathbf{t}}$. S obzirom na to da je $\mathbf{l}$ nepristrasna ocena, važi $E(\mathbf{l}) = \mathbf{t}$, odnosno $E(\mathbf{l}) = \mathbf{T}\boldsymbol{\beta}$. S druge strane,

$$E(\mathbf{l}) = \mathbf{L}E(\mathbf{Y}) = \mathbf{L}\mathbf{Z}'\boldsymbol{\beta}.$$

Otuda $\mathbf{L}\mathbf{Z}'\boldsymbol{\beta} = \mathbf{T}\boldsymbol{\beta}$. Poslednja jednakost mora da bude tačna za svaki vektor $\boldsymbol{\beta}$, pa odatle sledi da je $\mathbf{L}\mathbf{Z}' = \mathbf{T}$. Ostaje još da dokažemo da je u pitanju ocena sa najmanjom disperzijom među svim linearnim nepristrasnim ocenama.

Disperzija ocene $\mathbf{l}$ je

$$\mathbf{D}(\mathbf{l}) = \mathbf{D}(\mathbf{LY}) = \mathbf{L}\mathbf{D}(\mathbf{Y})\mathbf{L}' = \mathbf{L}\sigma^2\mathbf{I}_n\mathbf{L}' = \sigma^2\mathbf{L}\mathbf{L}'.$$

Minimum disperzija ocena $l_1, l_2, \ldots, l_m$ komponenata vektora $\mathbf{l}$ će se ostvariti ukoliko su dijagonalni elementi glavne dijagonale kovarijansne matrice $\mathbf{D}(\mathbf{l})$ minimalni mogući. To znači da je potrebno minimalizovati elemente glavne

dijagonale matrice $\mathbf{LL}'$. S obzirom na to da za ovu matricu važi razlaganje

$$\begin{aligned}\mathbf{LL}' &= \mathbf{TA}^{-1}\mathbf{T}' + \mathbf{LL}' - \mathbf{TA}^{-1}\mathbf{T}' - \mathbf{TA}^{-1}\mathbf{T}' + \mathbf{TA}^{-1}\mathbf{T}' = \\ &= \mathbf{TA}^{-1}\mathbf{AA}^{-1}\mathbf{T}' + \mathbf{LL}' - \mathbf{LZ}'\mathbf{A}^{-1}\mathbf{T}' - \mathbf{TA}^{-1}\mathbf{ZL}' + \mathbf{TA}^{-1}\mathbf{AA}^{-1}\mathbf{T}' = \\ &= (\mathbf{TA}^{-1}\mathbf{Z})(\mathbf{TA}^{-1}\mathbf{Z})' + (\mathbf{L} - \mathbf{TA}^{-1}\mathbf{Z})(\mathbf{L} - \mathbf{TA}^{-1}\mathbf{Z})'\,,\end{aligned}$$

to će njeni dijagonalni elementi dostići minimum ako i samo ako je drugi sabirak ovog razlaganja jednak nuli, tj. za

$$\mathbf{L} = \mathbf{TA}^{-1}\mathbf{Z}\,,$$

što je i trebalo dokazati. Zaista, množeći i levu i desnu stranu ove jednakosti sa desna sa $\mathbf{Y}$, dobija se $\mathbf{l} = \widehat{\mathbf{t}}$. □

Posledica 2.1. *Kovarijansna matrica ocene najmanjih kvadrata vektora $\boldsymbol{\beta}$ u slučaju nesingularne matrice $\mathbf{A}$ je oblika*

$$\mathbf{D}(\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{A}^{-1}.$$

Dokaz. Pođimo od kovarijansne matrice vektora $\widehat{\mathbf{t}}$

$$\mathbf{D}(\widehat{\mathbf{t}}) = \mathbf{D}(\mathbf{T}\widehat{\boldsymbol{\beta}}) = \mathbf{T}\mathbf{D}(\widehat{\boldsymbol{\beta}})\mathbf{T}'\,.$$

No, prema prethodnoj teoremi je

$$\mathbf{D}(\widehat{\mathbf{t}}) = \sigma^2\mathbf{TA}^{-1}\mathbf{T}'\,,$$

pa je

$$\sigma^2\mathbf{TA}^{-1}\mathbf{T}' = \mathbf{T}\mathbf{D}(\widehat{\boldsymbol{\beta}})\mathbf{T}'\,,$$

odnosno

$$\mathbf{D}(\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{A}^{-1}\,.\,\square$$

Dakle, poslednja teorema nam daje optimalnu linearnu ocenu u smislu ocene najmanjih kvadrata za proizvoljnu linearnu kombinaciju koeficijenata linearne regresije.

Preostalo je još da odredimo ocenu najmanjih kvadrata za disperziju os-

tataka regresionog modela u slučaju kada je ova nepoznata.

Teorema 2.4. *Nepristrasna ocena disperzije σ^2 dobijena po metodu najmanjih kvadrata, kada je matrica $\mathbf{A}$ nesingularna, je statistika*

$$\widetilde{\sigma}^2 = \frac{1}{n-k} S(\widehat{\boldsymbol{\beta}}) = \frac{1}{n-k} (\mathbf{Y} - \mathbf{Z}'\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}'\widehat{\boldsymbol{\beta}}) \,,$$

gde je, kao i do sada, $\widehat{\boldsymbol{\beta}}$ ocena najmanjih kvadrata vektora $\boldsymbol{\beta}$, odnosno proizvoljno rešenje normalne jednačine.

Dokaz. Označimo sa a_{ij} elemente matrice $\mathbf{A}$, a sa a^{ij} elemente matrice $\mathbf{A}^{-1}$.

Kako je

$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}'\boldsymbol{\beta}) = (\mathbf{Y} - E(\mathbf{Y}))'(\mathbf{Y} - E(\mathbf{Y})) \,,$$

to je

$$E(S(\boldsymbol{\beta})) = \sum_{i=1}^{n} \left(E(Y_i^2) - (E(Y_i))^2 \right) = n\sigma^2 \,.$$

Osim toga je

$$\begin{aligned} E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right) &= \sum_{j=1}^{k}\sum_{i=1}^{k} a_{ji} E\left((\widehat{\beta}_i - \beta_i)(\widehat{\beta}_j - \beta_j) \right) = \\ &= \sigma^2 \sum_{j=1}^{k}\sum_{i=1}^{k} a_{ji} a^{ij} = \sigma^2 tr(\mathbf{A}\mathbf{A}^{-1}) = \\ &= \sigma^2 tr(\mathbf{I}_k) = k\sigma^2 \,. \end{aligned}$$

Koristeći sada rezultat

$$S(\boldsymbol{\beta}) = S(\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \,,$$

dobija se da je

$$E(S(\widehat{\boldsymbol{\beta}})) = E(S(\boldsymbol{\beta})) - E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right) = n\sigma^2 - k\sigma^2 = (n-k)\sigma^2 \,,$$

odakle sledi tvrđenje teoreme.□

Primer 2.1. Ilustrujmo prethodnu teoriju na primeru regresije sa jednim kon-

trolisanim faktorom, tj. na primeru jednostruke regresije. Dakle, neka je u pitanju kontrolisani faktor x i njegov uticaj na slučajni ishod eksperimenta Y. Model linearne regresije podrazumeva određivanje najbolje linearne zavisnosti oblika

$$y = \beta_1 + \beta_2 x$$

na osnovu matrice plana reda $2 \times n$, čiji su vektori kolona $\mathbf{z}^{(i)} = (1, x^{(i)})'$, $i = 1, \ldots, n$. Uočimo da je u ovom slučaju $z(x) = x$, tj. da će rezultat eksperimenta biti registrovan u obliku $(x^{(i)}, y_i)$, $i = 1, \ldots, n$. Prema tome,

$$EY_i = \beta_1 + \beta_2 x^{(i)} \quad , \quad i = 1 \ldots, n\,.$$

Uobičajeno je da se koeficijent β_1 zove odsečak, a β_2 koeficijent pravca, zbog geometrijske interpretacije modela i naziva za ove koeficijente koji se koriste u analitičkoj geometriji.

Prateći dalje teoriju, treba definisati sve potrebne matrice i vektore:

$$\mathbf{Z} = \left\| \begin{array}{cccc} 1 & 1 & \cdots & 1 \\ x^{(1)} & x^{(2)} & \cdots & x^{(n)} \end{array} \right\|, \qquad \mathbf{A} = \left\| \begin{array}{cc} n & \sum_{i=1}^n x^{(i)} \\ \sum_{i=1}^n x^{(i)} & \sum_{i=1}^n {x^{(i)}}^2 \end{array} \right\| \quad \text{i}$$

$$\mathbf{V} = \left\| \begin{array}{c} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x^{(i)} Y_i \end{array} \right\| .$$

Pretpostavimo da je eksperiment sproveden bar na dva različita nivoa kontrolisanog faktora x. To bi za posledicu imalo da je rang matrice plana 2, tj. $rang\mathbf{Z} = 2$, a samim tim bi značilo da je matrica $\mathbf{A}$ regularna, odnosno nesingularna. Otuda

$$det\mathbf{A} = n \sum_{i=1}^{n} (x^{(i)} - \overline{x}_n)^2 \; > \; 0 \quad , \quad \overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}\,,$$

a njena inverzna matrica je

$$\mathbf{A}^{-1} = \frac{1}{det\mathbf{A}} \left\| \begin{array}{cc} \sum_{i=1}^n {x^{(i)}}^2 & -n\overline{x}_n \\ -n\overline{x}_n & n \end{array} \right\|,$$

što znači da postoji jedinstveno rešenje sistema normalnih jednačina i ono je dato vektorom statistika

$$\widehat{\boldsymbol{\beta}} = \left\| \begin{array}{c} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{array} \right\| = \left\| \begin{array}{c} \dfrac{\sum_{i=1}^n x^{(i)^2} \overline{Y}_n - \overline{x}_n \sum_{i=1}^n x^{(i)} Y_i}{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2} \\ \dfrac{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2} \end{array} \right\| \quad , \qquad \overline{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i \, .$$

Uočimo da su statistike $\widehat{\beta}_1$ i $\widehat{\beta}_2$ u sledećoj relaciji

$$\widehat{\beta}_1 = \overline{Y}_n - \overline{x}_n \widehat{\beta}_2 \, .$$

Disperziona matrica vektora $\widehat{\boldsymbol{\beta}}$ je

$$\mathbf{D}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{A}^{-1} \, ,$$

što znači da su disperzije komponenata

$$D(\widehat{\beta}_1) = \frac{\sum_{i=1}^n x^{(i)^2}}{n \sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2} \sigma^2 \qquad \text{i} \qquad D(\widehat{\beta}_2) = \frac{1}{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2} \sigma^2 \, ,$$

a njihova kovarijansa

$$Cov(\widehat{\beta}_1, \widehat{\beta}_2) = \frac{\overline{x}_n}{\sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2} \sigma^2 \, .$$

Ukoliko je disperzija grešaka, σ^2, nepoznata, treba i nju oceniti. Njena nepristrasna ocena data je relacijom

$$\widetilde{\sigma}^2 = \frac{S(\widehat{\boldsymbol{\beta}})}{n-2} \quad , \quad S(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^n (Y_i - \overline{Y}_n)^2 - \widehat{\beta}_2^2 \sum_{i=1}^n (x^{(i)} - \overline{x}_n)^2 \, . \triangle$$

Primer 2.2. Koristeći rezultat prethodnog primera, naći regresionu pravu kojom se može predvideti količina soli (Y) natrijumnitrata $NaNO_3$ koju je moguće rastvoriti u $100gr$ vode u zavisnosti od temperature (x) na osnovu sledećih eksperimentalnih podataka:

$x^{(i)}$	0	4	10	15	21	29	36	51	68
y_i	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1

i oceniti disperziju grešaka merenja.

Rešenje se dobija rešavanjem sistema jednačina $\mathbf{A}\boldsymbol{\beta} = \mathbf{V}$, tj.

$$\begin{aligned} 9\beta_1 \ + \ 234\beta_2 \ &= \ 811,3 \\ 234\beta_1 + 10144\beta_2 &= 24628,6 \end{aligned}$$

i ocene koeficijenata regresije su

$$\beta_1 = 67,5 \quad \text{i} \quad \beta_2 = 0,87\,.$$

Dakle, jednačina linearne regresije je

$$y = 67,5 + 0,87x\,,$$

a ocena disperzije slučajnih grešaka merenja je $\widetilde{\sigma}^2 = 0,92$.

Linearnu regresiju možemo odrediti koristeći programski jezik `R` i funkciju `lm` u njemu. Učitajmo najpre podatke iz tabele (pod pretpostavkom da se oni nalaze u nekom .csv fajlu)

```
> tabela = read.csv(file = "primer.csv",header = TRUE)
```

Dobili smo objekat `tabela` koji sadrži dve kolone, "So" i "Temp". Kako bismo stekli utisak da li je linearan model adekvatan da modelira posmatrani skup podataka, predstavićemo grafički posmatrane podatke. Pozivom funkcije `scatter.smooth` dobićemo grafik prikazan na slici 2.2.

```
> scatter.smooth(x=tabela$Temp, y=tabela$So, xlab = "Temperatura", ylab
= "So")
```

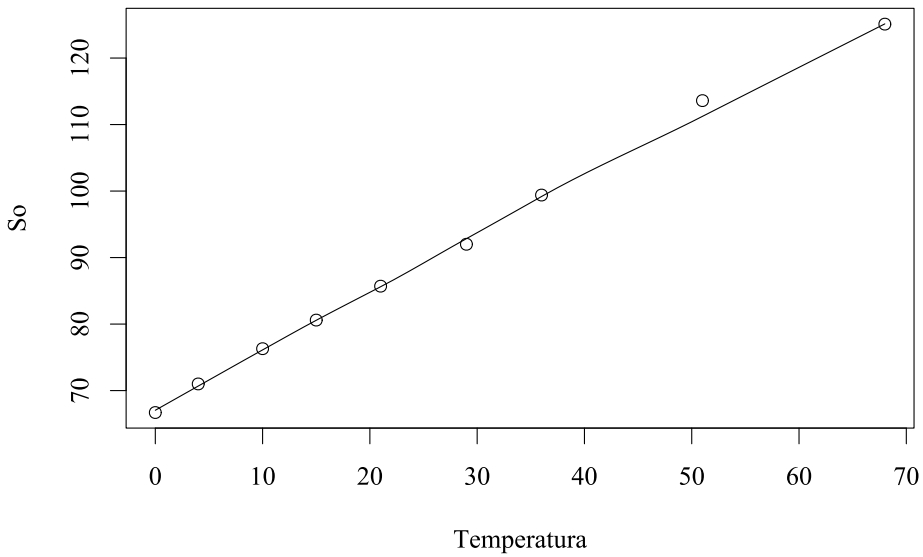
Možemo zaključiti da je zavisnosti temperature i količine rastvorene soli linearna pa nastavljamo sa određivanjem parametara modela.

Koristeći funkciju `lm` napravićemo objekat `lm.fit` koji sadrži sve potrebne informacije vezane za naš regresioni model

```
> lm.fit = lm(table$So~table$Temp)
```

Iskoristićemo funkciju `summary` kako bismo dobili osnovne informacije o modelu.

```
> summary(lm.fit)
...
```



Slika 2.2: Količina soli koju je moguće rastvoriti u zavisnosti od temperature vode.

```
Coefficients:
               Estimate  Std.  Error  t value    Pr(>|t|)
(Intercept)    67.42508    0.52320    128.87   4.47e-13***
tabela$Temp     0.86998    0.01558     55.82   1.55e-10***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Residual standard error:  0.993 on 7 degrees of freedom
Multiple R-squared:  0.9978, Adjusted R-squared:  0.9974
F-statistic:  3116 on 1 and 7 DF, p-value:  1.552e-10
```

Iz kolone `Estimate` pročitaćemo vrednosti za koeficijente našeg regresionog modela. Treba obratiti pažnju na kolonu `Pr(>|t|)` gde su nam date p-vrednosti testa za hipotezu $H_0 : \beta_i = 0, i = 1, 2$. Možemo zaključiti da se nulta hipoteza odbacuje kod oba koeficijenta regresionog modela.

Simulirane vrednosti regresionog modela dobijamo iz `lm.fit` objekta kao

```
> lm.fit$fitted.values
```

dok greške predviđanja dobijamo sa

```
> lm.fit$residuals
```

Koeficijent determinacije dat jednačinom (2.1) izračunaćemo pozivom

```
> 1-sum(lm.fit$residuals^2)/sum((tabela$So-mean(tabela$So))^2)
[1] 0.9977588
```

odakle zaključujemo da model sasvim dobro opisuje podatke iz posmatranog uzorka. Za dalje ispitivanje adekvatnosti modela možemo da iskoristimo funkcije `AIC(lm.fit)` i `BIC(lm.fit)` koje daju vrednosti za Akaike i Bajesov informacioni kriterijum (videti Dodatak) posmatranog modela, respektivno. △

Naglasimo da se prilikom korišćenja programskog jezika `R`, odnosno pozivom kodova iz ovog jezika, javlja tačka na mestu decimalnog zareza, jer je programski jezik prilagođen engleskom govornom području po čijem se pravopisu decimalna mesta odvajaju tačkom, a ne zarezom!

Primer 2.3. Razmotrimo sada najjednostavniji primer linearne paraboličke regresije. Posmatraćemo ponovo uticaj samo jednog kontrolisanog faktora x na slučajni ishod eksperimenta Y i pretpostaviti da se radi o modelu kod koga je

$$EY = \beta_0 + \beta_1 x + \beta_2 x^2 \,,$$

gde je sa x^2 označen kvadrat jedinog faktora x. Indekse za koeficijente β smo, iz tradicionalnih razloga vezanih za označavanje koeficijenata polinoma, označili počev od 0.

Sistem normalnih jednačina je, u ovom slučaju,

$$\begin{aligned} \beta_0 n \quad &+ \beta_1 \textstyle\sum x^{(i)} &+ \beta_2 \textstyle\sum {x^{(i)}}^2 &= \quad \textstyle\sum y_i \\ \beta_0 \textstyle\sum x^{(i)} &+ \beta_1 \textstyle\sum {x^{(i)}}^2 &+ \beta_2 \textstyle\sum {x^{(i)}}^3 &= \textstyle\sum x^{(i)} y_i \\ \beta_0 \textstyle\sum {x^{(i)}}^2 &+ \beta_1 \textstyle\sum {x^{(i)}}^3 &+ \beta_2 \textstyle\sum {x^{(i)}}^4 &= \textstyle\sum {x^{(i)}}^2 y_i \end{aligned} \,.$$

Koeficijente modela dobićemo iz programskog jezika `R` pozivom funkcije `lm` gde ćemo regresioni model definisati na sledeći način

```
> lm.fit2=lm(formula = table$So~table$Temp + I(table$Temp^2))
```

Analogno, možemo definisati linearni model koristeći polinom k-tog stepena po kontrolisanoj promenljivoj. $\triangle$

Istaknimo da će regresija uvek nositi naziv linearna regresija kada se radi o regresiji koja je linearna po svojim koeficijentima, odnosno po elementima vektora $\boldsymbol{\beta}$.

Primer 2.4. Navodimo još nekoliko modela regresije druge vrste koji ne spadaju u linearne, međutim jednostavnim transformacijama se mogu svesti na model linearne regresije i optimizovati metodom najmanjih kvadrata.

Na linearni model mogu se svesti, na primer, sledeći modeli jednostruke regresije

- $Y = \beta_1 x^{\beta_2} + \varepsilon$,
- $Y = \frac{1}{\beta_1 + \beta_2 x} + \varepsilon$,
- $Y = e^{\beta_1 + \beta_2 x} + \varepsilon$.

Označimo $EY = \overline{y}$, pa s obzirom na pretpostavku o očekivanju obeležja Y, posmatrajmo transformacije redom

- $\ln \overline{y} = v$, $\ln \beta_1 = b_1$, $\ln x = u$,
- $\frac{1}{\overline{y}} = v$,
- $\ln \overline{y} = v$

kojim dobijamo linearne modele

- $v = b_1 + \beta_2 u$,
- $v = \beta_1 + \beta_2 x$,
- $v = \beta_1 + \beta_2 x$.

Kako bismo u programskom jeziku `R` iskoristili funkciju `lm` za određivanje parametara modela, potrebno je prvo izvršiti pomenute transformacije obeležja, a zatim odrediti model kao

- `lm(formula = v~u)`

- `lm(formula = v∼x)`

- `lm(formula = v∼x)` △

Primer 2.5. Najjednostavniji model višestruke regresije je model sa dva različita kontrolisana faktora čija je matrica plana

$$\mathbf{Z} = \left\| \begin{array}{cccc} 1 & 1 & \cdots & 1 \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(n)} \\ x_3^{(1)} & x_3^{(2)} & \cdots & x_3^{(n)} \end{array} \right\| .$$

Odgovarajući sistem normalnih jednačina za ovaj model je

$$\begin{array}{rcrcrcl} \beta_1 n & + & \beta_2 \sum x_2^{(i)} & + & \beta_3 \sum x_3^{(i)} & = & \sum y_i \\ \beta_1 \sum x_2^{(i)} & + & \beta_2 \sum {x_2^{(i)}}^2 & + & \beta_3 \sum x_2^{(i)} x_3^{(i)} & = & \sum x_2^{(i)} y_i \\ \beta_1 \sum x_3^{(i)} & + & \beta_2 \sum x_2^{(i)} x_3^{(i)} & + & \beta_3 \sum {x_3^{(i)}}^2 & = & \sum x_3^{(i)} y_i \end{array} . \tag{2.11}$$

Model višestruke regresije dobićemo u programskom jeziku `R` pomoću funkcije `lm` koja bi bila pozvana kao

```
> lm.fit2=lm(formula = y ∼ x2 + x3)
```

i analogno, kada imamo više od dve nezavisne promenljive. Ako bismo želeli da izbegnemo slobodan član (tj. β_1), funkciju bismo pozvali kao `lm(formula = y ∼ 0 + x2 + x3)`. △

Primer 2.6. Pri proizvodnji žica od čelika važno je da li kvalitet žice odgovara propisanom kvalitetu. U tu svrhu se, između ostalog, vrši ispitivanje zatezne čvrstoće žice u zavisnosti od njenog spoljašnjeg prečnika i količine molibdena u čeliku od koga je žica izrađena. Uzeta su četiri parčeta žice, sa različitih koturova, iste dužine i izvršena su merenja. Rezultati tih merenja dati su tabelom

x_2, spoljašnji prečnik u *cm*	3	2	4	3	5	4	3	4	6	2	3
x_3, količina molibdena u *mol*	5	5	8	7	7	5	6	8	8	6	8
y, zatezna čvrstoća	8	8	16	12	14	10	11	12	17	9	12

.

Izračunati koeficijente višestruke regresije, a zatim proceniti kakva će biti čvrstoća žice spoljašnjeg prečnika 3,5cm sa količinom molibdena 6,4 mola. Izračunati koeficijent determinacije.

Rešavanjem sistema (2.11), gde su dva data kontrolisana faktora spoljašnji prečnik žice i količina molibdena u leguri, dobijaju se koeficijenti regresionog modela $\beta_1 = -1,41$, $\beta_2 = 1,29$ i $\beta_3 = 1,28$. Međutim, može se pokazati da koeficijent β_1 nije statistički značajan. Prema tome, sistem jednačina koji treba rešiti sastoji se od poslednje dve jednačine sistema (2.11) gde smo fiksirali $\beta_1 = 0$. Svi koeficijenti koje smo ovako dobili su statistički značajni pa je regresiona jednačina

$$y = 1,28x_2 + 1,08x_3\,,$$

gde su sa x_2 i x_3 označeni redom spoljašnji prečnik i količina molibdena. Procenjena vrednost čvrstoće je 11,45. Koeficijent determinacije iznosi 0,87. (Indeksi kontrolisanih promenljivih su 2 i 3 samo iz razloga da bi bili u saglasnosti sa prethodnom teorijom. Inače se prilikom primene koriste indeksi redom, tj. ako su, kao što je u ovom primeru, dva faktora čiji se uticaj posmatra, sasvim je prirodno, a tako se najčešće i označava, indeksi su 1 i 2.)

Ovaj primer možemo rešiti primenom programskog jezika `R`. Najpre uvezimo podatke iz tabele u objekat `tabela`

```
> tabela<-read.csv("primer.csv", header = TRUE)
```

a zatim pozovimo funkciju `lm` na sledeći način

```
> lm.fit<-lm(formula=
   tabela$Cvrstoca~0+tabela$Precnik+tabela$Molidben)
> summary(lm.fit)

 Coefficients:
                  Estimate  Std.  Error  t value    Pr(>|t|)
 tabela$Precnik    1.2847      0.3438      3.737    0.004650 **
 tabela$Molidben   1.0878      0.1900      5.726    0.000285 ***
---
 Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

Koeficijent determinacije dobićemo pozivom funkcije

```
> 1-sum(lm.fit.v$residuals^2)/
   sum((tabela$Cvrstoca-mean(tabela$Cvrstoca))^2)
[1] 0.8737016
```

△

Zadržimo se kratko na vektoru $\mathbf{U}$ definisanom sa

$$\mathbf{U} = \mathbf{Y} - \mathbf{Z}'\widehat{\boldsymbol{\beta}}\,,$$

poznatom pod nazivom **vektor ostataka**, a njegove komponente se nazivaju ostaci. Označimo sa

$$\mathbf{B} = \mathbf{I}_n - \mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}\,.$$

Tada je

$$\mathbf{U} = \mathbf{B}\mathbf{Y}\,.$$

Lako je konstatovati da je matrica $\mathbf{B}$ simetrična i idempotentna, kao i da je

$$\widetilde{\sigma}^2 = \frac{1}{n-k}\mathbf{Y}'\mathbf{B}\mathbf{Y}\,, \quad E(\mathbf{U}) = \mathbf{0} \quad \text{i} \quad \mathbf{D}(\mathbf{U}) = \sigma^2\mathbf{B}\,.$$

Uloga vektora ostataka biće razjašnjena u poglavlju o analizi rasipanja.

Nadalje ćemo razmatrati specijalan slučaj modela linearne regresije druge vrste.

2.1.2 Model normalne regresije

Sa uvođenjem dodatnih pretpostavki o raspodeli vektora grešaka $\boldsymbol{\varepsilon}$ moguće je dobiti još neke ocene u vezi sa koeficijentima linearne regresije. Najčešća pretpostavka je pretpostavka o normalnosti vektora $\boldsymbol{\varepsilon}$ tj. $\boldsymbol{\varepsilon} : \mathcal{N}(\mathbf{0}\,, \sigma^2\mathbf{I}_n)$. U tom slučaju se govori o normalnoj regresiji. Direktna posledica te pretpostavke i definicije modela linearne regresije je da vektor $\mathbf{Y}$ ima takođe normalnu raspodelu sa vektorom očekivanja $\mathbf{Z}'\boldsymbol{\beta}$ i kovarijansnom matricom $\sigma^2\mathbf{I}_n$ tj.

$$\mathbf{Y} : \mathcal{N}(\mathbf{Z}'\boldsymbol{\beta}\,,\, \sigma^2\mathbf{I}_n)\,. \tag{2.12}$$

Model normalne regresije će ovde biti razmatran isključivo na prostom uzorku.

2.1.3 Ocena maksimalne verodostojnosti parametara modela normalne regresije

Normalni model (2.12) sadrži $(k+1)$ parametar, odnosno, definisan je parametrom Θ dimenzije $(k+1)$:

$$\Theta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \\ \sigma^2 \end{pmatrix},$$

čije moguće vrednosti pripadaju Euklidovom poluprostoru $\mathbf{\Theta} \subset E_{k+1}$,

$$\mathbf{\Theta} = \{\Theta : -\infty < \beta_j < +\infty, j = 1, \ldots, k, \sigma^2 > 0\}.$$

Ako je $\boldsymbol{y} = (y_1, \ldots, y_n)'$ realizacija vektora $\mathbf{Y}$ tada će funkcija verodostojnosti biti:

$$L(\Theta; \boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})\right).$$

Vidimo da se maksimum funkcije verodostojnosti po $\boldsymbol{\beta}$ pri svakom fiksiranom σ^2 postiže kada je kvadratna forma $(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})$ minimalna, tj. kada za $\boldsymbol{\beta}$ uzmemo ocenu najmanjih kvadrata dobijenu kod opšteg modela linearne regresije. Otuda se ocena dobijena metodom maksimalne verodostojnosti i ocena dobijena metodom najmanjih kvadrata za $\boldsymbol{\beta}$ poklapaju kod modela normalne regresije.

Kod opšteg modela linearne regresije ocena najmanjih kvadrata je bila najbolja među svim linearnim nepristrasnim ocenama vektora $\boldsymbol{\beta}$. Kod normalne linearne regresije ova ocena je (na osnovu ranije navedenog o oceni maksimalne verodostojnosti) najbolja među svim nepristrasnim ocenama za $\boldsymbol{\beta}$.

Do ocene maksimalne verodostojnosti za nepoznatu disperziju σ^2 dolazi se na uobičajeni način:

$$\ln L(\Theta; \boldsymbol{y}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta}),$$

$$\frac{\partial \ln L}{\partial \sigma^2} = 0\,,$$

$$-\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})\frac{1}{(\sigma^2)^2} = 0\,.$$

Rešavanjem poslednje jednačine po σ^2 dobija se

$$\sigma^2 = \frac{1}{n}(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{Z}'\boldsymbol{\beta})\,.$$

Otuda, ako umesto $\boldsymbol{\beta}$ koristimo ocenu $\widehat{\boldsymbol{\beta}}$, dobijamo statistiku

$$\widehat{\sigma}^2 = \frac{1}{n}S(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}(\mathbf{Y} - \mathbf{Z}'\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}'\widehat{\boldsymbol{\beta}})$$

kao ocenu maksimalne verodostojnosti za σ^2.

Kao što znamo, nepristrasna ocena za σ^2 ima oblik:

$$\widetilde{\sigma}^2 = \frac{S(\widehat{\boldsymbol{\beta}})}{n-k}\,,$$

pa je ocena $\widehat{\sigma}^2$ pristrasna i njena pristrasnost iznosi:

$$\begin{aligned} E\widehat{\sigma}^2 - \sigma^2 &= E\left(\frac{S(\widehat{\boldsymbol{\beta}})}{n}\right) - \sigma^2 \\ &= \frac{1}{n}(n-k)\sigma^2 - \sigma^2 \\ &= -\frac{k}{n}\sigma^2\,. \end{aligned}$$

Dakle, pristrasnost opada sa porastom broja posmatranja n.

2.1.4 Osnovna teorema teorije normalne regresije

Teorema 2.5. *Slučajne veličine $\widehat{\boldsymbol{\beta}}$ i $S(\widehat{\boldsymbol{\beta}})$ su nezavisne, a takve su i $S(\widehat{\boldsymbol{\beta}})$ i $Q = S(\boldsymbol{\beta}) - S(\widehat{\boldsymbol{\beta}})$. Pri tome su njihove raspodele redom:*

$$\widehat{\boldsymbol{\beta}} : \mathcal{N}(\boldsymbol{\beta}\,, \sigma^2\mathbf{A}^{-1})$$

$$\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2} : \chi^2_{(n-k)}$$

$$\frac{Q}{\sigma^2} : \chi^2_k \,.$$

Dokaz. Kompletan dokaz teoreme može se naći na pr. u [Ivčenko G. I., Medvedev J. I., 1984], a mi ćemo ovde dokazati samo nezavisnost navedenih slučajnih veličina.

Uvedimo normirani vektor grešaka

$$\boldsymbol{\varepsilon}^* = \left(\frac{\varepsilon_1}{\sigma}, \ldots, \frac{\varepsilon_n}{\sigma}\right)'$$

čija je raspodela data sa:

$$\boldsymbol{\varepsilon}^* : \mathcal{N}(\mathbf{0}\,, \mathbf{I}_n).$$

Tada vektor $\mathbf{Y}$ dobija oblik

$$\mathbf{Y} = \mathbf{Z}'\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}^*.$$

Odavde se dobija da je:

$$\widehat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{Z}\mathbf{Y} = \mathbf{A}^{-1}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}^*) = \mathbf{A}^{-1}\mathbf{Z}\mathbf{Z}'\boldsymbol{\beta} + \mathbf{A}^{-1}\mathbf{Z}\sigma\boldsymbol{\varepsilon}^* = \boldsymbol{\beta} + \sigma\mathbf{A}^{-1}\mathbf{Z}\boldsymbol{\varepsilon}^*.$$

Kako je

$$\widetilde{\sigma}^2 = \frac{1}{n-k}\mathbf{Y}'\mathbf{B}\mathbf{Y},$$

gde je, kao i do sada,

$$\mathbf{B} = \mathbf{I}_n - \mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}\,.$$

Sledi da je

$$\frac{S(\widehat{\boldsymbol{\beta}})}{n-k} = \widetilde{\sigma}^2 \quad = \quad \frac{1}{n-k}(\mathbf{Z}'\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}^*)'\mathbf{B}(\mathbf{Z}'\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}^*)\,.$$

Odavde je očigledno

$$S(\widehat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{\varepsilon}^{*\prime}\mathbf{B}\boldsymbol{\varepsilon}^*\,,$$

odnosno

$$\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2} = \boldsymbol{\varepsilon}^{*\prime}\mathbf{B}\boldsymbol{\varepsilon}^*\,.$$

Dakle, treba utvrditi nezavisnost sledećih slučajnih promenljivih

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \sigma \mathbf{A}^{-1}\mathbf{Z}\boldsymbol{\varepsilon}^*,$$

i

$$S(\widehat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{\varepsilon}^{*\prime}\mathbf{B}\boldsymbol{\varepsilon}^*.$$

Njihova nezavisnost sledi iz činjenice da je $\mathbf{A}^{-1}\mathbf{ZB} = \mathbf{0}$, odnosno činjenice da je proizvod matrica kvadratne i linearne forme vektora slučajnih grešaka $\boldsymbol{\varepsilon}^*$ jednak nula matrici. To ima za posledicu upravo nezavisnost kvadratne i linearne forme ([Ivčenko G. I., Medvedev J. I., 1984], lema 1.2).

Kako slučajna veličina Q zavisi od uzorka $\mathbf{Y}$ samo preko statistike $\widehat{\boldsymbol{\beta}}$, a slučajne promenljive $\widehat{\boldsymbol{\beta}}$ i $S(\widehat{\boldsymbol{\beta}})$ su nezavisne, to su i Q i $S(\widehat{\boldsymbol{\beta}})$ takođe nezavisne. □

Kao direktne posledice prethodne teoreme mogu se navesti sledeća tvrđenja:
Za svako $j = 1, \ldots, k$ je:

- $$\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{a^{jj}}} : \mathcal{N}(0,1) \tag{2.13}$$

  pri čemu je slučajna promenljiva $\widehat{\beta}_j$ nezavisna od slučajne promenljive $S(\widehat{\boldsymbol{\beta}})$, a a^{jj} je j–ti element glavne dijagonale matrice $\mathbf{A}^{-1}$.

- $$\frac{\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{a^{jj}}}}{\sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2(n-k)}}} = \frac{(\widehat{\beta}_j - \beta_j)\sqrt{n-k}}{\sqrt{a^{jj}S(\widehat{\boldsymbol{\beta}})}} : t_{n-k}\,, \tag{2.14}$$

  dakle, u pitanju je slučajna promenljiva sa Studentovom raspodelom sa $n-k$ stepeni slobode.

- $$\frac{\frac{Q}{k\sigma^2}}{\frac{S(\widehat{\boldsymbol{\beta}})}{(n-k)\sigma^2}} = \frac{n-k}{k} \cdot \frac{Q}{S(\widehat{\boldsymbol{\beta}})} : F_{k,n-k}\,, \tag{2.15}$$

tj. ova slučajna promenljiva ima Fišerovu raspodelu sa k i $(n-k)$ stepeni slobode.

2.1.5 Skupovi poverenja za parametre normalne regresije

Odredićemo najpre interval poverenja za pojedini koeficijent β_j linearne regresije.

Na osnovu posledice (2.13), slučajna promenljiva $\widehat{\beta}_j$ ima raspodelu $\mathcal{N}(\beta_j, \sigma^2 a^{jj})$, $j = 1, \ldots, k$. Dakle, zadatak se svodi na ocenjivanje nepoznatog matematičkog očekivanja normalne raspodele kada je disperzija nepoznata. U tu svrhu možemo koristiti centralnu statistiku:

$$\frac{\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{a^{jj}}}}{\sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2(n-k)}}} : t_{n-k}\,.$$

Uvedimo oznaku

$$\sqrt{\frac{n-k}{a^{jj}}} \cdot \frac{1}{\sqrt{S(\widehat{\boldsymbol{\beta}})}}(\widehat{\beta}_j - \beta_j) = t\,.$$

Odavde se dobija da je:

$$(\widehat{\beta}_j - \beta_j) = t \cdot \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}}\,.$$

Interval poverenja sa nivoom poverenja $1-\alpha$, $0 < \alpha < 1$, sledi iz

$$P\left\{|t| \le t_{n-k,\frac{1-\alpha}{2}}\right\} = 1-\alpha\,, \tag{2.16}$$

gde je $t_{n-k,\frac{1-\alpha}{2}}$ konstanta koja zadovoljava uslov (2.16) pri čemu je t slučajna promenljiva koja ima Studentovu raspodelu sa $n-k$ stepeni slobode. Zaista, iz

$$\left|(\widehat{\beta}_j - \beta_j)\sqrt{\frac{n-k}{S(\widehat{\boldsymbol{\beta}})a^{jj}}}\right| \le t_{n-k,\frac{1-\alpha}{2}}$$

sledi

$$I_{\beta_j} = \left[\widehat{\beta}_j - t_{n-k,\frac{1-\alpha}{2}} \cdot \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}} \quad , \quad \widehat{\beta}_j + t_{n-k,\frac{1-\alpha}{2}} \cdot \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}} \right].$$

Interval poverenja za nepoznatu disperziju σ^2 dobija se na osnovu centralne statistike

$$\frac{S(\widehat{\boldsymbol{\beta}})}{\sigma^2}$$

koja ima $\chi^2_{(n-k)}$ raspodelu. Dakle, dvostrani interval poverenja je, po pravilu, oblika:

$$I_{\sigma^2} = \left[\frac{S(\widehat{\boldsymbol{\beta}})}{\chi^2_{n-k,1-\frac{\alpha}{2}}} \quad , \quad \frac{S(\widehat{\boldsymbol{\beta}})}{\chi^2_{n-k,\frac{\alpha}{2}}} \right],$$

gde je $\chi^2_{n-k,\,1-\frac{\alpha}{2}}$ kvantil reda $1-\frac{\alpha}{2}$ raspodele χ^2 sa $n-k$ stepeni slobode i $\chi^2_{n-k,\,\frac{\alpha}{2}}$ kvantil reda $\frac{\alpha}{2}$ iste raspodele.

Na gore opisani način je moguće postaviti intervale poverenja za svaki od koeficijenata regresije $\beta_1, \ldots, \beta_k$. Ako nađemo k takvih intervala sa jednim istim nivoom poverenja $\gamma = 1-\alpha$, tada će očekivana vrednost broja intervala koji prekrivaju vrednost β_j biti $k\gamma$. Ocenimo sa kojom verovatnoćom svi intervali poverenja istovremeno pokrivaju svoj odgovarajući parametar β_j.

Označimo sa A_j događaj da slučajni interval

$$\left[\widehat{\beta}_j - t_{n-k,\frac{1-\alpha_j}{2}} \cdot \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}} \quad , \quad \widehat{\beta}_j + t_{n-k,\frac{1-\alpha_j}{2}} \cdot \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{jj}}{n-k}} \right]$$

obuhvata pravu vrednost parametra β_j. Sledi da je

$$P(A_j) = 1-\alpha_j, \quad 0 < \alpha_j < 1.$$

Događaj čija nas verovatnoća zanima je $A_1 A_2 \ldots A_k$, pa je:

$$P(A_1 A_2 \ldots A_k) = 1 - P(\cup_{j=1}^{k} A_j^c)$$

$$P(\cup_{j=1}^{k} A_j^c) \leq \sum_{j=1}^{k} P(A_j^c) = \sum_{j=1}^{k} \alpha_j$$

$$P(A_1 A_2 \dots A_k) \geq 1 - \sum_{j=1}^{k} \alpha_j.$$

Ako izaberemo $\alpha_1 = \alpha_2 = \ldots = \alpha_k = \frac{\alpha}{k}$, za neko $\alpha,\ 0 < \alpha < 1$, dobijamo da je:

$$P(A_1 A_2 \dots A_k) \geq 1 - k\frac{\alpha}{k} = 1 - \alpha.$$

Međutim, moguće je postaviti i ovakav zadatak:

U euklidskom prostoru E_k naći oblast poverenja G_γ sa nivoom poverenja γ, $0 < \gamma < 1$, koja sa verovatnoćom γ prekriva nepoznatu parametarsku tačku $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)$.

U tu svrhu koristi se centralna statistika

$$\frac{n-k}{k} \cdot \frac{Q}{S(\widehat{\boldsymbol{\beta}})} = F$$

koja ima Fišerovu raspodelu $F_{k,n-k}$ na sledeći način:

$$\gamma = 1 - \alpha = P\{F \leq F_{k,n-k;1-\alpha}\} = P\{\boldsymbol{\beta} \in G_\gamma(\mathbf{Y})\},$$

gde je $F_{k,\,n-k;\,1-\alpha}$ kvantil reda $1-\alpha$ Fišerove raspodele sa k i $n-k$ stepeni slobode i

$$\begin{aligned} G_\gamma(\mathbf{Y}) &= \left\{ \boldsymbol{\beta} \;\middle|\; \frac{n-k}{k} \cdot \frac{Q}{S(\widehat{\boldsymbol{\beta}})} \leq F_{k,n-k;1-\alpha} \right\} = \\ &= \left\{ \boldsymbol{\beta} \;\middle|\; \frac{S(\boldsymbol{\beta}) - S(\widehat{\boldsymbol{\beta}})}{S(\widehat{\boldsymbol{\beta}})} \leq \frac{k}{n-k} F_{k,n-k;1-\alpha} \right\} = \\ &= \left\{ \boldsymbol{\beta} \;\middle|\; (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{A} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \frac{k}{n-k} S(\widehat{\boldsymbol{\beta}}) F_{k,n-k;1-\alpha} \right\} \end{aligned}$$

i predstavlja oblast elipsoida sa centrom u $\boldsymbol{\beta}$ i granicom datom jednačinom

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{A} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \frac{k}{n-k} S(\widehat{\boldsymbol{\beta}}) F_{k,n-k;1-\alpha}\,.$$

Primer 2.7. Razmotrićemo intervale poverenja i oblast poverenja za model

jednostruke linearne regresije

$$EY_i = \beta_1 + \beta_2 x^{(i)}\,, \quad i = 1, \ldots, n\,.$$

Interval poverenja za β_2 sa nivoom poverenja $\gamma = 1 - \alpha$ ukoliko je disperzija nepoznata, biće:

$$I_{\beta_2} = \left[\widehat{\beta}_2 - t_{n-2,\frac{1-\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{22}}{n-2}} \quad , \quad \widehat{\beta}_2 + t_{n-2,\frac{1-\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})a^{22}}{n-2}} \right]$$

pri čemu je, u ovom specijalnom slučaju,

$$a^{22} = \frac{1}{\sum_{i=1}^{n}(x^{(i)} - \overline{x}_n)^2}\,,$$

pa je

$$I_{\beta_2} = \left[\widehat{\beta}_2 - t_{n-2,\frac{1-\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{(n-2)\sum_{i=1}^{n}(x^{(i)} - \overline{x}_n)^2}}\,, \right.$$
$$\left. \widehat{\beta}_2 + t_{n-2,\frac{1-\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{(n-2)\sum_{i=1}^{n}(x^{(i)} - \overline{x}_n)^2}} \right]\,.$$

Interval poverenja možemo dobiti iz programskog jezika `R` pozivom funkcije `confint`. Ako bismo posmatrali model iz Primera 2.2, interval poverenja za $\alpha = 0.05$ dobićemo kao

```
> confint(object=lm.fit, level = 0.95)

                   2.5%       97.5%
(Intercept)  66.1879192  68.662251
table$Temp    0.8331248   0.906826
```

Ove vrednosti možemo i sami da izračunamo pomoću gore navedene jednačine. U tom slučaju treba najpre izračunati $S(\widehat{\boldsymbol{\beta}})$ kao

```
> S=t(lm.fit$residuals) %*% B %*% lm.fit$residuals
```

zatim a^{22} kao

```
> a22 = 1/(var(table$Temp)*(n-1))
```

gde je B matrica definisana u prethodnoj diskusiji, a $n = 9$. Traženi kvantil za studentovu raspodelu dobićemo kao

```
> qt(p=1-.05/2, df=(n-2))
```

Ako tražimo oblast poverenja $G_\gamma(\mathbf{Y})$ koja natkriva tačku (β_1, β_2) u euklidskom prostoru E_2 sa verovatnoćom γ , dobijamo oblast ograničenu elipsom:

$$\begin{aligned}
G_\gamma(\mathbf{Y}) &= \\
&= \left\{ \boldsymbol{\beta} \left| \|\widehat{\beta}_1 - \beta_1, \widehat{\beta}_2 - \beta_2\| \right\| \begin{matrix} n & \sum_{i=1}^n x^{(i)} \\ \sum_{i=1}^n x^{(i)} & \sum_{i=1}^n {x^{(i)}}^2 \end{matrix} \right\| \begin{matrix} \widehat{\beta}_1 - \beta_1 \\ \widehat{\beta}_2 - \beta_2 \end{matrix} \right\| \leq \\
&\leq \left. \frac{2}{n-2} S(\widehat{\boldsymbol{\beta}}) F_{2,n-2;\gamma} \right\} = \\
&= \left\{ \boldsymbol{\beta} \left| (\widehat{\beta}_1 - \beta_1)^2 n + 2(\widehat{\beta}_1 - \beta_1)(\widehat{\beta}_2 - \beta_2) \sum x^{(i)} + (\widehat{\beta}_2 - \beta_2)^2 \sum {x^{(i)}}^2 \leq \right. \right. \\
&\leq \left. \frac{2}{n-2} S(\widehat{\boldsymbol{\beta}}) F_{2,n-2;\gamma} \right\} = \\
&= \left\{ \boldsymbol{\beta} \left| (\widehat{\beta}_1 - \beta_1)^2 + 2\overline{x}_n (\widehat{\beta}_1 - \beta_1)(\widehat{\beta}_2 - \beta_2) + \frac{1}{n} (\widehat{\beta}_2 - \beta_2)^2 \sum {x^{(i)}}^2 \leq \right. \right. \\
&\leq \left. \left. \frac{2}{n(n-2)} S(\widehat{\boldsymbol{\beta}}) F_{2,n-2;\gamma} \right\} . \triangle \right.
\end{aligned}$$

Moguće je tražiti oblast poverenja i za linearnu kombinaciju parametara regresije, tj. za vektor $\boldsymbol{t} = \mathbf{T}\boldsymbol{\beta}$, gde je matrica $\mathbf{T}$ dimenzije $m \times k$, a $rang\mathbf{T} = m$. U tu svrhu koristi se činjenica da $\widehat{\boldsymbol{t}} = \mathbf{T}\widehat{\boldsymbol{\beta}}$ ima raspodelu $\mathcal{N}(\boldsymbol{t}, \sigma^2\mathbf{D})$, gde je matrica $\mathbf{D}$ definisana sa $\mathbf{D} = \mathbf{T}\mathbf{A}^{-1}\mathbf{T}'$. Odavde sledi da $Q_\mathbf{T} = (\widehat{\boldsymbol{t}} - \boldsymbol{t})'\mathbf{D}^{-1}(\widehat{\boldsymbol{t}} - \boldsymbol{t})$ ne zavisi od $S(\widehat{\boldsymbol{\beta}})$ i količnik $\frac{Q_\mathbf{T}}{\sigma^2}$ ima χ^2 raspodelu sa m stepeni slobode, $\frac{Q_\mathbf{T}}{\sigma^2} : \chi^2_m$.

2.1.6 Testiranje hipoteza o ocenama parametara normalne regresije

Testiraćemo nultu hipotezu

$$H_0 : (\beta_1, \beta_2, \ldots, \beta_k) \in \mathcal{B}_0 \subset E_k$$

protiv odgovarajuće alternativne. Najčešće je oblast $\mathcal{B}_0$ linearni potprostor od E_k oblika:

$$\mathcal{B}_0 = \{\boldsymbol{\beta} \,|\, \mathbf{T}\boldsymbol{\beta} = \boldsymbol{t}_0\},$$

gde je $\mathbf{T}$, kao i u prethodnom, matrica dimenzije $m \times k$, a $\boldsymbol{t}_0$ vektor dimenzije $m \times 1$.

Preciznije, najopštiji oblik hipoteze vezane za nepoznati višedimenzioni parametar $\Theta = (\boldsymbol{\beta}' \vdots \sigma^2)'$ linearne regresije koju treba testirati je

$$H_0 : \Theta \in \boldsymbol{\Theta} = \{\Theta \,|\, \boldsymbol{\beta} \in \mathcal{B}_0, \sigma^2 > 0\}\,.$$

Kritična oblast veličine α za testiranje ove složene nulte hipoteze protiv alternativne

$$H_1 : \Theta \in \boldsymbol{\Theta} = \{\Theta \,|\, \boldsymbol{\beta} \in \mathcal{B}_0^c, \sigma^2 > 0\}\,,$$

koja je takođe složena, je

$$C = \left\{ \boldsymbol{y} \quad \middle| \quad \frac{n-k}{m} \frac{(\widehat{\boldsymbol{t}} - \boldsymbol{t}_0)' \mathbf{D}^{-1} (\widehat{\boldsymbol{t}} - \boldsymbol{t}_0)}{S(\widehat{\boldsymbol{\beta}})} \geq F_{m,n-k;1-\alpha} \quad \right\}\,,$$

gde je

$$\mathbf{D} = \mathbf{T}\mathbf{A}^{-1}\mathbf{T}'\,, \quad \widehat{\boldsymbol{t}} = \mathbf{T}\widehat{\boldsymbol{\beta}}\,,$$

dok je oblast prihvatanja H_0 njen komplement.

Primer 2.8. Razmotrićemo ponovo model jednostruke linearne regresije $Y = \beta_1 + \beta_2 x + \varepsilon$. Testiraćemo hipotezu

$$H_0 : \beta_2 = \beta_{20}\,,$$

što znači da se testira samo koeficijent pravca prave kojom je predstavljen regresioni model.

S obzirom na to da je u ovom primeru $m = 1$ (dimenzija matrice $\mathbf{T}$ je 1×2), granica kritične oblasti će biti kvantil reda $1-\alpha$ slučajne promenljive sa $F_{1,n-2}$ raspodelom. Međutim, poznato je da se Fišerova raspodela sa $(1, n-k)$ stepeni slobode poklapa sa raspodelom kvadrata slučajne promenljive koja ima Studentovu raspodelu sa $n-k$ stepeni slobode. Zbog toga i na osnovu (2.14), za alternativnu hipotezu $H_1 : \beta_2 \neq \beta_{20}$, dobili bismo kritičnu oblast iz uslova:

$$|\widehat{\beta}_2 - \beta_{20}| \geq t_{n-2,\frac{1-\alpha}{2}} \sqrt{\frac{S(\widehat{\boldsymbol{\beta}})}{(n-2)\sum_{i=1}^{n}(x^{(i)} - \overline{x}_n)^2}}\,,$$

gde je

$$S(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \left(Y_i - \beta_{20} x^{(i)} - \widehat{\beta}_1 \right)^2 , \quad \widehat{\beta}_1 = \overline{Y}_n - \beta_{20} \overline{x}_n \, . \, \triangle$$

Razmotrimo i opšti problem testiranja hipoteza vezanih za koeficijente normalne linearne regresije.

Razmatraćemo samo osnovni model. Preciznije, ako na obeležje Y utiče l neslučajnih faktora x_1, x_2, ..., x_l gde je $l > 2$, razmatraćemo model (A):

$$(A) \qquad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_g x_g + \beta_{g+1} x_{g+1} + \ldots + \beta_l x_l + \varepsilon.$$

Podsetimo se da razmatramo ovaj model pod pretpostavkom da je ε slučajna promenljiva sa normalnom raspodelom. To bi za posledicu imalo da Y ima normalnu raspodelu, kao i to da statistike $\widehat{\beta}_j$, $j = 0, 1, \ldots, l$ kojima se ocenjuju parametri regresije β_j, $j = 0, 1, \ldots, l$ imaju normalnu raspodelu. Pri fitovanju obeležja Y na ovaj način, važno je utvrditi da li svi pobrojani faktori utiču na obeležje sa istim utvrđenim pragom značajnosti ili se za neke od njih može utvrditi da nisu od značaja, čime bi se smanjio broj sabiraka modela (A). Hipoteza o tome da neki od faktora nisu od dovoljnog značaja da bi bitno uticali na obeležje Y postavlja se na sledeći način:

$$H_0 : \beta_{g+1} = \beta_{g+2} = \ldots = \beta_l = 0$$

za neko $g < l$, gde se bez smanjenja opštosti uzima da su krajnjih $l - g$ faktora modela (A) manje značajnosti od postavljenog praga.

Ukoliko se testiranjem postavljene hipoteze prihvati hipoteza H_0, prelazi se na model (B), tzv. redukovani model u odnosu na model (A), koji se smatra potpunim modelom:

$$(B) \qquad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_g x_g + \varepsilon.$$

Odgovarajuće sume kvadrata odstupanja za uzorak obima n za model (A) i model (B) su redom:

$$Q_A = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \ldots + \beta_g x_g^{(i)} + \beta_{g+1} x_{g+1}^{(i)} + \ldots + \beta_l x_l^{(i)}) \right)^2 ,$$

$$Q_B = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \ldots + \beta_g x_g^{(i)}) \right)^2 ,$$

gde smo, da se podsetimo, i–ti nivo pojedinog faktora uticaja označili gornjim indeksom u zagradi.

Potpuni model, model (A), će davati predviđanje sa manjom greškom u odnosu na redukovani model, model (B), tj. $Q_A < Q_B$. Otuda se Q_B može predstaviti na sledeći način:

$$Q_B = Q_A + (Q_B - Q_A),$$

gde su Q_A i $Q_B - Q_A$ nezavisne slučajne veličine, za koje, pri uslovu normalne regresije, slučajne promenljive

$$\frac{Q_B}{\sigma^2} \ , \quad \frac{Q_A}{\sigma^2} \ \text{ i } \ \frac{Q_B - Q_A}{\sigma^2}$$

imaju χ^2 raspodele sa odgovarajućim brojem stepeni slobode, a statistike koje se dobijaju kao njihovi količnici imaju Fišerove raspodele. Otuda se za testiranje postavljene nulte hipoteze koristi test statistika

$$F_{\nu_1,\nu_2} = \frac{(n-l-1)(Q_B - Q_A)}{(l-g)Q_A} ,$$

koja ima Fišerovu raspodelu sa $\nu_1 = l - g$ i $\nu_2 = n - l - 1$ stepeni slobode, na osnovu koje se određuje kritična oblast testa.

Napomenimo da se kvantili za Fišerovu raspodelu mogu dobiti iz programskog jezika `R` kao

```
> qf(p=1-α, df1=(l-g), df2=(n-l-1))
```

gde nam je α prag značajnosti.

2.2 Regresija prve vrste (regresija i korelacija)

Neka su dati obeležje Y i slučajni vektor $\mathbf{X} = (X_1, \ldots, X_p)$, gde su $X_1, \ldots, X_p$, obeležja koja utiču na obeležje Y, tj. obeležja Y i $\mathbf{X}$ su povezana nekom statističkom zavisnošću. Pretpostavimo da nam je poznata i zajednička

raspodela $F_{\mathbf{X}Y}(x_1, \ldots, x_p, y)$. Vektor $\mathbf{X}$ je dostupan ispitivanju, tj. možemo da registrujemo njegove vrednosti tokom eksperimenta, dok to nije slučaj sa veličinom Y. Sve informacije o obeležju Y dobijamo preko vektora $\mathbf{X}$, pa u tom smislu komponente vektora $\mathbf{X}$ nazivamo **prediktori** ili **prediktorske (prognostičke, predviđajuće) promenljive**. Ideja o predviđanju se matematički ostvaruje pretpostavkom da je moguće odrediti statistiku $\psi(\mathbf{X})$ kojom ćemo na zadovoljavajući način ocenjivati vrednosti obeležja Y. Takva statistika se zove **predikcija** ili **predviđanje** za obeležje Y na osnovu $\mathbf{X}$.

Razradom metoda nalaženja optimalnog predviđanja prema pojedinim kriterijumima bavi se teorija statističke regresije.

2.2.1 Najbolje predviđanje za obeležje Y na osnovu vektora X

Vratimo se zajedničkoj raspodeli $F_{\mathbf{X}Y}(\mathbf{x}, y)$, $\mathbf{x} = (x_1, \ldots, x_p) \in R^p$, $y \in R$, za koju za sada pretpostavimo da nam je poznata. U tom slučaju je često moguće odrediti uslovnu raspodelu, tj. funkciju raspodele $F_{Y|\mathbf{X}=\mathbf{x}}(y|\mathbf{x})$, kao i odgovarajuću funkciju gustine raspodele za slučaj raspodele apsolutno neprekidnog, odnosno funkciju mase kod slučajnih promenljivih diskretnog tipa.

Model regresije prve vrste se bazira na uslovnom matematičkom očekivanju. Uslovno matematičko očekivanje $E(Y|\mathbf{X} = \mathbf{x})$ se može posmatrati kao funkcija od $\mathbf{x}$. Označimo je sa

$$M(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})\,.$$

Neka je $\psi(\mathbf{X})$ proizvoljno predviđanje za Y na osnovu $\mathbf{X}$. Srednjekvadratna greška tog predviđanja će biti

$$E\left(Y - \psi(\mathbf{X})\right)^2.$$

Optimalno (najbolje) predviđanje za Y, u oznaci $\psi^*(\mathbf{X})$, će biti ono koje minimizira srednjekvadratno odstupanje, tj.

$$E\left(Y - \psi^*(\mathbf{X})\right)^2 = \inf_{\psi} E\left(Y - \psi(\mathbf{X})\right)^2.$$

Dokažimo da je $\psi^*(\mathbf{X}) = M(\mathbf{X})$. Zaista,

$$\begin{aligned} E\left(Y-\psi(\mathbf{X})\right)^2 &= E\left(Y-M(\mathbf{X})+M(\mathbf{X})-\psi(\mathbf{X})\right)^2 = \\ &= E\left(Y-M(\mathbf{X})\right)^2 + 2E[(Y-M(\mathbf{X}))(M(\mathbf{X})-\psi(\mathbf{X}))] + \\ &+ E\left(M(\mathbf{X})-\psi(\mathbf{X})\right)^2 \geq E\left(Y-M(\mathbf{X})\right)^2 , \end{aligned} \tag{2.17}$$

jer je

$$\begin{aligned} & E[(Y-M(\mathbf{X}))(M(\mathbf{X})-\psi(\mathbf{X}))] = \\ = \; & E\{E[(Y-M(\mathbf{X}))(M(\mathbf{X})-\psi(\mathbf{X}))|\mathbf{X}]\} = \\ = \; & E\{(M(\mathbf{X})-\psi(\mathbf{X}))E[(Y-M(\mathbf{X}))|\mathbf{X}]\} = \\ = \; & E\left\{(M(\mathbf{X})-\psi(\mathbf{X}))[E(Y|\mathbf{X})-M(\mathbf{X})]\right\} = 0\,. \end{aligned}$$

U (2.17) važi jednakost ako i samo ako je

$$E\left(M(\mathbf{X})-\psi(\mathbf{X})\right)^2 = 0\,,$$

što će biti tačno ako i samo ako je

$$M(\mathbf{X}) = \psi(\mathbf{X})$$

skoro sigurno, što je i trebalo dokazati.

Ako sa Δ označimo minimalnu grešku predviđanja po kriterijumu srednje-kvadratnog odstupanja, biće $\Delta = E(Y-M(\mathbf{X}))^2$. Tada je

$$\Delta = E\left\{E[(Y-M(\mathbf{X}))^2|\mathbf{X}]\right\}\,.$$

DEFINICIJA 2.2. Funkcija po $\mathbf{x}$ u oznaci $D(Y|\mathbf{X}=\mathbf{x})$ ili kraće $\sigma^2_{Y\mathbf{X}}$, definisana kao

$$D(Y|\mathbf{X}=\mathbf{x}) = \sigma^2_{Y\mathbf{X}} = E\left((Y-M(\mathbf{X}))^2|\mathbf{X}=\mathbf{x}\right)$$

zove se *uslovna disperzija* za Y, na osnovu vektora $\mathbf{X}=\mathbf{x}$. $\diamondsuit$

Važno svojstvo predviđanja $M(\mathbf{X})$ daje sledeća teorema.

Teorema 2.6. *Veličina $M(\mathbf{X})$ ima maksimalnu koreliranost sa Y među svim predviđanjima za Y na osnovu vektora $\mathbf{X}$.*

Dokaz. Treba naći koeficijent korelacije slučajnih veličina Y i $M(\mathbf{X})$. Da bismo to našli, pođimo od proizvoljnog predviđanja za Y na osnovu $\mathbf{X}$, $\psi(\mathbf{X})$, i odredimo najpre kovarijansu

$$\begin{aligned} Cov(\psi(\mathbf{X}), Y) &= E[(\psi(\mathbf{X}) - E\psi(\mathbf{X}))(Y - EY)] = \\ &= E\{E[(\psi(\mathbf{X}) - E\psi(\mathbf{X}))(Y - EY)|\mathbf{X}]\} = \\ &= E\{(\psi(\mathbf{X}) - E\psi(\mathbf{X}))[E(Y|\mathbf{X}) - E(E(Y|\mathbf{X}))]\} = \\ &= Cov(\psi(\mathbf{X}), M(\mathbf{X}))\,. \end{aligned} \tag{2.18}$$

Sada možemo da procenimo traženi koeficijent korelacije

$$\rho^2_{\psi Y} = \frac{Cov^2(\psi(\mathbf{X}), Y)}{\sigma^2_\psi \sigma^2_Y} = \frac{Cov^2(\psi(\mathbf{X}), M)}{\sigma^2_\psi \sigma^2_Y} \cdot \frac{\sigma^2_M}{\sigma^2_M} = \rho^2_{\psi M} \frac{\sigma^2_M}{\sigma^2_Y}\,,$$

gde su σ^2_ψ, σ^2_Y i σ^2_M disperzije redom slučajnih promenljivih $\psi(\mathbf{X})$, Y i $M(\mathbf{X})$.

Ako je

$$\psi(\mathbf{X}) = M(\mathbf{X})$$

skoro sigurno, tada je na osnovu (2.18)

$$Cov(M, Y) = Cov(M, M) = \sigma^2_M\,.$$

Odavde je

$$\rho_{MY} = \frac{Cov(M, Y)}{\sigma_M \sigma_Y} = \frac{\sigma_M}{\sigma_Y}$$

$$\rho^2_{\psi Y} = \rho^2_{\psi M} \rho^2_{MY} \leq \rho^2_{MY}\,.$$

Zaključujemo da M ima maksimalnu koreliranost sa Y tj. važi:

$$|\rho_{\psi Y}| \leq |\rho_{MY}|. \square$$

Kvadrat koeficijenta korelacije predviđanja $M(\mathbf{X})$ i predviđene slučajne veličine Y ima posebno mesto u regresionoj analizi, te se iz tog razloga uvodi sledeća teorijska definicija.

Definicija 2.3. Veličina

$$\eta^2_{Y\mathbf{X}} = \frac{\sigma^2_M}{\sigma^2_Y} = \rho^2_{MY}$$

se naziva *korelacioni količnik* ili *koeficijent determinacije.* $\diamondsuit$

Uzorački analogon ove veličine je, da se podsetimo, dat sa (2.1). On daje udeo modelom objašnjenih varijacija u ukupnim varijacijama za Y.

Zadržimo se sada na linearnim predviđanjima za Y na osnovu vektora $\mathbf{X}$.

Teorema 2.7. *Funkcija*

$$\psi^*(\mathbf{X}) = \beta^*_0 + \boldsymbol{\beta}^{*'}\mathbf{X}$$

predstavlja najbolje linearno predviđanje za Y na osnovu vektora $\mathbf{X}$, gde je

$$\beta^*_0 = EY - \boldsymbol{\beta}^{*'}E\mathbf{X}\,, \quad \boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}\mathbf{a}\,, \quad \boldsymbol{\Sigma} = \|Cov(X_i, X_j)\|_{p\times p}\,,$$

$$\mathbf{a} = (Cov(Y, X_1), \ldots, Cov(Y, X_p))'\,,$$

među svim linearnim predviđanjima, i maksimalno je korelirana sa Y među svim linearnim predviđanjima za Y na osnovu vektora $\mathbf{X}$.

Dokaz. Pođimo od linearne funkcije $\psi(\mathbf{X}) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}$. Prema uvedenom kriterijumu minimalnog srednjekvadratnog odstupanja optimalne vrednosti za β_i će biti β^*_i koje minimalizuju srednjekvadratno odstupanje:

$$\begin{aligned} E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{X})^2 &= E(Y - EY + EY - \beta_0 - \boldsymbol{\beta}'\mathbf{X} + \boldsymbol{\beta}'E\mathbf{X} - \boldsymbol{\beta}'E\mathbf{X})^2 = \\ &= E[(Y - EY) + (EY - \beta_0 - \boldsymbol{\beta}'E\mathbf{X}) - \boldsymbol{\beta}'(\mathbf{X} - E\mathbf{X})]^2\,. \end{aligned} \tag{2.19}$$

Uvedimo sledeće oznake

$$b_0 = \beta_0 - EY + \boldsymbol{\beta}'E\mathbf{X}\,, \quad \sigma^2_Y = E(Y - EY)^2.$$

Smenom u jednakosti (2.19), dobićemo da je

$$E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{X})^2 = \sigma^2_Y + b^2_0 + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{a}\,. \tag{2.20}$$

Dokažimo da su

$$\beta_0^* = EY - \boldsymbol{\beta}^{*'} E\mathbf{X} \quad \text{i} \quad \boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}\mathbf{a},$$

traženi optimalni koeficijenti linearne regresije. Posmatrajmo vektor $\boldsymbol{\beta}$ definisan sa

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* + \boldsymbol{\delta}$$

i uvrstimo ga u kriterijumsku funkciju (2.20). Sada je

$$\begin{aligned} E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{X})^2 &= \sigma_Y^2 + b_0^2 + (\boldsymbol{\beta}^* + \boldsymbol{\delta})'\boldsymbol{\Sigma}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - 2(\boldsymbol{\beta}^* + \boldsymbol{\delta})'\mathbf{a} = \\ &= \sigma_Y^2 + b_0^2 + \mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a} + \boldsymbol{\delta}'\boldsymbol{\Sigma}\boldsymbol{\delta} - 2\mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a} = \\ &= \sigma_Y^2 + b_0^2 + \boldsymbol{\delta}'\boldsymbol{\Sigma}\boldsymbol{\delta} - \mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a} \geq \\ &\geq \sigma_Y^2 - \mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a}\,. \end{aligned}$$

Jednakost će se postići ako i samo ako je $b_0 = 0$ i $\boldsymbol{\delta} = \mathbf{0}$, tj. nula vektor, odnosno ako i samo ako je

$$\beta_0 = \beta_0^* \quad \text{i} \quad \boldsymbol{\beta} = \boldsymbol{\beta}^*\,.$$

Dokažimo još deo tvrđenja koji se odnosi na koreliranost:

$$\begin{aligned} Cov(Y, \boldsymbol{\beta}'\mathbf{X}) &= E[(Y - EY)(\boldsymbol{\beta}'\mathbf{X} - E(\boldsymbol{\beta}'\mathbf{X}))] = \\ &= \boldsymbol{\beta}' E[(Y - EY)(\mathbf{X} - E\mathbf{X})] = \\ &= (\beta_1, \ldots, \beta_p) \begin{pmatrix} Cov(Y, X_1) \\ Cov(Y, X_2) \\ \\ Cov(Y, X_p) \end{pmatrix} = \\ &= \boldsymbol{\beta}'\mathbf{a} = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^*. \end{aligned}$$

S druge strane je

$$Cov(Y, \boldsymbol{\beta}^{*'}\mathbf{X}) = \boldsymbol{\beta}^{*'}\boldsymbol{\Sigma}\boldsymbol{\beta}^* = D(\boldsymbol{\beta}^{*'}\mathbf{X}) \geq 0.$$

Dakle, korelacija između Y i $\boldsymbol{\beta}^{*'}\mathbf{X}$ će iznositi

$$\rho^2_{Y,\beta^{*'}\mathbf{X}} = \frac{Cov^2(Y, \boldsymbol{\beta}^{*'}\mathbf{X})}{\sigma^2_Y \sigma^2_{\beta^{*'}\mathbf{X}}},$$

a odavde sledi da je

$$\sigma^2_Y \rho^2_{Y,\beta^{*'}\mathbf{X}} = Cov(Y, \boldsymbol{\beta}^{*'}\mathbf{X}) = \boldsymbol{\beta}^{*'}\boldsymbol{\Sigma}\boldsymbol{\beta}^*,$$

te je

$$\rho^2_{Y,\beta'\mathbf{X}} = \frac{(\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^*)^2}{\sigma^2_Y \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}.$$

Ako primenimo nejednakost Koši–Švarc–Bunjakovskog, dobićemo da je

$$\begin{aligned}
\sigma^2_Y \rho^2_{Y,\beta'\mathbf{X}} &= \frac{(\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^*)^2}{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}} \leq \\
&\leq \frac{(\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})(\boldsymbol{\beta}^{*'}\boldsymbol{\Sigma}\boldsymbol{\beta}^*)}{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}} = \\
&= \boldsymbol{\beta}^{*'}\boldsymbol{\Sigma}\boldsymbol{\beta}^* = \sigma^2_Y \rho^2_{Y,\beta^{*'}\mathbf{X}}.
\end{aligned}$$

Dakle, dobija se da je

$$\rho^2_{Y,\beta'\mathbf{X}} \leq \rho^2_{Y,\beta^{*'}\mathbf{X}}, \quad \text{odnosno} \quad |\rho_{Y,\beta'\mathbf{X}}| \leq |\rho_{Y,\beta^{*'}\mathbf{X}}|,$$

a odavde je konačno

$$|\rho_{Y,\psi(\mathbf{X})}| \leq |\rho_{Y,\psi^*(\mathbf{X})}|,$$

što je i trebalo dokazati. □

Imajući u vidu Teoremu 2.6, da je uslovno matematičko očekivanje $M(\mathbf{X})$ statistika koja je maksimalno korelirana sa Y u odnosu na sva druga predviđanja za Y na osnovu $\mathbf{X}$, to ako je $M(\mathbf{X})$ linearna funkcija, na osnovu Teoreme 2.7 sledi da je $M(\mathbf{X})$ oblika

$$M(\mathbf{X}) = EY + \mathbf{a}'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - E\mathbf{X}). \tag{2.21}$$

Zaista, ako je $M(\mathbf{X}) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}$, a optimalne vrednosti za β_0 i $\boldsymbol{\beta}$ su date teoremom 2.7, zamenom je lako proveriti da je $M(\mathbf{X})$ dato sa (2.21).

Primer 2.9. Neka na slučajni ishod eksperimenta Y utiče samo jedan slučajni faktor X i neka je poznata njihova zajednička raspodela koja je normalna

$$(X,Y) : \mathcal{N}\left(\mu_X\,,\,\mu_Y\,,\,\sigma_X^2\,,\,\sigma_Y^2\,,\,\rho\right)\,,\quad \rho = Cov\left(\frac{X-\mu_X}{\sigma_X}\,,\,\frac{Y-\mu_Y}{\sigma_Y}\right)\,.$$

U tom slučaju je njihova zajednička gustina

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \right.\right.$$
$$\left.\left. -2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}\,,$$

a uslovna gustina za Y pod uslovom $X = x$ je

$$f_{Y|X}(y|x) = \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} exp\left\{-\frac{(y-m(x))^2}{2\sigma_Y^2(1-\rho^2)}\right\}\,,$$

gde je

$$m(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x-\mu_X)\,.$$

Dakle, uslovna raspodela je takođe normalna i to $\mathcal{N}\Big(m(x), \sigma_Y^2(1-\rho^2)\Big)$, pa je

$$M(x) = E(Y|X=x) = m(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x-\mu_X)\,.$$

Možemo da konstatujemo da je funkcija regresije Y po X linearna:

$$M(X) = EY + \frac{Cov(X,Y)}{\sigma_X^2}(X-EX)\,.$$

Uslovna disperzija, odnosno srednjekvadratna greška predviđanja, je

$$\sigma_{YX}^2 = \sigma_Y^2 - \sigma_M^2$$

a disperzija za M i, kao posledica, korelacioni količnik su

$$\sigma_M^2 = E\left(M(X) - EM(X)\right)^2 = \frac{Cov^2(X,Y)}{\sigma_X^2}\,,\quad \eta_{YX}^2 = \frac{\sigma_M^2}{\sigma_Y^2} = \rho^2\,.$$

Kao što se vidi, najbolje predviđanje u smislu srednjekvadratnog odstupanja kod normalne raspodele je *linearno*. △

2.3 Logistička regresija

Modeli linearne regresije o kojima je bilo reči podrazumevaju da je zavisna promenljiva Y apsolutno neprekidnog tipa. Kao takvi, ovi modeli nisu pogodni za rešavanje problema gde je Y kategorijskog tipa. Recimo, zavisna promenljiva Y nam opisuje da li pada kiša. To zavisi od mnogo faktora, a oni su elementi vektora nezavisne promenljive $\boldsymbol{X}$ koji može da se sastoji iz elemenata koji su svi slučajne promenljive, ili su svi neslučajne promenljive, ili, pak, može da sadrži i slučajne i neslučajne faktore. Model logističke regresije ima za cilj da opiše funkcionalnu vezu između promenljivih Y i $\boldsymbol{X}$, ako ona postoji. Logistička regresija je jedan od najpopularnijih modela za podatke kategorijskog tipa. Naročito važno mesto ima u klasifikaciji binarnih podataka. Prema broju klasa koje model može da opiše, logističku regresiju možemo svrstati u dva osnovna tipa *binarnu* i *višeklasnu*. Binarna logistička regresija opisuje samo dva ishoda 0 ili 1. Višeklasna logistička regresija je adaptacija binarnog modela kako bi se uključilo više od dve klase, odnosno više od dve vrednosti obeležja Y. Mi ćemo se nadalje detaljnije baviti binarnom logističkom regresijom.

2.3.1 Binarna logistička regresija

Binarna logistička regresija je model regresije koji na osnovu realizacije nezavisne promenljive $\boldsymbol{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)})'$ vrši klasifikaciju zavisne promenljive Y_i koja se može svrstati u klasu 0 ili 1. Dakle, Y_i je ishod eksperimenta za i-ti nivo posmatranih faktora. Izlaz iz modela je vrednost između 0 i 1. Kao takva, može se interpretirati kao verovatnoća da zavisna promenljiva Y_i pripada jednoj ili drugoj klasi. Ako se za granicu koja određuje pripadnost klasi uzme broj 0,5 , onda je sasvim intuitivna interpretacija preko verovatnoća. U skladu sa problemom koji se rešava granica može biti bilo koji broj između 0 i 1.

Pored dvoklasne klasifikacije, model binarne logističke regresije nam opisuje i kakav je uticaj pojedinog obeležja (elementa) iz vektora $\boldsymbol{X}$ na ishod. Na

primer, dobićemo odgovor za koliko je veća verovatnoća da će pasti kiša ako se vlažnost vazduha poveća za deset posto. Kao takav, model binarne logističke regresije je našao široku primenu kako u statistici tako i u mašinskom učenju.

Dakle, osnovna pretpostavka modela je da posmatrana obeležja dovode do ishoda 0 ili 1. Ideja je da odredimo verovatnoću događaja na koji posmatrana obeležja imaju uticaj. Kao i kod Bernulijeve raspodele, neka je p_i verovatnoća događaja 1, a $1 - p_i$ verovatnoća događaja 0, gde je p_i određeno u zavisnosti od nivoa kontrolisanih faktora, tj. vektora $\boldsymbol{x}^{(i)}$. Kod modela linearne regresije imali smo da je za n nivoa posmatranih faktora uticaja

$$EY_i = \boldsymbol{x}^{(i)'}\boldsymbol{\beta}, \quad i = 1, 2 \ldots, n.$$

Napomenimo da je $\boldsymbol{\beta}$ vektor nepoznatih parametara, tj. koeficijenata modela. Cilj je da od ove jednačine dođemo do modela logističke regresije. Da bismo to uradili definisaćemo **šansu** događaja kao

$$odds(p_i) = \frac{p_i}{1 - p_i}, \quad i = 1, 2, \ldots, n.$$

Model logističke regresije bazira se na pretpostavci da je prirodni logaritam šanse opisan pomoću nezavisne promenljive $\boldsymbol{x}$, odnosno

$$\ln odds(p_i) = \boldsymbol{x}^{(i)'}\boldsymbol{\beta}, \quad i = 1, 2 \ldots, n,$$

pa je dakle

$$\ln \frac{p_i}{1 - p_i} = \boldsymbol{x}^{(i)'}\boldsymbol{\beta}, \quad i = 1, 2 \ldots, n.$$

Funkcija na levoj strani ove jednačine se naziva *logit* od p_i, pa odatle i naziv logistička regresija. Rešavanjem jednačine po p_i dobićemo da je

$$p_i = \frac{e^{\boldsymbol{x}^{(i)'}\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}^{(i)'}\boldsymbol{\beta}}} = \frac{1}{1 + e^{-\boldsymbol{x}^{(i)'}\boldsymbol{\beta}}}, \quad i = 1, 2 \ldots, n,$$

što je zapravo **sigmoidna funkcija**, tj. funkcija čiji je grafik u obliku slova S, ograničena, diferencijabilna, realna funkcija, definisana na celom skupu realnih brojeva, sa nenegativnim izvodom u svakoj tački oblasti definisanosti. Grafik

sigmoidne funkcije za ovaj slučaj dat je na slici 2.3. Dakle, verovatnoću i-tog nivoa smo izrazili u funkciji linearne kombinacije kontrolisanih faktora. Promenljivu p_i posmatraćemo kao funkciju $p_i = p_i(\boldsymbol{x}^{(i)}, \boldsymbol{\beta})$, pri čemu ona predstavlja parametar Bernulijeve raspodele za slučajnu promenljivu Y_i.

Dakle, za skup opserviranih vrednosti $(\boldsymbol{X}, \boldsymbol{y})$, funkcija verodostojnosti modela, pod pretpostavkom da su posmatranja nezavisna među sobom, biće

$$L(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i},$$

pri čemu je $\boldsymbol{X}$ matrica plana dimenzije $(k+1)\times n$, gde svaka kolona predstavlja opservaciju, a prva vrsta je jedinični vektor. Vektor $\boldsymbol{y}$ ima n elemenata gde je $y_i \in \{0, 1\}$. U opštem slučaju, lakše je razmatrati logaritam ovog izraza koji ima oblik

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i \ln p_i + (1-y_i)\ln(1-p_i)).$$

Vrednost $D = -2\mathcal{L}(\boldsymbol{y}|\boldsymbol{\beta})$ naziva se **devijacija modela** i koristi se kao mera valjanosti dobijene ocene. Kod modela linearne regresije imali smo koeficijent determinacije, dok kod logističke regresije koristimo **pseudo koeficijent determinacije**, $pseudo - R^2$. On se računa kao

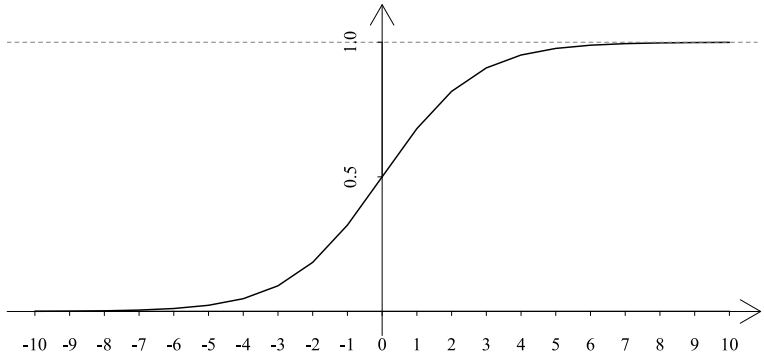
$$pseudo - R^2 = 1 - \frac{D}{D_{null}},$$

gde je D_{null} devijacija modela koji sadrži samo slobodan član (tj. samo prvu vrstu matrice $\boldsymbol{X}$, dakle bez prediktora).

Gradijent funkcije $\mathcal{L}(\boldsymbol{y}|\boldsymbol{\beta})$ u odnosu na $\boldsymbol{\beta}$ ima oblik

$$\nabla_{\boldsymbol{\beta}}\mathcal{L} = \sum_{i=1}^{n}(y_i - p_i)\boldsymbol{x}^{(i)\prime}.$$

Možemo primetiti da kod gradijenta ne figuriše sigmoidna funkcija, već on samo zavisi od razlike između opservirane vrednosti i vrednosti dobijene predviđanjem modela pomnožene opserviranim vrednostima nezavisnog vektora $\boldsymbol{x}^{(i)}$.



Slika 2.3: Sigmoidna funkcija za parametar β=0,75

2.3.2 Iterativni metod najmanjih kvadrata za ocenjivanje parametara logističke regresije

Kako smo videli u poglavlju 2.1.3, kod modela linearne regresije ocena parametara dobijena metodom maksimalne verodostojnosti jednaka je sa ocenom dobijenom metodom najmanjih kvadrata, pod pretpostavkom da šum (ostatak, "greška") ima normalnu raspodelu. Dakle, model linearne regresije ima eksplicitno rešenje za određivanje parametara metodom maksimalne verodostojnosti. Kod logističke regresije to nije slučaj zbog nelinearnosti sigmoidne funkcije. Ipak, kvadratna forma (2.5) koju smo razmatrali kod modela linearne regresije slična je onoj koja figuriše i kod modela logističke regresije. Može se pokazati da funckija $\boldsymbol{\mathcal{L}}$ ima jedinstvenu tačku maksimuma. Funkcija $\mathcal{L}$ se maksimizira u odnosu na $\boldsymbol{\beta}$ iterativnim putem koji se bazira na Njutn-Rapsonovom optimizacionom postupku

$$\boldsymbol{\beta}_{l+1} = \boldsymbol{\beta_l} + \boldsymbol{\Delta}_l, \tag{2.22}$$

gde je $\boldsymbol{\Delta}_l = \boldsymbol{H}^{-1}\nabla_{\boldsymbol{\beta}}\mathcal{L}$, a $\boldsymbol{H}$ Hesijan čiji su elementi dobijeni kao drugi izvod funkcije $\mathcal{L}$ po vektoru parametara $\boldsymbol{\beta}$. Hesijan glasi

$$\boldsymbol{H} = \nabla_{\boldsymbol{\beta}}\nabla_{\boldsymbol{\beta}}\mathcal{L} = -\sum_{i=1}^{n} \boldsymbol{x}^{(i)}\nabla_{\boldsymbol{\beta}}p_i = -\sum_{i=1}^{n} \boldsymbol{x}^{(i)}p_i(1-p_i)\boldsymbol{x}^{(i)\prime} = -\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}'.$$

Matrica $\boldsymbol{W}$ je dijagonalna, čiji su elementi $p_i(1-p_i)$ što je zapravo izvod funkcije p_i po $\boldsymbol{x}^{(i)'}\boldsymbol{\beta}$. Ova matrica se još naziva **težinska matrica**. Dakle, iterativni korak može se izračunati kao

$$\boldsymbol{\Delta}_l = (-\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}')^{-1}\boldsymbol{X}(\boldsymbol{y}-\boldsymbol{p}_l),$$

pri čemu smo sa $\boldsymbol{p}_l$ označili vektor verovatnoća p_i izračunatih za opservirane vrednosti $\boldsymbol{x}^{(i)}$ sa vektorom parametara $\boldsymbol{\beta}_l$. Podsećanja radi, kod modela linearne regresije parametre određujemo iz jednačine $\boldsymbol{\beta} = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{y}$. Dakle, možemo primetiti da je iterativni korak $\boldsymbol{\Delta}_l$ zapravo ocena parametara težinskog linearnog modela kod koga je zavisna promenljiva definisana kao razlika između opservirane i predviđene vrednost. Kako težinska matrica $\boldsymbol{W}$ nije konstanta, već zavisi od vektora parametara $\boldsymbol{\beta}_l$, ovaj metod se još naziva **iterativni težinski metod najmanjih kvadrata**.

Konvergencija jednačine (2.22) postiže se obično za manje od 10 koraka, a tvrđenje da će metod konvergirati ka rešenju daje nam sledeća teorema.

Teorema 2.8 *Postoji jedinstvena ocena parametra $\boldsymbol{\beta}$ određena kao granična vrednost jednačine (2.22) kada l teži beskonačnosti.*□

Striktan dokaz teoreme nećemo navoditi. On sledi iz činjenice da je, zbog osobina sigmoidne funkcije, $p_i(1-p_i) > 0$, pa možemo zaključiti da je Hesijan $\boldsymbol{H}$ negativno definitna matrica. Dakle funkcija $\mathcal{L}(\boldsymbol{y}|\boldsymbol{\beta})$ je konkavna za svako $\boldsymbol{\beta}$. Konvergencija ka tački maksimuma sledi iz osobina Njutn-Rapsonovog metoda.

Primer 2.10. Banci je potreban model koji će joj dati odgovor na pitanje da li će pojedini njen klijent bankrotirati ili će uspeti da izmiri svoja dugovanja. U drugoj, trećoj i četvrtoj koloni donje tabele su dati podaci o klijentima iz prethodnog perioda, gde je svaki klijent opisan sa tri nezavisne promenljive: da li je u stalnom radnom odnosu, koliko mu je dugovanje, kolika mu je prosečna plata. Na osnovu ovih podataka treba predvideti zavisnu promenljivu koja nam kaže da li je klijent bankrotirao (1 za da, 0 za ne).

Za potrebe rešavanja bančinog problema, napravićemo model logističke regresije. Kao i u prethodnim primerima, najpre ćemo učitati podatke iz tabele pozivom

```
banka<-read.csv("dataset.csv")
```

Model logističke regresije ćemo dobiti pomoću funkcije `glm` na sledeći način

```
logm.fit<-glm(banka$bankrotirao∼
   banka$sro+banka$dugovanje+banka$plata,
   family=binomial(link="logit"))
```

Bitno je da specifikujemo ulazni parametar `family`, jer nam u suprotnom funkcija `glm` može dati model linearne regresije. Proveravamo dobijeni model pozivom funkcije `summary` gde, između ostalog, dobijamo

```
> summary(logm.fit)
 Coefficients:
                  Estimate    Std.  Error  z value  Pr(>|z|)
 (Intercept)      -7.527e+00  2.268e+00    -3.319   0.000904 ***
 banka$sro        -4.220e+00  1.640e+00    -2.573   0.010072 *
 banka$dugovanje  3.996e-04   1.096e-04    3.645    0.000267 ***
 banka$plata      1.450e-04   5.983e-05    2.423    0.015404 *
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

Možemo zaključiti da su svi koeficijenti statistički značajni, pa bi jednačina modela bila

$$y = \frac{1}{1 + e^{-(-7{,}257 - 4{,}22x_1 + 0.0003996x_2 + 0.000145x_3)}},$$

gde promenljiva x_1 opisuje stalni radni odnos, x_2 visinu dugovanja i x_3 prosečnu platu. Ako uzmemo podatke iz prve vrste posmatrane tabele, vrednost koju će izabrani model logističke regresije ddati je 0,0965. Ovu vrednost možemo interpretirati kao verovatnoću da posmatrani klijent banke bankrotira.

Dobijeni koeficijenti daju nam sledeću informaciju. Recimo, koeficijent uz promenljivu x_1 nam govori da je logaritmovana vrednost šanse klijenta da bankrotira manja za 4,22 ako je klijent u stalnom radnom odnosu (za nepromenjeni nivo duga i prosečne plate).

Modelom predviđene vrednosti, dobićemo pozivom

```
> logm.fit$fitted.values
```

Ako želimo da definisani model `logm.fit` predvidi verovatnoću da dođe do bankrota za neki drugi skup klijenata, pozvali bismo funkciju `predict` na sledeći način

```
> predict(logm.fit, newdata = noviSkupPodataka, type="response")
```

Bankrotirao	U stalnom radnom odnosu	Dugovanje	Prosečna plata
0	0	8172	12106
0	0	12206	13269
1	0	15043	13965
1	0	22058	14271
1	0	15729	14930
0	0	4489	15799
1	0	17710	15976
0	0	11196	16556
0	0	9196	17492
0	0	8087	17600
0	0	5275	17637
1	0	14870	17854
1	0	18719	18077
1	0	15510	19028
1	0	17747	20360
1	0	18994	20655
1	0	11187	21848
0	0	15849	22430
0	1	11130	23810
0	1	8255	24905
0	1	18550	25211
0	1	10951	26465
1	0	19815	28128
0	1	2370	28252
1	0	15304	30004
0	1	6420	30466
1	1	17006	30489
0	1	10735	31767
0	1	14549	32189
0	1	9543	32458
0	1	7732	34353
1	0	13289	34710
0	1	25293	35704
1	1	11191	37225
0	1	11611	37469
1	1	17171	38409
0	1	7857	38463
1	1	19645	39055
0	1	6157	39376
0	1	6430	41474
1	1	19916	42133
1	1	15317	43930
0	1	7295	44362
0	1	6067	44995
1	0	17636	46227
1	0	16428	46857
0	1	9136	46907
1	1	18896	48956
0	1	2290	50500
0	1	10560	51318
1	1	7802	51657
0	1	4948	54385
1	1	15504	56274
1	0	14652	58700
0	1	4859	61566

Detaljnije testiranje preciznosti modela zahteva podelu skupa podataka na dva disjunktna podskupa, trening skup i test skup. Kada se definiše model pomoću podataka iz trening skupa, preciznost predviđanja modelom proverava se na test skupu. Ova metodologija, kao i način merenja dobijenih rezultata, spada u oblast mašinskog učenja, tako da je u ovoj knjizi nećemo razmatrati. △

Glava 3

Analiza rasipanja

Analiza rasipanja ili **analiza disperzija** ili **disperziona analiza** ili **analiza varijansi** ili **analiza odstupanja** je metod za razlučivanje ostvarenih (opserviranih) variranja (odstupanja, rasturanja,rasipanja, disperzije) u eksperimentalnim podacima u odnosu na izvore variranja. Ovaj metod se u literaturi često nalazi pod skraćenicom ANOVA što inače potiče od engleskog naziva za ovaj metod[1].

Metod se sastoji u tome što se ukupne varijacije predstavljaju kao zbir varijacija za koje se može odrediti izvor, odnosno uzrok ili faktor koji ih prouzrokuje i onih za koje se ne može odrediti izvor.

Ako već treba dovesti u vezu različite delove sveukupne varijacije sa uzročnicima, onda eksperiment mora da bude projektovan tako da se ovo povezivanje obavi na logički strog način.

Metod je razvio Fišer i izložio ga 1923. godine.

Okosnica ovog metoda je statistika Q koja se koristi u ocenjivanju disperzije obeležja:

$$Q = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

na osnovu uzorka $\mathbf{X} = (X_1, \ldots, X_n)$, gde je $\overline{X}_n$ sredina tog uzorka. Ova statistika će se u disperzionoj analizi razbijati na sume iz kojih će se određivati uzroci variranja podataka u odnosu na očekivanu vrednost.

Analiza disperzija je skup statističkih postupaka koji se bave uglavnom

[1] Analysis of Variance

analizom uticaja dejstva jednog ili više faktora na ishod eksperimenta – posmatrano obeležje. Samo ime **analiza disperzija** (ili **varijansi**) potiče otuda što se, pre svega, koriste statistike koje su zbirovi kvadrata nekih odstupanja.

Ovde će biti reči o jednofaktorskoj i dvofaktorskoj analizi.

Uvešćemo najpre nekoliko pojmova koji se koriste kod ove analize.

DEFINICIJA 3.1. *Objekti na kojima se vrše merenja se nazivaju* **eksperimentalne jedinke**. ♢

Primetimo da to mogu da budu ne samo ljudi ili životinje, već i, recimo, parcele zemljišta i sl.

DEFINICIJA 3.2. *Promenljiva veličina koja je potpuno kontrolisana od strane eksperimentatora se zove* **faktor**. *Različiti intenziteti, odnosno različite kategorije, posmatranog faktora se nazivaju* **nivoi**. ♢

Nivoi se mogu izražavati bilo kvalitativno bilo kvantitativno.

3.1 Jednofaktorski problem

U okviru ovog problema ispituje se uticaj jednog neslučajnog (kontrolisanog) faktora koji u eksperimentu ima k različitih vrednosti, $k \geq 2$, odnosno nivoa, na slučajni ishod eksperimenta Y.

Primer 3.1. Na tržištu se nude tri kvaliteta kafe: minas, santos i afrička vrsta. Statističkim postupkom, na osnovu eksperimenta sprovedenog u više gradova, treba utvrditi da li će vrsta kafe uticati na prihod ostvaren prodajom kafe (ukoliko je cena svih vrsta kafe ista) u tim gradovima.

Dakle, posmatra se jedan kvalitativni faktor – vrsta kafe (na tri različita nivoa), na slučajni ishod eksperimenta – prihod u pojedinim gradovima. Ti gradovi se posmatraju kao različite populacije.△

Primer 3.2. Eksperimentom treba utvrditi dozu leka u miligramima koju treba primenjivati u terapiji određene bolesti kod pacijenata. Pacijenti se, po pravilu, dele na grupe i na pojedinu grupu pacijenata se primenjuje ista doza leka. U ovom slučaju posmatrane grupe pacijenata predstavljaju, u statističkom smislu, različite populacije.

Ovde posmatramo kvantitativni faktor – doza leka, na slučajni ishod eksperimenta – stanje pacijenta u posmatranoj bolesti.△

Na nivou j, $1 \leq j \leq k$, uzima se prost uzorak obima n_j : (Y_{j1}, Y_{j2}, ..., Y_{jn_j}). Tako se dolazi do k nezavisnih uzoraka koji ne moraju biti istog obima. Ako za obeležje Y važi $E(Y) = m$, a pod uticajem uočenog faktora na nivou j u populaciji se dobija $E(Y_{ji}) = m_j$, $i = 1, 2, \ldots, n_j$, tada se veličina $\mu_j = m_j - m$ naziva efekat j–tog nivoa. Osim toga, pretpostavlja se da elementi uzorka Y_{ji} imaju $\mathcal{N}(m_j, \sigma^2)$ raspodelu, $j = 1, 2, \ldots, k$, za svako i, tj. kao i kod modela normalne regresije druge vrste, matematički model za jednofaktorski problem je

$$Y_{ji} = m + \mu_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots n_j, \ j = 1, 2, \ldots, k\,, \tag{3.1}$$

gde su ε_{ij} nezavisne slučajne promenljive sa istom raspodelom, naravno, ε_{ij} : $\mathcal{N}(0, \sigma^2)$. Uočimo da je u modelu sadržana i pretpostavka o konstantnosti disperzije σ^2.

S obzirom na definiciju efekta pojedinog nivoa, sledi da je $\sum_{j=1}^{k} \mu_j = 0$.

Testira se hipoteza da uočeni faktor ne utiče na obeležje Y, što se najčešće izražava preko testiranja efekata pojedinačnih nivoa uočenog faktora

$$H_0(\mu_1 = \mu_2 = \ldots = \mu_k = 0)$$

ili direktno

$$H_0(m_1 = m_2 = \ldots = m_k),$$

protiv alternative

$$H_1(\exists j; j \in \{1, 2, \ldots, k\}, \ \mu_j \neq 0)$$

odnosno

$$H_1(\exists j, l; \ j, l \in \{1, 2, \ldots, k\}, \ m_j \neq m_l).$$

Kako se matematičko očekivanje može da oceni sredinom uzorka, uvode se statistike:

- sredina celog uzorka

$$\overline{Y} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} Y_{ji} = \frac{1}{n}\sum_{j=1}^{k} n_j \overline{Y}_j, \quad n = \sum_{j=1}^{k} n_j,$$

- sredina poduzorka koji odgovara j-tom nivou

$$\overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ji}, \quad j = 1, 2, \ldots, k,$$

- totalna (ukupna) suma kvadrata

$$Q = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} Y_{ji}^2 - n\overline{Y}^2,$$

- rezidualna suma kvadrata, ili suma kvadrata ostataka, odnosno suma kvadrata grešaka

$$Q_u = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y}_j)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} Y_{ji}^2 - \sum_{j=1}^{k} n_j \overline{Y}_j^2,$$

- suma kvadrata odstupanja od hipoteze H_0 – varijacije nastale usled dejstva posmatranog faktora

$$Q_s = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_j - \overline{Y})^2 = \sum_{j=1}^{k} n_j \overline{Y}_j^2 - n\overline{Y}^2.$$

Lako je proveriti da je $Q = Q_u + Q_s$. Zaista,

$$\begin{aligned} Q &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y}_j + \overline{Y}_j - \overline{Y})^2 = \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}\Big[(Y_{ji} - \overline{Y}_j) + (\overline{Y}_j - \overline{Y})\Big]^2 = \\ &= \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y}_j)^2 + 2(\overline{Y}_j - \overline{Y})\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y}_j) + n_j(\overline{Y}_j - \overline{Y})^2\right] = \end{aligned}$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji}-\overline{Y}_j)^2 + \sum_{j=1}^{k} n_j(\overline{Y}_j-\overline{Y})^2 = Q_u + Q_s\,.$$

Statistika Q predstavlja zbir svih varijacija, Q_u–varijacije u okviru istog nivoa (zbir kvadrata odstupanja unutar nivoa–"unutrašnje" varijacije), a Q_s–varijacije među različitim nivoima (zbir kvadrata odstupanja među nivoima – "spoljašnje" varijacije). Takođe, lako je dokazati da su Q_u i Q_s nezavisne slučajne promenljive.

Pod uslovom da je nulta hipoteza tačna, statistika

$$F_{k-1,\ n-k} = \frac{(n-k)Q_s}{(k-1)Q_u}$$

ima Fišerovu raspodelu sa $(k-1, n-k)$ stepeni slobode. Ako se za realizovani uzorak $(y_{j1}, y_{j2}, \ldots, y_{jn_j})$, $j = 1, 2, \ldots, k$, realizovana vrednost statistike $F_{k-1,\ n-k}$ označi sa $f_{k-1,\ n-k}$, a kvantil reda $1-\alpha$ slučajne promenljive sa $F_{k-1,\ n-k}$ raspodelom označi sa $F_{k-1,\ n-k;\ 1-\alpha}$, najbolja kritična oblast veličine α je

H_0	H_1	C
$\mu_1 = \mu_2 = \ldots = \mu_k = 0$	$\exists j,\quad j \in \{1, 2, \ldots, k\},$ $\mu_j \neq 0$	$f_{k-1,\ n-k} \geq$ $F_{k-1,\ n-k;\ 1-\alpha}$

.

Primer 3.3. Data je tabela uspeha 20 ispitanika koji su bili podeljeni u 4 grupe približno jednakih sposobnosti. Svaku od grupa podučavao je po jedan instruktor. Po obavljenoj obuci ispitanici su bili testirani, a rezultati koje su pokazali predstavljeni su tabelom:

Instruktor	Uspeh ispitanika				
I	80	50	45	30	90
II	78	59	67	83	100
III	65	75	33	88	56
IV	72	99	51	86	23

Testirati hipotezu da različiti instruktori nisu uticali na uspeh ispitanika sa pragom značajnosti $\alpha = 0,05$.

Testira se hipoteza H_0(različiti instruktori ne utiču na uspeh ispitanika). Na osnovu datih podataka dobija se sledeća tabela suma:

Instruktor	$\sum_i y_{ji}$	$\sum_i y_{ji}^2$	n_j	$\bar{y}_j$
I	295	19925	5	59
II	387	30943	5	77,4
III	317	21819	5	63,4
IV	331	25511	5	66,2
$\sum_j$	1330	98198	20	266

Sredina celog uzorka iznosi 66,5. Takođe, dobija se da je

$$\begin{aligned} Q_s &= \left(5 \cdot 59^2 + 5 \cdot 77,4^2 + 5 \cdot 63,4^2 + 5 \cdot 66,2^2\right) - 20 \cdot 66,5^2 = \\ &= 89368,8 - 88445 = 923,8 \\ Q_u &= 98198 - 89368,8 = 8829,2 \\ Q &= 98198 - 88445 = 9753. \end{aligned}$$

Realizovana vrednost statistike $F_{k-1,n-k}$ je

$$f_{3,16} = \frac{16 \cdot 923,8}{3 \cdot 8829,2} = 0,558,$$

koja, s obzirom na to da je $F_{3,16;0,95} = 3,24$, ne pripada kritičnoj oblasti, te se nulta hipoteza prihvata sa pragom značajnosti $\alpha = 0,05$.

Pomenuta hipoteza se može testirati i upotrebom programskog jezika `R`. Uvezimo podatke iz .csv fajla

```
> tabela <- read.csv("primer.csv", header = TRUE)
```

Prva kolona, `tabela$Instruktor`, sadrži vrednosti nezavisne promenljive i ona mora biti kategorijskog tipa. Najpre, treba ispitati da li je kolona `tabela$Instruktor` kreirana na adekvatan način, pa ćemo pozvati funkciju

```
> is.factor(tabela$Instruktor)
[1] TRUE
```

Kako je rezultat `TRUE`, možemo nastaviti sa radom. Pomoću funkcije `aov` napravićemo objekat `av.test` koji predstavlja ANOVA model

```
av.test = aov(tabela$Uspeh ~ tabela$Instruktor)
```

Rezultat možemo videti pozivom funkcije `summary`

```
> summary(av.test)
```

```
                    Df  Sum Sq  Mean Sq  F value  Pr(>F)
 tabela$Instruktor   3   924     307.9    0.558    0.65
 Residuals          16  8829     551.8
```

Dobili smo da je realizovana vrednost statistike (`F value`) 0,558, a p-vrednost (`Pr(>F)`) 0,65, pa možemo zaključiti da se prihvata nulta hipoteza. Ako želimo da izračunamo granicu kritične oblasti za $\alpha = 0,05$, to možemo učiniti pozivom funkcije

```
> qf(p = 0.95, df1 = 3, df2 = 16)
[1] 3.238872
```

Napomena 1: Treba proveriti da li dobijeni ANOVA model zadovoljava polaznu pretpostavku da je $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Prikažimo grafički vrednosti ε_{ij} za svaku vrednost dobijenu predviđanjem modelom. Pozvaćemo funkciju `residualPlot`, koja je implementirana na sledeći način

```
residualPlot<-function(model){
   par(family="serif", cex=1.25)
   plot(x=c(1:length(model$residuals)), y=model$residuals,
      xlab="Ispitanik", ylab=expression(epsilon))
   abline(h=0)
}
```

Funkciju ćemo pozvati kao

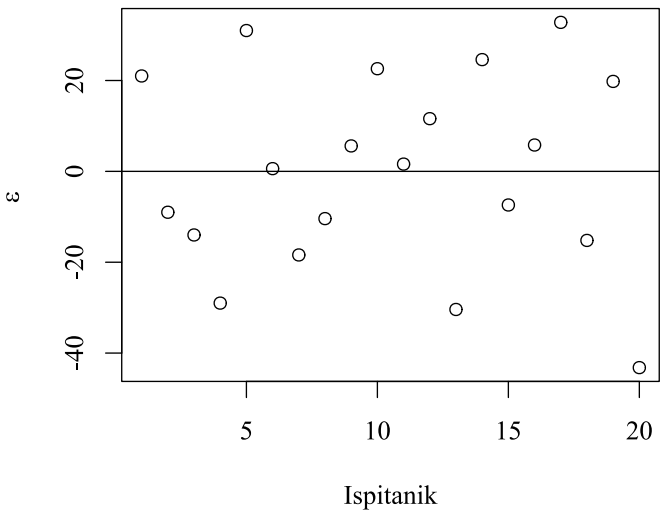
```
> residualPlot(av.test)
```

i dobićemo dijagram dat na slici 3.1. Možemo uočiti konstantnu simetričnu varijaciju grešaka u okolini nule, što ukazuje da je model adekvatan. Kako bismo potvrdili valjanost ANOVA modela, testiraćemo nultu hipotezu da reziduali imaju normalnu raspodelu, a za testiranje koristimo Kolmogorov–Smirnov test čije rezultate dobijamo pozivom

```
> ks.test(av.test$residuals, "pnorm", 0, sd(av.test$residuals))

   One-sample Kolmogorov-Smirnov test
data:  av.test$residuals
D = 0.12082, p-value = 0.8989
alternative hypothesis:  two-sided
```

Kako je p-vrednost veća od praga značajnosti prihvatamo nultu hipotezu i možemo zaključiti da je dobijeni model dobar, a samim tim i zaključci koji iz



Slika 3.1: Razlike između stvarnih vrednosti i vrednosti dobijenih predviđanjem modelom.

njega slede.

Napomena 2: U svrhu testiranja nulte hipoteze o normalnosti ostataka, možemo koristiti i Šapiro–Uilkov[2] test koji je baš namenjen proveri da li posmatrano obeležje ima normalnu raspodelu. On postiže veću moć u odnosu na test Kolmogorov–Smirnova za isti prag značajnosti. Pored toga, u ovom primeru je uzorak relativno malog obima, a poznato je da u takvim slučajevima, kada se na osnovu uzorka (kao što je ovde slučaj sa disperzijom ostataka) vrši ocenjivanje nepoznatih parametara na osnovu istog uzorka, zaključivanje na osnovu testa Kolmogorov–Smirnova može da bude nepouzdano. To bi se dogodilo iz tog razloga što bi ocenjivanje moglo da prouzrokuje pomeranje granice kritične oblasti.

Rezultat Šapiro–Uilkovog testa primenom programskog jezika `R`, dobijamo pozivom

```
> shapiro.test(av.test$residuals)
Shapiro-Wilk normality test
```

[2]Shapiro–Wilk

```
data:  av.test$residuals
W = 0.96796, p-value = 0.7114
```

Kako je p-vrednost veća od praga značajnosti, prihvatamo nultu hipotezu i možemo zaključiti da je dobijeni model adekvatan. △

Pretpostavka o normalnoj raspodeli je suštinska za primenu disperzione analize. Međutim, u ovom primeru nije vršeno prethodno testiranje takve hipoteze s obzirom na opšteprihvaćenu činjenicu da kognitivna (saznajna) obeležja (za kakvo se može smatrati korišćeno u primeru) imaju normalnu raspodelu. Slično će biti i u nekoliko narednih primera.

Ako se statističkom analizom dođe do zaključka da treba odbaciti nultu hipotezu, tada treba izvršiti testiranje pojedinačnih hipoteza po svim parovima nivoa:

$$H_0(m_j = m_l) \quad \text{protiv} \quad H_1(m_j \neq m_l) \text{ za } j, l \in \{1, 2, \ldots, k\},\ j \neq l,$$

postupkom testiranja jednakosti matematičkih očekivanja kod dva obeležja sa normalnim raspodelama koja imaju jednake disperzije, da bi se utvrdilo koji nivoi posmatranog faktora utiču na obeležje Y. Treba napomenuti da postoje i preciznije metode za proveru ovih relacija od uobičajenog t–testa.

Primer 3.4. Zgrada jednog fakulteta se nalazi u jednoj prometnoj ulici. Neke njene učionice su okrenute ulici, dok su druge okrnute dvorištu koje je tiho. Eksperiment koji je vršen, imao je za cilj da utvrdi da li bučnost učionice može da utiče na pamćenje studenata. Eksperiment je vršen sa istom grupom od ukupno 8 studenata, s tim što je ta grupa studenata premeštana tako da je jednom bila u tihoj, a drugi put u bučnoj učionici. Svi studenti su dobili da uče iste tekstove. U tihoj i bučnoj učionici su studentima zadavani tekstovi približno istih težina. Tekst koji su učili je trebalo da reprodukuju posle istog utvđenog vremena. Dobijeni su sledeći rezultati procenata zapamćenog materijala:

Tip učionice	Procenat zapamćenog materijala							
Tiha	88	75	68	59	65	62	92	81
Bučna	50	54	62	48	25	36	32	60

.

Testirati hipotezu da buka ne utiče na pamćenje studenata sa pragom značajnosti $\alpha = 0,05$.

Dakle, nulta hipoteza je H_0(Buka ne utiče na pamćenje studenata.). Na osnovu datih podataka dobija se da je $Q_s = 3108,06$, $Q_u = 2328,38$ i $Q = 5436,44$. Realizovana vrednost test statistike je

$$f_{1,14} = 18,688.$$

S obzirom na to da je $F_{1,14;0,95} = 4,60$, to realizovana vrednost pripada kritičnoj oblasti, te se nulta hipoteza odbacuje sa pragom značajnosti $\alpha = 0,05$. Dakle, buka utiče na pamćenje (pa se dalje formalno proverava da li na smanjenje ili na poboljšanje pamćenja), ali s obzirom da su ispitivana samo dva nivoa faktora uticaja, nema drugih parova da bi bilo potrebe produžiti testiranje po parovima.

Rešimo sada ovaj primer upotrebom programskog jezika `R`. Slično kao u prethodnom primeru, uvezimo podatke iz .csv fajla u objekat `tabela`, a zatim ANOVA model dobijamo pozivom

```
> av.test<-aov(tabela$Zapamceno ~ tabela$Ucionica)
> summary(av.test)
                  Df  Sum Sq  Mean Sq  F value     Pr(>F)
 tabela$Ucionica   1    3108    3108.1    18.69   0.000702 ***
 Residuals        14    2328     166.3
```

Zaključujemo da tip učionice ima uticaj na procenat zapamćenog materijala kod studenata.

Napomena: Da je bilo više od dva tipa učionice, testiranje bismo obavili pomoću Takijevog HSD[3] testa, čiji se rezultati dobijaju pozivom funkcije

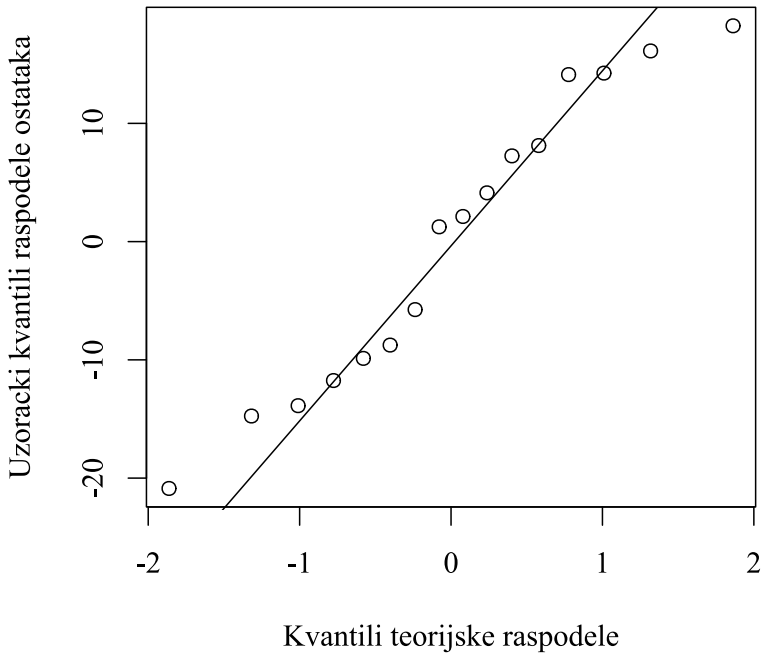
```
> TukeyHSD(av.test, conf.level = 0.95)
```

pri čemu smo za prag značajnosti testa uneli $\alpha = 0.05$.

Takijev HSD test je test značajnosti i služi za poređenje svih parova m_i i m_j, $i, j = 1, \ldots, n$, tako što posmatra razliku između najveće i najmanje uzoračke sredine po nivoima u odnosu na uzoračku standardnu devijaciju sume svih očekivanja.

U Primeru 3.3 naveli smo dva načina kako da u programskom jeziku `R`

[3]Tuckey's HSD (honestly significant difference) test



Slika 3.2: Q-Q dijagram reziduala modela u poređenju sa normalnom raspodelom.

proverimo adekvatnost modela za posmatrani uzorak. Sada ćemo iskoristiti **Q-Q dijagram** u tu svrhu, koji se može generisati funkcijom

```
qqresidualPlot<-function(model){
   par(family="serif", cex=1.25)
   qqnorm(model$residuals, xlab = "Kvantili teorijske raspodele",
      ylab = "Uzoracki kvantili raspodele ostataka", main = "")
   qqline(model$residuals)
}
```

preko poziva

```
> qqresidualPlot(av.test)
```

Q-Q dijagram je grafički metod za proveru da li dva skupa podataka imaju

istu raspodelu. Kod provere da li posmatrano obeležje iz koga je uzet uzorak ima određenu (teorijsku) raspodelu, upoređuju se teorijski kvantili pretpostavljene raspodele sa uzoračkim kvantilima ispitivane raspodele. Dakle, u našem slučaju proveravamo saglasnost raspodele reziduala (ostataka) sa normalnom raspodelom. Što su tačke na grafiku dobijene na osnovu uzoračkih kvantila reziduala, recimo definisane koordinatama (a, b), bliže pravoj liniji $x = y$, to je raspodela testiranog obeležja bliža zadatoj teorijskoj raspodeli. Pri tome je a kvantil reda p teorijske raspodele sa kojom se upoređuje raspodela posmatranog obeležja, a b uzorački kvantil istog reda p posmatranog obeležja. Prava $x = y$ je, ukoliko se radi o raspodeli apsolutno neprekidnog tipa, nastala kada se i na ordinatnu osu takođe nanesu teorijski kvantili, kod nas, normalne raspodele i dobijene tačke, (M_p, M_p), za kvantile reda p teorijske raspodele, spoje. Dakle, možemo zaključiti sa slike 3.2 da reziduali uglavnom prate normalnu raspodelu. $\triangle$

3.2 Dvofaktorski problem

3.2.1 Dvofaktorski problem na prostom uzorku

Dvofaktorski problem nastaje kada treba da se ispita uticaj dva faktora, recimo A i B, na obeležje Y. Neka se uticaj faktora A ispituje na k ($k \geq 2$) nivoa, a faktora B na l ($l \geq 2$) nivoa. Prost slučajni uzorak, u smislu analize rasipanja, nastaje kada se na svakom ukrštenom nivou dva posmatrana faktora uzima, tj. posmatra, tačno jedan element. Dakle, obim takvog uzorka je $k \times l$ i takav uzorak se najčešće predstavlja dvodimenzionalno, tabelom:

$A \downarrow \backslash B \rightarrow$	1	2	...	l	
1	Y_{11}	Y_{12}	...	Y_{1l}	$Y_{1\bullet}$
2	Y_{21}	Y_{22}	...	Y_{2l}	$Y_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
k	Y_{k1}	Y_{k2}	...	Y_{kl}	$Y_{k\bullet}$
	$Y_{\bullet 1}$	$Y_{\bullet 2}$	...	$Y_{\bullet l}$	$Y_{\bullet\bullet}$

gde je Y_{ij} vrednost obeležja Y na elementu uzorka izloženom i–tom nivou faktora A i j–tom nivou faktora B, $Y_{i\bullet} = \sum_{j=1}^{l} Y_{ij}$, $Y_{\bullet j} = \sum_{i=1}^{k} Y_{ij}$, $Y_{\bullet\bullet} =$

$\sum_{i=1}^{k}\sum_{j=1}^{l} Y_{ij}$. Dakle, svaki ukršteni nivo (i,j) ispitivanih faktora primenjuje se na tačno jedan element uzorka.

Za definisanje modela koriste se sledeće oznake. Kao i kod jednofaktorskog problema, $m = E(Y)$, a zatim, $m_{i\bullet} = E(Y_{ij})$, $j = 1, 2, \ldots, l$, $m_{\bullet j} = E(Y_{ij})$, $i = 1, 2, \ldots, k$, tj. $m_{i\bullet}$ je matematičko očekivanje obeležja Y u populaciji koja je od faktora A izložena samo i–tom nivou, a $m_{\bullet j}$ je matematičko očekivanje obeležja Y u populaciji koja je od faktora B izložena samo j–tom nivou. Sa $\mu_i = m_{i\bullet} - m$ označava se efekat i–tog nivoa faktora A, a sa $\nu_j = m_{\bullet j} - m$ efekat j–tog nivoa faktora B.

Matematički linearni model dvofaktorske analize disperzija na prostom uzorku je

$$Y_{ij} = m + \mu_i + \nu_j + \varepsilon_{ij},$$

gde su ε_{ij} nezavisne identički raspodeljene $\mathcal{N}(0, \sigma^2)$ slučajne promenljive, pri čemu je σ^2 nepoznato. U okviru ovog modela testiraju se sledeće nulte hipoteze:

- $H_{0A}\,(\mu_1 = \mu_2 = \ldots = \mu_k = 0)$ — efekti nivoa faktora A na obeležje Y su bez bitnih razlika, tj. faktor A ne utiče na ishod eksperimenta;

- $H_{0B}\,(\nu_1 = \nu_2 = \ldots = \nu_l = 0)$ — efekti nivoa faktora B na obeležje Y su bez bitnih razlika, tj. faktor B ne utiče na ishod eksperimenta;

- $H_{0AB}\,(\mu_1 = \mu_2 = \ldots = \mu_k = \nu_1 = \nu_2 = \ldots = \nu_l = 0)$ — efekti nivoa oba posmatrana faktora A i B na slučajni ishod eksperimenta Y su bez bitnih razlika, tj. nijedan od posmatranih faktora A i B ne utiče na ishod eksperimenta;

protiv alternativnih redom:

- $H_{1A}\,(\exists i,\ i \in \{1, 2, \ldots, k\},\ \mu_i \neq 0)$

- $H_{1B}\,(\exists j,\ j \in \{1, 2, \ldots, l\},\ \nu_j \neq 0)$

- $H_{1AB}\,(\exists (i,j)$ tako da $(i,j) \in \{1, 2, \ldots, k\} \times \{1, 2, \ldots, l\}, (\mu_i, \nu_j) \neq (0, 0))$.

Za sprovođenje testova potrebne su sledeće statistike:

- sredina (celog) uzorka

$$\overline{Y} = \frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l} Y_{ij},$$

- sredina poduzorka koji odgovara i-tom nivou faktora A

$$\overline{Y}_{i\bullet} = \frac{1}{l}\sum_{j=1}^{l} Y_{ij},$$

- sredina poduzorka koji odgovara j-tom nivou faktora B

$$\overline{Y}_{\bullet j} = \frac{1}{k}\sum_{i=1}^{k} Y_{ij},$$

- ukupna suma kvadrata odstupanja od srednje vrednosti, odnosno, uzoračke sredine celog uzorka

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{l}(Y_{ij} - \overline{Y})^2,$$

- suma kvadrata odstupanja za faktor A

$$Q_A = l\sum_{i=1}^{k}(\overline{Y}_{i\bullet} - \overline{Y})^2,$$

- suma kvadrata odstupanja za faktor B

$$Q_B = k\sum_{j=1}^{l}(\overline{Y}_{\bullet j} - \overline{Y})^2,$$

- slučajna suma kvadrata

$$Q_S = Q - Q_A - Q_B.$$

Pod uslovom da su nulte hipoteze tačne, statistike

$$F_{k-1,(k-1)(l-1)} = \frac{(k-1)(l-1)Q_A}{(k-1)Q_S},$$

$$F_{l-1,(k-1)(l-1)} = \frac{(k-1)(l-1)Q_B}{(l-1)Q_S} \quad \text{i}$$

$$F_{k+l-2,(k-1)(l-1)} = \frac{(k-1)(l-1)(Q_A+Q_B)}{(k+l-2)Q_S}$$

imaju Fišerove raspodele sa naznačenim brojem stepeni slobode. Ako se, kao i kod jednofaktorske analize, koriste oznake

$$f_{k-1,(k-1)(l-1)}\,, \quad f_{l-1,(k-1)(l-1)} \quad \text{i} \quad f_{k+l-2,(k-1)(l-1)}$$

za realizovane vrednosti poslednjih statistika redom, najbolje kritične oblasti veličine α su:

H_0	H_1	C
$\mu_1 = \mu_2 = \ldots$ $\ldots = \mu_k = 0$	$\exists i, i \in \{1,2,\ldots,k\}$, $\mu_i \neq 0$	$f_{k-1,(k-1)(l-1)} \geq$ $F_{k-1,(k-1)(l-1);\ 1-\alpha}$
$\nu_1 = \nu_2 = \ldots$ $\ldots = \nu_l = 0$	$\exists j, j \in \{1,\ldots,l\}$, $\nu_j \neq 0$	$f_{l-1,(k-1)(l-1)} \geq$ $F_{l-1,(k-1)(l-1);\ 1-\alpha}$
$\mu_1 = \ldots = \mu_k =$ $= \nu_1 = \ldots = \nu_l = 0$	$\exists (i,j), (i,j) \in$ $\{1,\ldots,k\} \times \{1,\ldots,l\}$ $\wedge(\mu_i,\ \nu_j) \neq (0,0)$	$f_{k+l-2,(k-1)(l-1)} \geq$ $F_{k+l-2,(k-1)(l-1);\ 1-\alpha}$

.

Primer 3.5. Iskoristićemo eksperiment iz prethodnog primera i usložniti ga dalje, tako što će svako od studenata morati da pamti četiri vrste teksta u svakoj od dve učionice. Ispituje se da li buka i vrsta teksta utiču na pamćenje studenta. Dobijeni su sledeći procenti zapamćenog materijala:

Učionica \ Materijal	Besmisleni slogovi	Proza	Poezija	Formule
Tiha	58	85	73	61
Bučna	25	48	52	28

.

Testiraćemo hipotezu da buka ne utiče na pamćenje studenata, hipotezu da vrsta materijala ne utiče na pamćenje studenata i hipotezu da buka i vrsta

materijala ne utiču na pamćenje studenata sve sa pragom značajnosti $\alpha = 0,05$.

Realizovane vrednosti odgovarajućih F statistika su:

Suma kvadrata	f
$Q_A = 1922$	80,083
$Q_B = 949,5$	13,188
$Q_S = 72$	///

Za prag značajnosti 0,05 granica kritične oblasti je kvantil $F_{1,3;0,95} = 10,1$. S obzirom na realizovanu vrednost test statistike 80,083 , H_{0A} se odbacuje, tj. može se zaključiti da buka utiče na pamćenje studenata.

Što se tiče druge nulte hipoteze vezane za vrstu teksta, zna se da je $F_{3,3;0,95} = 9,28$ i kako je realizovana vrednost test statistike jednaka 13,188 , to se i druga nulta hipoteza odbacuje, tj. može da se zaključi da pamćenje studenata zavisi i od vrste materijala koji se uči.

Za treću hipotezu je $F_{4,3;0,95} = 9,12$ i kako je realizovana vrednost odgovarajuće test statistike jednaka 29,911 , to se i treća nulta hipoteza odbacuje, tj. zaključuje se da buka i vrsta materijala koji se uči itekako utiču na pamćenje studenata.

Zadržaćemo se ovde kratko na redosledu kojim smo vršili testiranje. U radu smo koristili redosled navođenja hipoteza u prethodnom tekstu, međutim, u praksi treba činiti upravo obrnuto. Dakle, kod modela dvofaktorske analize na prostom uzorku, najpre treba proveriti hipotezu H_{0AB}, jer njenim prihvatanjem bismo istovremeno utvrdili da će i H_{0A} i H_{0B} biti prihvaćene. Drugim rečima, H_{0A} i H_{0B} se proveravaju tek pošto se utvrdi da H_{0AB} treba odbaciti.

Iskoristimo sada programski jezik `R` kako bismo rešili ovaj primer. Kao i u ranijim primerima, u objekat `tabela` uvezimo podatke iz gornje tabele koja je struktuirana na sledeći način:

```
> tabela
```

```
 Ucionica  Materijal  Zapamceno
   tiha     slogovi       58
   tiha      proza        85
   tiha     poezija       73
    ...
  bucna     poezija       55
  bucna     formule       23
```

Ispitujemo da li faktori `tabela$Ucionica` i `tabela$Materijal` utiču na procenat zapamćenog materijala tako što formiramo ANOVA model pozivom funkcije

```
> av.test=aov(tabela$Zapamceno ~ tabela$Ucionica+tabela$Materijal)
```

čiji rezultat dobijamo kao

```
> summary(av.test)
                    Df  Sum Sq  Mean Sq  F value    Pr(>F)
 tabela$Ucionica     1  1922.0   1922.0    80.08  0.00294 **
 tabela$Materijal    3   949.5    316.5    13.19   0.03109 *
 Residuals           3    72.0     24.0
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

Iz kolone `F value` možemo pročitati realizovane vrednosti test statistike, dok p-vrednosti čitamo iz kolone `Pr(>F)`. Zaključujemo da, za prag značajnosti $\alpha = 0,05$, oba faktora imaju uticaj na procenat zapamćenog materijala. Međutim, ako bismo prag značajnosti podigli na $\alpha = 0,01$ zaključili bismo da vrsta teksta (`tabela$Materijal`) nema uticaj na procenat zapamćenog materijala.

Rezultat testiranja treće hipoteze, H_{0AB}, izračunaćemo pomoću vrednosti iz kolona `Df` i `Sum Sq`. Tako dobijamo da je $\frac{3 \cdot 1 \cdot (1922+949,5)}{(3+1) \cdot 72} = 29,91$, dok je odgovarajući kvantil za $\alpha = 0,05$

```
> qf(p=0.95, df1=4, df2=3)
[1] 9.117182
```

pa zaključujemo da treba odbaciti nultu hipotezu, H_{0AB}. $\triangle$

Dvofaktorski problem ispitivan na prostom uzorku ne omogućava ispitivanje međuzavisnosti dva ispitivana faktora. Testiranje međuzavisnosti dva ispitivana kontrolisana faktora vrši se na uzorku sa ponavljanjem.

3.2.2 Dvofaktorski problem na uzorku sa ponavljanjem

Uzorak dvofaktorskog problema koji ima više elemenata koji odgovaraju svakom uređenom paru nivoa (i, j) posmatranih faktora je uzorak sa ponavljanjem. Pri tome broj elemenata (ponavljanja) ne mora biti jednak u svakoj "ćeliji", odnosno na svakom ukrštenom nivou (i, j) posmatranih faktora. Tada je prikaz uzorka u tabeli trodimenzionalan.

Mi ćemo se ovde baviti samo slučajem jednakog broja elemenata uzorka u svakoj ćeliji:

$A\downarrow \backslash B\rightarrow$	1	2	...	l
1	Y_{111}	Y_{121}	...	Y_{1l1}
	Y_{112}	Y_{122}	...	Y_{1l2}
	$\vdots$	$\vdots$	...	$\vdots$
	Y_{11r}	Y_{12r}	...	Y_{1lr}
2	Y_{211}	Y_{221}	...	Y_{2l1}
	Y_{212}	Y_{222}	...	Y_{2l2}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	Y_{21r}	Y_{22r}	...	Y_{2lr}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
k	Y_{k11}	Y_{k21}	...	Y_{kl1}
	Y_{k12}	Y_{k22}	...	Y_{kl2}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	Y_{k1r}	Y_{k2r}	...	Y_{klr}

.

Koristićemo oznaku $E(Y_{ija}) = m_{ij}$, $a = 1, 2, \ldots, r$, a međusobni efekat faktora A na nivou i i faktora B na nivou j predstavljaćemo sa

$$\eta_{ij} = m_{ij} - (m + \mu_i + \nu_j)$$

i važi

$$\sum_{i=1}^{k} \mu_i = \sum_{j=1}^{l} \nu_j = \sum_{i=1}^{k} \eta_{ij} = \sum_{j=1}^{l} \eta_{ij} = 0\,.$$

Matematički linearni model dvofaktorske analize disperzija na uzorku sa

ponavljanjem je

$$Y_{ija} = m + \mu_i + \nu_j + \eta_{ij} + \varepsilon_{ija}\,,$$

gde su ε_{ija} nezavisne identički raspodeljene $\mathcal{N}(0, \sigma^2)$ slučajne promenljive, pri čemu se podrazumeva da je σ^2 nepoznato. U okviru ovog modela testiraju se sledeće nulte hipoteze:

- $H_{0A}\,(\mu_1 = \mu_2 = \ldots = \mu_k = 0)$ — efekti nivoa faktora A na obeležje Y su bez bitnih razlika, tj. faktor A ne utiče na ishod eksperimenta;
- $H_{0B}\,(\nu_1 = \nu_2 = \ldots = \nu_l = 0)$ — efekti nivoa faktora B na obeležje Y su bez bitnih razlika, tj. faktor B ne utiče na ishod eksperimenta;
- $H_{0AB}\,(\eta_{ij} = 0,\ \forall(i,j),\ i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, l)$ — nema interaktivnog dejstva faktora A i B na obeležje Y.

protiv alternativnih redom:

- $H_{1A}\,(\exists i,\ i \in \{1, 2, \ldots, k\},\ \mu_i \neq 0)$
- $H_{1B}\,(\exists j,\ j \in \{1, 2, \ldots, l\},\ \nu_j \neq 0)$
- $H_{1AB}\,(\exists(i,j),\ i \in \{1, 2, \ldots, k\},\ j \in \{1, 2, \ldots, l\},\ \eta_{ij} \neq 0.)$

Za sprovođenje testova definišu se sledeće statistike:

- uzoračka sredina celog uzorka

$$\overline{Y} = \frac{1}{k \cdot l \cdot r} \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} Y_{ija},$$

- uzoračka sredina ćelije (i, j)

$$\overline{Y}_{ij} = \frac{1}{r} \sum_{a=1}^{r} Y_{ija},$$

- uzoračka sredina na nivou i faktora A

$$\overline{Y}_{i\bullet} = \frac{1}{l \cdot r} \sum_{j=1}^{l} \sum_{a=1}^{r} Y_{ija},$$

- uzoračka sredina na nivou j faktora B

$$\overline{Y}_{\bullet j} = \frac{1}{k \cdot r} \sum_{i=1}^{k} \sum_{a=1}^{r} Y_{ija},$$

- ukupna suma kvadrata odstupanja od srednje vrednosti celog uzorka

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} (Y_{ija} - \overline{Y})^2,$$

- suma kvadrata odstupanja za faktor A

$$Q_A = \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} (\overline{Y}_{i\bullet} - \overline{Y})^2 = l \cdot r \sum_{i=1}^{k} (\overline{Y}_{i\bullet} - \overline{Y})^2,$$

- suma kvadrata odstupanja za faktor B

$$Q_B = \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} (\overline{Y}_{\bullet j} - \overline{Y})^2 = k \cdot r \sum_{j=1}^{l} (\overline{Y}_{\bullet j} - \overline{Y})^2,$$

- suma kvadrata interaktivnog dejstva faktora A i B

$$\begin{aligned} Q_I &= \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} (\overline{Y}_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y})^2 = \\ &= r \sum_{i=1}^{k} \sum_{j=1}^{l} (\overline{Y}_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y})^2, \end{aligned}$$

- slučajna suma kvadrata

$$Q_S = \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{a=1}^{r} (Y_{ija} - \overline{Y}_{ij})^2.$$

Očigledno je

$$Q = Q_A + Q_B + Q_I + Q_S.$$

Pod uslovom da su nulte hipoteze tačne, statistike

$$F_{k-1,kl(r-1)} = \frac{kl(r-1)Q_A}{(k-1)Q_S},$$

$$F_{l-1,kl(r-1)} = \frac{kl(r-1)Q_B}{(l-1)Q_S} \quad \text{i}$$

$$F_{(k-1)(l-1),kl(r-1)} = \frac{kl(r-1)Q_I}{(k-1)(l-1)Q_S}$$

imaju Fišerove raspodele sa naznačenim brojem stepeni slobode. Za realizovane vrednosti poslednjih statistika, najbolje kritične oblasti veličine α su date narednom tabelom.

H_0	H_1	C
$\mu_1 = \mu_2 = \ldots$ $\ldots = \mu_k = 0$	$\exists i, i \in \{1,2,\ldots,k\}$, $\mu_i \neq 0$	$f_{k-1,kl(r-1)} \geq$ $F_{k-1,kl(r-1);\ 1-\alpha}$
$\nu_1 = \nu_2 = \ldots$ $\ldots = \nu_l = 0$	$\exists j, j \in \{1,2,\ldots,l\}$, $\nu_j \neq 0$	$f_{l-1,kl(r-1)} \geq$ $F_{l-1,kl(r-1);\ 1-\alpha}$
$\forall(i,j)$, $i = 1,2,\ldots,k$; $j = 1,2,\ldots,l$, $\eta_{ij} = 0$	$\exists(i,j)$, $i \in \{1,2,\ldots,k\}$, $j \in \{1,2,\ldots,l\}$, $\eta_{ij} \neq 0$	$f_{(k-1)(l-1),kl(r-1)} \geq$ $F_{(k-1)(l-1),kl(r-1);\ 1-\alpha}$

.

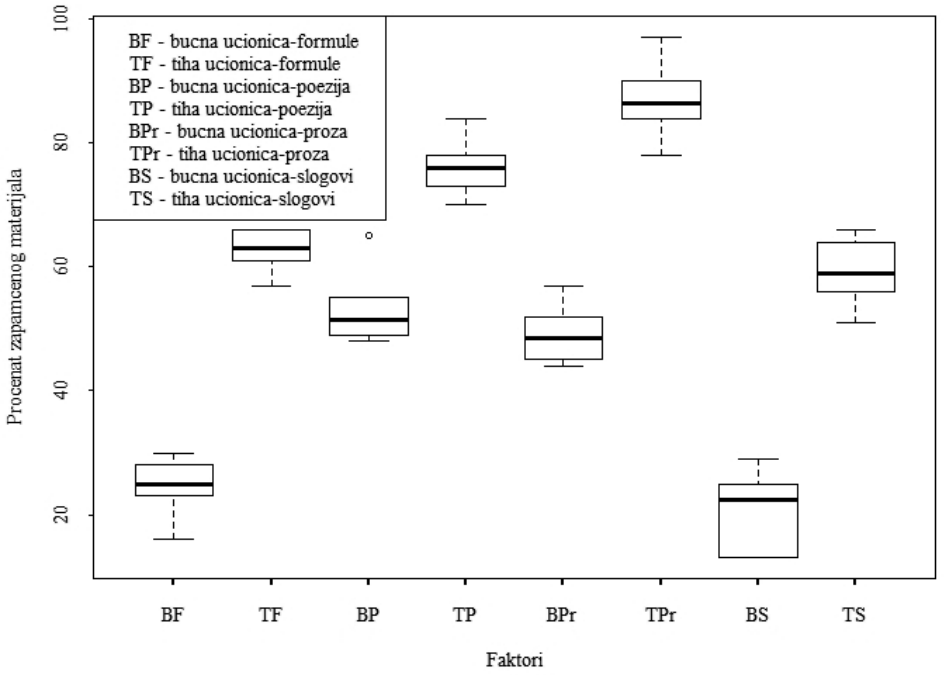
Ako se testiranje sprovodi bez upotrebe računara, formiraju se tabele za postepeno izračunavanje pomenutih suma kvadrata.

Moguće je analizirati disperzije i kada se ispituje istovremeni uticaj dva faktora, a u svakoj ćeliji se ne nalazi isti broj podataka (elemenata uzorka).

Disperziona analiza se primenjuje i kod ispitivanja istovremenog uticaja više od dva faktora uticaja.

Na kraju naglasimo još jednom da se, u slučaju odbacivanja bilo koje od postavljenih nultih hipoteza, postupak testiranja po pravilu nastavlja. Ovo s toga što je odbacivanjem nulte hipoteze konstatovano da jedan ili više neslučajnih faktora utiče na ishod eksperimenta, pa je dalje od interesa utvrditi nivo na kome posmatrani faktor utiče na ishod eksperimenta. Postupak za ovo testiranje je već objašnjen kod jednofaktorskog problema.

Primer 3.6. Pretpostavimo da smo testiranje iz Primera 3.5 vršili svakog



Slika 3.3: Kutija dijagrami prikupljenih podataka o procentu zapamćenog materijala.

dana od ponedeljka do subote. Dakle, imamo po šest opservacija za svaki par faktora (tip učionice i vrsta teksta). Podaci su prikazani kutija dijagramom na slici 3.3. Testiraćemo hipoteze navedene u Primeru 3.5.

	Učionica \ Materijal	Besmisleni slogovi	Proza	Poezija	Formule
Ponedeljak	Tiha	58	85	73	61
	Bučna	25	48	52	28
Utorak	Tiha	51	90	70	64
	Bučna	23	44	48	30
Sreda	Tiha	60	88	75	66
	Bučna	22	49	55	23
Četvrtak	Tiha	56	97	78	62
	Bučna	13	45	65	16
Petak	Tiha	66	84	84	66
	Bučna	13	57	51	26
Subota	Tiha	64	78	77	57
	Bučna	29	52	49	24

Na slici 3.3 pored srednje vrednosti i standardne devijacije primećujemo i da postoji autlajer kod podataka BP.

Kutija dijagram na slici 3.3 generisan je funcijom `niceBoxplot` koju smo definisali na sledeći način

```
niceBoxplot<-function(){
   par(family="serif", cex=0.85)
   boxplot(tabela$Zapamceno ~ tabela$Ucionica + tabela$Materijal +
   tabela$Materijal:tabela$Ucionica, xaxt="n")
   axis(1, at=seq(1, 8, by=1), lwd=0, lwd.ticks=2, labels = c(
      "BF","TF","BP","TP","BPr","TPr","BS","TS"))
   title(sub="Faktori",line=3,
      ylab="Procenat zapamcenog materijala")
   legend("topleft",inset = c(-0.001,-0.001), xjust=1, legend=c(
      "BF - bucna ucionica-formule", "TF - tiha ucionica-formule",
      "BP - bucna ucionica-poezija", "TP - tiha ucionica-poezija",
      "BPr - bucna ucionica-proza", "TPr - tiha ucionica-proza",
      "BS - bucna ucionica-slogovi", "TS - tiha ucionica-slogovi"))
}
```

Tačne vrednosti svih autlajera dobićemo pozivom funkcije

```
> boxplot(tabela$Zapamceno ~ tabela$Ucionica*tabela$Materijal)$out
[1] 65
```

U našem primeru to je samo jedna opservacija čija je vrednost 65. Nećemo se baviti izbacivanjem autlajera, pa nastavljamo test sa celim skupom podataka. Napravimo sada ANOVA model i sumirajmo rezultate

```
> av.test=aov(tabela$Zapamceno ~ tabela$Ucionica*tabela$Materijal)
> summary(av.test)
                                Df  Sum Sq  Mean Sq  F value    Pr(>F)
 tabela$Ucionica                 1   14111    14111   483.461  <2e-16 ***
 tabela$Materijal                3    7420     2473    84.744  <2e-16 ***
 tabela$Ucionica:tabela$Materijal  3    526      175     6.002  0.00178 **
 Residuals                      40   1168       29
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

Posmatrajući kolonu `Pr(>F)`, možemo zaključiti da buka i vrsta materijala utiču na procenat zapamćenog teksta, a takođe i svaki od ovih faktora pojedinačno ima značajan uticaj. Primetimo da je u pozivu funkcije znak "*" a ne "+", jer želimo da testiramo i zajednički uticaj oba faktora. △

Glava 4

Plan uzorka sa slučajnim blokovima

Do sada smo koristili isključivo uzorke *potpuno slučajnog plana*. Sada ćemo se upoznati sa osnovnim pojmovima u vezi sa *planovima uzoraka sa slučajnim blokovima*, kao i elementarnim načinima njihovog korišćenja. Za ovaj plan uzorka se često koristiti, jednostavno, naziv *slučajni blokovi* ili *uzorak blokova* ili *blokovski uzorak*.

Da bismo približili čitaocu primenu ovakvog uzorka, vratimo se na Primer 3.3. U tom primeru je navedeno da su ispitanici bili podeljeni u četiri grupe približno istih sposobnosti. Takva homogenost unutar grupa je bila ključna za primenu jednofaktorske analize rasipanja. Međutim, iako zvuči dosta jednostavno, takva podela na grupe nije uvek moguća, a i zahteva ozbiljan rad pre same primene eksperimenta. Da bi se te komplikacije izbegle, primenjuje se plan uzorka sa slučajnim blokovima. Izložićemo nadalje u čemu se on sastoji.

U eksperimentu se često posmatra istovremeno više faktora, kao što je bilo reči u prethodnoj glavi. Zbog toga se uvodi i sledeći termin.

Definicija 4.1. **Tretman** *čini specifična kombinacija nivoa posmatranih faktora.* $\diamondsuit$

Različite tretmane ćemo označavati nadalje sa T_1, T_2, ...

Vratimo se ponovo na Primer 3.3. Da bismo izbegli razmatranje o približno jednakim sposobnostima četiri grupe ispitanika, ispitanici, odnosno jedinke, se

mogu podeliti u četiri grupe na slučajan način! Tako možemo doći do potpuno slučajnog plana:

DEFINICIJA 4.2. **Potpuno slučajan plan** *za poređenje k tretmana je onaj u kome se n jedinki na kojima se vrši merenje na slučajan način podeli u k grupa obima $n_1, n_2, \ldots, n_k$, $n_1 + n_2 + \cdots + n_k = n$. Posle toga se sve jedinke jedne grupe izlažu istom tretmanu. Dakle, svaki tretman se primenjuje na samo jednu od dobijenih grupa, ali na sve jedinke u toj grupi.* $\diamondsuit$

Kao što se može da zapazi, ovaj plan je korišćen u teoriji jednofaktorskog problema u prethodnoj glavi.

Naglasimo činjenicu da se posmatranja (merenja) dobijena primenom različitih nivoa posmatranih faktora na osnovu potpuno slučajnog plana smatraju nezavisnim poduzorcima, odnosno populacijama (kao što je objašnjeno u prethodnoj glavi). Ovakav plan ima svoje posebno mesto u statistici.

DEFINICIJA 4.3. **Jednosmerni plan** *za upoređivanje k populacija je razmeštaj uzorka u kome se dobijaju nezavisni slučajni uzorci – poduzorci izborom iz svake od posmatranih populacija.* $\diamondsuit$

Dakle, jednosmerni plan je samo proširenje pojma slučajnog uzorka koji smo do sada koristili ne uvodeći pojam jednosmernog plana.

Uzorak sa slučajnim blokovima je uopštenje plana *sparenih uzoraka*. Ukoliko treba ispitati k tretmana, tada treba da nam bude na raspolaganju, recimo, b blokova savaki sa po tačno k jedinki.

DEFINICIJA 4.4. **Uzorak sa slučajnim blokovima** *koji sadrži b blokova i k tretmana se sastoji od b blokova svaki sa po k eksperimentalnih jedinki. Tretmani se na slučajan način primenjuju na jedinke unutar bloka, ali tako da se svaki od tretmana primenjuje na tačno jednu jedinku.* $\diamondsuit$

Drugim rečima, svaki od tretmana se javlja tačno jedanput u svakom bloku.

Blok može da čini i samo jedan subjekt.

Primer 4.1. Treba ispitati uticaj četiri različita dozvoljena stimulativna sredstva na rezultat sportiste jednog sporta.

Četiri stimulativna sredstva koja ispitujemo će predstavljati, zapravo, četiri različita tretmana T_1, T_2, T_3 i T_4. Pretpostavimo da dejstvo ovih tretmana ispitujemo na osam sportista. Na svakog od sportista delovaćemo svakim

od četiri tretmana u slučajnom redosledu. Odnosno, svaki od sportista će predstavljati jedan blok. Razumljivo da prilikom primene raznih tretmana na jednog sportistu treba da protekne odgovarajuće, ali isto vreme za svakog od njih da bi se izbegao uticaj prethodnog tretmana. △

Kod uzorka sa slučajnim blokovima, izraz *slučajni blokovi* predstavlja činjenicu da se unutar pojedinog bloka tretmani primenjuju u slučajnom redosledu, na slučajan način.

Blokovi mogu da budu vreme, mesto, eksperimentalni materijal itd. Kada su, na primer u pitanju životinje, obično se životinje iz istog legla uzimaju za jedan blok.

Uzorak slučajnih blokova je samo jedan od mnogih blokovskih planova. Na primer *Latinski kvadrat* je takođe jedan blokovski plan. Međutim, ovde se nećemo njime dalje baviti, jer je cilj ove glave samo da predstavi postojanje i drugih planova uzorka osim potpuno slučajnog plana. Nadalje ćemo se kratko baviti primenom analize rasipanja na uzorak sa slučajnim blokovima.

4.1 Analiza rasipanja kod uzoraka sa slučajnim blokovima

Da bismo izgradili matematički model za analizu rasipanja kod uzorka sa slučajnim blokovima, primetimo da, kada imamo tačno b blokova i k tretmana, uzorak koji tom prilikom nastaje je obima $n = bk$. Takođe primetimo da kod ovog plana uzorka imamo posla sa dve nezavisne kvalitativne promenljive – blokovi i tretmani. Ako sa Y_{ij} označimo ishod eksperimenta pri i–tom tretmanu, $i = 1, \ldots, k$ u j–om bloku, $j = 1, \ldots b$, sa m ukupno očekivanje, tj. za obeležje Y, ishod eksperimenta, važi $E(Y) = m$, sa τ_i (neslučajni) efekat i–tog tretmana, sa β_j (neslučajni) efekat j–tog bloka, a sa ε_{ij} slučajnu grešku pri primeni i–tog tretmana u j–tom bloku, imaćemo model

$$Y_{ij} = m + \tau_i + \beta_j + \varepsilon_{ij}\,,\ i = 1, \ldots, k\,,\ j = 1, \ldots, b\,, \tag{4.1}$$

uz uslov da su greške ε_{ij} nezavisne među sobom i sve sa istom raspodelom $\mathcal{N}(0, \sigma^2)$. Ukupni efekat tretmana posebno, posebno blokova je 0, tj. $\sum_{i=1}^{k} \tau_i = \sum_{j=0}^{b} \beta_j = 0$.

Uzorak koji se u ovom modelu razmatra se može da predstavi sledećom tabelom

Tretmani ↓ \Blokovi →	1	2	...	b	
1	Y_{11}	Y_{12}	...	Y_{1b}	$Y_{1\bullet}$
2	Y_{21}	Y_{22}	...	Y_{2b}	$Y_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
k	Y_{k1}	Y_{k2}	...	Y_{kb}	$Y_{k\bullet}$
	$Y_{\bullet 1}$	$Y_{\bullet 2}$	...	$Y_{\bullet b}$	$Y_{\bullet\bullet}$

gde je Y_{ij} vrednost obeležja Y na elementu uzorka izloženom i–tom tretmanu u j–tom bloku, $Y_{i\bullet} = \sum_{j=1}^{b} Y_{ij}$, $Y_{\bullet j} = \sum_{i=1}^{k} Y_{ij}$, $Y_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{b} Y_{ij}$.

Parametri $m, \tau_1, \ldots, \tau_k, \beta_1, \ldots, \beta_b$ i σ^2 su nepoznati. Razmatraju se i modeli kod kojih je σ^2 poznato, ali mi se nećemo zadržavati na takvim modelima. Pri tome posebno uočimo sabirke $\beta_1, \ldots, \beta_b$ u modelu. Ukoliko su oni neke konstante, tj. efekti blokova se pretpostavlja da su fiksirane nepoznate konstante, imaćemo model sa *fiksiranim efektima blokova*. Za razliku od ovakvog modela, posebno se razmatra model kod koga su *slučajni efekti blokova*, dakle, gde se $\beta_1, \ldots, \beta_b$ smatraju slučajnim promenljivim. Nadalje ćemo posmatrati samo model sa fiksiranim efektima blokova. Jasno da su posmatranja (ishodi eksperimenta) unutar jednog bloka *zavisna* među sobom. To je i ključna razlika u odnosu na model (3.1) kod koga su poduzorci bili nezavisni.

Model (4.1) se razlikuje od modela (3.1) samo po sabirku β_j. Drugi se odnosi na *potpuno slučajan plan*, koji je zapravo specijalan slučaj jednostranog plana, dok prvi sadrži i sabirak koji "meri" efekat svakog pojedinog bloka.

Neke od osnovnih osobina modela (4.1) su:

$$E(Y_{ij}) = m + \tau_i + \beta_j\,,\ \ D(Y_{ij}) = \sigma^2\,,\ \ i = 1, 2, \ldots, k, \,,\ j = 1, 2, \ldots, b\,,$$
$$E(Y_{ij_1}) - E(Y_{ij_2}) = \beta_{j_1} - \beta_{j_2}\,, \quad E(Y_{i_1 j}) - E(Y_{i_2 j}) = \tau_{i_1} - \tau_{i_2}\,,$$
$$E(Y_{i_1 j_1}) - E(Y_{i_2 j_2}) = (\tau_{i_1} - \tau_{i_2}) + (\beta_{j_1} - \beta_{j_2})\,,$$
$$i_1, i_2 \in \{1, 2, \ldots, k\}\,,\ j_1, j_2 \in \{1, 2, \ldots, b\}\,.$$

Kada je obeležje Y izloženo posmatranom tretmanu na nivou i, onda je, jasno, $E(Y_{i\bullet}) = m + \tau_i = m_i$. Uzimajući u obzir takvo označavanje, model se

može da zapiše i u obliku.

$$Y_{ij} = m_i + \beta_j + \varepsilon_{ij} \tag{4.2}$$

sa svim do sada objašnjenim značenjima.

Analiza rasipanja kod uzorka sa slučajnim blokovima se sprovodi na sličan način kao kod potpuno slučajnog plana za koji smo postupak već opisali.

Testiraju se sledeće hipoteze:

$$H_{0T}(\tau_1 = \tau_2 = \cdots = \tau_k = 0) \quad \text{odnosno} \quad H_{0T}(m_1 = m_2 = \cdots = m_k) \tag{4.3}$$

protiv alternativne

$$H_{1T}(\exists i; i \in \{1, 2, \ldots, k\},\ \tau_i \neq 0) \quad \text{odnosno}$$
$$H_{1T}(\exists i, l;\ i, l \in \{1, 2, \ldots, k\},\ m_i \neq m_l)\,,$$

ali i

$$H_{0B}(\beta_1 = \beta_2 = \cdots = \beta_b = 0) \tag{4.4}$$

protiv alternativne

$$H_{1B}(\exists j; j \in \{1, 2, \ldots, b\},\ \beta_j \neq 0)\,.$$

Hipoteza (4.3) znači da su različiti tretmani bez uticaja na ishod eksperimenta, dok hipoteza (4.4) znači da je podela na blokove bila bez uticaja na ishod eksperimenta, tj. da smo eksperiment mogli da sprovedemo bez podele na blokove koristeći potpuno slučajan plan.

Da bismo izvršili ova testiranja posmatramo statistike:

- sredina celog uzorka

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{b} Y_{ij}\,, \quad n = bk\,,$$

- sredina j–tog bloka

$$\overline{Y}_{\bullet j} = \frac{1}{k}\sum_{i=1}^{k} Y_{ij}\,,$$

- sredina i–tog tretmana

$$\overline{Y}_{i\bullet} = \frac{1}{b}\sum_{j=1}^{b} Y_{ij}\,,$$

- ukupna suma kvadrata

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \overline{Y})^2\,,$$

- suma kvadrata unutar blokova

$$Q_B = k\sum_{j=1}^{b}(\overline{Y}_{\bullet j} - \overline{Y})^2\,,$$

- suma kvadrata tretmana

$$Q_T = b\sum_{i=1}^{k}(\overline{Y}_{i\bullet} - \overline{Y})^2\,,$$

- suma kvadrata grešaka

$$Q_E = \sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \overline{Y}_{\bullet j} - \overline{Y}_{i\bullet} + \overline{Y})^2\,.$$

Tada se ukupna suma kvadrata Q može da zapiše kao zbir

$$Q = Q_B + Q_T + Q_E\,,$$

što se lako pokazuje. Takođe se može pokazati da su sva tri sabirka nezavisna među sobom.

Ako je H_{0T} tačna, onda statistika

$$F_{(k-1),(b-1)(k-1)} = \frac{\frac{Q_T}{k-1}}{\frac{Q_E}{(b-1)(k-1)}}$$

ima Fišerovu raspodelu sa $(k-1)$ i $(b-1)(k-1)$ stepeni slobode.

Ako je H_{0B} tačna, statistika

$$F_{(b-1),(b-1)(k-1)} = \frac{\frac{Q_B}{b-1}}{\frac{Q_E}{(b-1)(k-1)}}$$

ima Fišerovu raspodelu sa $(b-1)$ i $(b-1)(k-1)$ stepeni slobode.

Ako se, kao i do sada, koriste oznake

$$f_{(k-1),(b-1)(k-1)} \quad \text{i} \quad f_{(b-1),(b-1)(k-1)}$$

za realizovane vrednosti poslednjih statistika redom, najbolje kritične oblasti veličine α su:

H_0	H_1	C
$\tau_1 = \tau_2 = \ldots$ $\ldots = \tau_k = 0$	$\exists i, i \in \{1, 2, \ldots, k\}$, $\tau_i \neq 0$	$f_{(k-1),(b-1)(k-1)} \geq$ $F_{(k-1),(b-1)(k-1);1-\alpha}$
$\beta_1 = \beta_2 = \ldots$ $\ldots = \beta_b = 0$	$\exists j, j \in \{1, \ldots, b\}$, $\beta_j \neq 0$	$f_{(b-1),(b-1)(k-1)} \geq$ $F_{(b-1),(b-1)(k-1);1-\alpha}$

gde su $F_{(k-1),(b-1)(k-1);1-\alpha}$ i $F_{(b-1),(b-1)(k-1);1-\alpha}$ kvantili reda $1-\alpha$ Fišerove raspodele sa naznačenim stepenima slobode.

Primer 4.2. Na 15 dobrovoljaca koji su bolovali od iste bolesti trebalo je izvršiti eksperiment primene tri vrste terapija da bi se utvrdilo koja od terapija je najuspešnija u izlečenju, odnosno poboljšanju stanja pacijenata. Testiranje vršiti sa $\alpha = 0,01$ nivoom značajnosti.

Ovih 15 dobrovoljaca je podeljeno u 5 blokova, svaki po tri pacijenta. Zahvaljujući tome što svi dobrovoljci nisu bili u istom stadijumu bolesti, naprotiv, blokovi su bili formirani tako da u bloku 1 budu pacijenti u najnižem stadijumu bolesti, u bloku 2 u nešto višem, odnosno naprednijem stadijumu i tako dalje do bloka broj 5 u kome su se našla tri pacijenta sa najdalje poodmaklom bolešću.

Na skali kojom se identifikuje bolest, izvršena su merenja za svakog ispitanika ponaosob pre i posle primene terapije/tretmana. Razlike ta dva merenja su prikazane tabelom

Tretmani ↓ \Blokovi →	1	2	3	4	5	
1	8	11	9	16	24	68
2	2	1	12	11	19	45
3	-2	0	6	2	11	17
	8	12	27	29	54	130

Računanjem na osnovu tabele se dobija

$$Q = 767,3\,, \quad Q_T = 260,9\,, \quad Q_B = 438,0\,,$$

odakle se dobija da je $Q_E = 68,4$.

Testiraćemo najpre hipotezu $H_{0T}(\tau_1 = \tau_2 = \tau_3 = 0)$, protiv odgovarajuće alternativne. Realizovana vrednost test statistike je $f_{2,8} = 15,26$, a granica kritične oblasti je $F_{2,\,8;\,0,99} = 8,65$. Prema tome, nultu hipotezu treba odbaciti, tj. postoji tretman ili više njih koji utiču na poboljšanje stanja pacijenta. Da bi se utvrdilo koji je tretman taj koji ima uticaj na poboljšanje stanja pacijenata, treba nastaviti postupak testiranja testiranjem po dva tretmana, dakle, poređenjem svaka dva tretmana.

Što se tiče hipoteze o uticaju blokova na ishod eksperimenta, testiraćemo nultu hipotezu $H_{0B}(\beta_1 = \cdots = \beta_5 = 0)$, protiv odgovarajuće alternativne. Realizovana vrednost test statistike je $f_{4,8} = 12,807$, a granica kritične oblasti je $F_{4,\,8;\,0,99} = 7,01$ pa i tu nultu hipotezu treba odbaciti. Međutim, to je i bilo za očekivati s obzirom na to da su blokovi pravljeni tako da budu različiti među sobom (koliko je to bilo moguće).

Kako bismo rešili ovaj primer u programskom jeziku `R`, implemetiraćemo funkciju `blok.anova` koja nam daje realizovane vrednosti statistika $f_{2,8}$ i $f_{4,8}$ kao i odgovarajuće p-vrednosti. Implementacija funkcije `blok.anova` glasi

```
blok.anova<-function(imeFajla, nivoiFaktora){
   tabela<-read.csv(imeFajla, header = TRUE)
   r = c(t(as.matrix(tabela$merenje)))
   k = length(nivoiFaktora) #broj nivoa
   n = length(r) / k #broj blokova
   tretmani = gl(k, 1, n*k, factor(nivoiFaktora))
   blokovi = gl(n, k, k*n)
   av = aov(r ~ tretmani + blokovi)
   summary(av)}
```

Funkcija `blok.anova` se bazira na funkciji `aov` o kojoj je ranije bilo reči. Bitno je samo kako se formira .csv fajl koji se učitava u objekat `tabela`, a da bi funkcija ispravno radila, vrednosti iz gornje tabele se u .csv fajl upisuju po kolonama. Takođe kolona koja sadrži vrednosti merenja mora biti naslovljena "merenje". Pozivom funkcije `blok.anova` dobićemo sledeći rezultat

```
> blok.anova("data.csv", c("Tretman_1","Tretman_2","Tretman_3"))
            Df  Sum Sq  Mean Sq  F value    Pr(>F)
 tretmani    2   260.9   130.47    15.26   0.00186**
 blokovi     4   438.0   109.50    12.81   0.00148**
 Residuals   8    68.4     8.55
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

Kolona `F value` nam vraća vrednosti za tražene statistike, dok iz kolone `Pr(>F)` čitamo njihove p-vrednosti. Kao što smo to i ranije zaključili, nulta hipoteza se odbacuje u oba slučaja.

Napomena: Primetimo da smo ovde iskoristili funkciju `gl` koja daje niz čiji se članovi raspoređuju prema obrascu koji smo definisali. Elementi koji formiraju niz su zadati kroz promenljivu `factor(nivoiFaktora)`, a ako je ta promenljiva izostavljena, onda su ti elementi prirodni brojevi (počevši od jedinice pa redom). $\triangle$

Naglasimo sledeće. Ukoliko rezultat testiranja kaže da H_{0B} treba odbaciti, onda je bila opravdana takva podela na blokove sa kojom je rađeno. Ukoliko H_{0B} treba prihvatiti, odnosno ne treba je odbaciti, sledeće eksperimente koji upoređuju iste te tretmane verovatno treba sprovesti ili (1) koristeći drugačiju podelu na blokove, ili (2) koristeći jednosmerni plan.

U stvari, da li treba koristiti uzorak sa slučajnim blokovima ili potpuno slučajan plan, reći će nam vrednost Q_B. Ako je ta vrednost "velika", prednost treba dati uzorku sa slučajnim blokovima. Ako je ta vrednost "mala", treba koristiti jednosmerni plan. Upravo takav slučaj smo imali u poslednjem primeru.

4.2 Definisanje blokova

Plan uzorka sa slučajnim blokovima se razlikuje od potpuno slučajnog plana po tome što su eksperimentalne jedinke grupisane u blokove na osnovu

poznatih razlika, ili makar na osnovu razlika za koje se sumnja da postoje. Kada su te razlike kvantitativne prirode, jednostavno je izvršiti grupisanje, recimo, ako grupišemo jedinke prema broju godina. S druge strane, ako bi se radilo o grupisanju na osnovu nečije sklonosti ka muzici, idealan raspored po blokovima nije moguće izvršiti. Kako bilo, cilj je postići što veću homogenost unutar bloka, a između blokova što veću razliku. Ono što je dobro kod ovog pristupa je što možemo sami odrediti broj blokova. Najvažnije je da ovakav pristup daje preciznije rezultate od potpuno slučajnog plana.

Plan uzorka sa slučajnim blokovima se koristi kako bi se povećala preciznost eksperimenta putem smanjenja varijanse slučajne promenljive ε_{ij}. Manja varijansa greške postiže se grupisanjem jedinki u homogene poduzorke (blokove), i sprovođenjem eksperimenta nezavisno od toga o kom se bloku radi. Blokovi se ne smeju definisati nakon obavljenog eksperimenta.

Ovaj metod se pokazao efikasnim ako postoji jasno definisana osobina po kojoj se razlikuju jedinke, koju možemo uočiti pre nego što otpočnemo eksperiment.

Postoji gotovo beskonačno mnogo načina kako jedinke možemo razvrstati po blokovima. Najbolje bi bilo kad bismo mogli da grupišemo jedinke tako da je ishod eksperimenta homogen unutar bloka. Kako ovo po pravilu nije moguće, grupisanje se vrši na osnovu sličnosti za koje smatramo da bi mogle da utiču na rezultat testiranja.

Nekada je pravljenje blokova sasvim očigledno. Recimo da u fabriku sira svaki dan stigne jedna cisterna mleka od koje se proizvode različite vrste sireva. Od mleka iz cisterne svaki dan dobijamo malo drugačiji sir. Ovde je prirodno da se blokovi definišu prema cisternama sa mlekom.

Grupisanje se može izvršiti i prema lokaciji. Recimo da police na kojima stoje proizvodi definišu blokove, ili da gradovi u kojima se nalaze jedinke određuju blokove. Za očekivati je da što su lokacijski bliskije jedinke, to ima više sličnosti između njih.

Jedinke se mogu grupisati i prema vremenu kada je izvršeno testiranje. Ako možemo da uradimo samo pet merenja dnevno, onda bismo blokove pravili prema danima. Slično kao i kod grupisanja prema lokaciji i kod grupisanja prema vremenu rezultati koji su vremenski bliski međusobno su sličniji.

Kada smo definisali uzorak sa slučajnim blokovima, naveli smo da je k tret-

mana primenjeno na k jedinki, pri čemu svaki od b blokova sadrži k jedniki. Međutim u praksi nailazimo na određene poteškoće. Na primer, možemo precizno da razvrstamo jedinke po blokovima, ali nema dovoljno jedinki u svakom bloku za svaki tretman. Zapravo, postoji blok sa manje od k jedinki. To nas dovodi do nekompletnog dizajna blokova, što zahteva posebnu analizu.

S druge strane, može se desiti da neki od blokova ima više od k jedinki. Prevazilaženje ovog problema zavisi od više faktora. Ako jedinke nisu previše "skupe", možemo zanemariti višak jedinki koje se javljaju u blokovima. Na ovaj način najlakše smo sveli uzorak na onaj iz definicije 4.4. Ako se teško dolazi do jedinke nad kojom se vrši ispitivanje, ovakav pristup nije prihvatljiv. U najboljem slučaju imaćemo u svakom bloku l puta više jedinki nego tretmana (gde je l neki prirodan broj). Tada ćemo slučajnim izborom tretman primeniti na l jedinki jednog bloka, drugi tretman na drugih l jedinki i tako dalje.

Postoji mogućnost da broj jedinki unutar bloka nije jednak broju tretmana pomnoženom brojem l. Tada se opet radi o nekompletnom dizajnu blokova, a ovaj problem možemo prevazići na sledeći način. Na primer, posmatramo tri tretmana T_1, T_2 i T_3 na uzorku koji smo rasporedili u tri bloka sa po pet jedinki. U bloku 1 pimenićemo tretmane (T_1,T_2,T_3,T_1,T_2), u bloku 2 (T_1,T_2,T_3,T_1,T_3) i u bloku 3 (T_1,T_2,T_3,T_2,T_3). Dakle, u svakom bloku su primenjeni svi tretmani plus dva dodatna.

Poslednja mogućnost je da imamo blokove sa različitim brojem jedinki, odnosno neki blokovi imaju više jedinki od drugih. U ovom slučaju najbolje rezultate dobićemo optimizacijom blokova na osnovu nekog kriterijuma. Algoritmi koji rade ovu optimizaciju su komplikovani pa ih sada nećemo navoditi, a mogu se naći kao deo programskih paketa za statistiku.

Glava 5

Statistička analiza slučajnih procesa

U statistici su izgrađeni mnogi modeli kojima se modeliraju podaci koji imaju karakter tzv. slučajnih procesa. Zbog toga ćemo se ovde upoznati sa osnovama statističke analize slučajnih procesa i samo sa njima, jer dublje ulaženje u modele slučajnih procesa prevazilazi obim naše analize.

Da bismo se upoznali sa osnovama statističke analize slučajnih procesa, definišimo najpre sam slučajni proces i njemu srodan vremenski niz.

Definicija 5.1. *Neka je* $(\Omega, \mathcal{F}, P)$ *prostor verovatnoća i* $T \subseteq R$. *Neka je funkcija* $X(\omega, t) : \Omega \times T \to R$, *takva da je za svako fiksirano* t *ona jedna slučajna promenljiva i to jedno ili višedimenziona. Skup* $X = \{X(\omega, t), t \in T\}$ *(kraće* $X_t(\omega)$*) je (realni)* **slučajni proces** *sa neprekidnim vremenom* t. *Skup* T *se zove* **indeksni skup** *ili* **parametarski skup** *slučajnog procesa. Ukoliko* T *nije interval već diskretan skup tačaka,* $T \subseteq Z$, *skup* X *se zove* **slučajni niz** *ili* **vremenski niz** *ili* **vremenska serija** *ili* **slučajni proces sa diskretnim vremenom**. $\diamondsuit$

Primetimo da smo t nazvali *vremenom*. Skup T ne mora da bude skup vremena, ali se taj termin odomaćio pre svega zbog toga što skup T najčešće baš i jeste skup vremena.

Definicija 5.2. *Pri fiksiranom* $\omega \in \Omega$, $X(\omega, t)$ *postaje funkcija samo od vremena (parametra)* t *i kao takva zove se* **trajektorija** *ili* **realizacija slučajnog**

procesa X *koja odgovara elementarnom ishodu* ω. $\diamondsuit$

DEFINICIJA 5.3. *Pri fiksiranom* $t \in T$, $X(\omega, t)$ *je slučajna promenljiva koja se zove* **zasek slučajnog procesa**. $\diamondsuit$

Dakle, za fiksirano ω, u pitanju je realna funkcija argumenta t. Kodomen slučajnog procesa može da bude i skup kompleksnih brojeva (ili njegov pravi kompleksni podskup), ali se mi nećemo baviti takvim procesima.

Naglasimo još jednom da ako je skup T neprebrojiv (najčešće je u pitanju neki interval iz skupa realnih brojeva), kaže se da je dati proces sa neprekidnim vremenom. Ukoliko je $T \subseteq Z$, gde je sa Z označen skup celih brojeva, kaže se da je u pitanju proces sa diskretnim vremenom. Za slučajni proces sa neprekidnim vremenom se, jednostavno, koristi termin **slučajni proces**, dok se za onaj sa diskretnim vremenom koristi termin **slučajni niz**, ili **vremenski niz**, ili **vremenska serija**.

U teoriji postoje mnoga uopštenja definicije slučajnog procesa, ali mi ćemo se zadržati na gornjoj definiciji.

Jedan od osnovnih zadataka pri obradi rezultata merenja koja imaju karakter sukcesivne zavisnosti, statističkih podataka sa takvim svojstvom, pojava koje imaju karakter slučajnih procesa, je određivanje statističkih parametara ili funkcija koje karakterišu statistička svojstva tih procesa. U takve zadatke spadaju: ocenjivanje srednje vrednosti, ocenjivanje disperzije, ocenjivanje autokorelacione funkcije, spektralne funkcije, spektralne gustine i dr. Pri tome je i u ovom slučaju jedno od najvažnijih pitanja kvalitet predložene ocene u odnosu na postavljeni kriterijum valjanosti. Statistika slučajnih promenljivih, u najvećem delu razrađena za prost slučajni uzorak, ne može se direktno "preneti" na slučajne procese. Razlog za to je, pre svega taj, što su ordinate realizacija slučajnog procesa $\{X(t), t \in T\}$, po pravilu, realizacije međusobno zavisnih slučajnih veličina. U ovom kratkom osvrtu na statistiku slučajnih procesa bavićemo se uglavnom slučajnim procesima *stacionarnim u širokom smislu* ili *slabo stacionarnim*. Da bismo uveli definiciju slabo stacionarnog procesa, uvešćemo najpre definiciju *autokovarijansne funkcije*.

DEFINICIJA 5.4. **Autokovarijansna funkcija** *slučajnog procesa* $\{X(t), t \in$

$T\}$ *je funkcija*

$$K(t,s) = Cov(X(t), X(s)) = E\left[(X(t) - EX(t))\,(X(s) - EX(s))\right] .\diamondsuit$$

Kao što vidimo iz definicije, autokovarijansna funkcija je u opštem slučaju funkcija od dve promenljive.

DEFINICIJA 5.5. *Slučajni proces* $\{X(t), t \in T\}$ *je* **slabo stacionaran** *ako ima momente drugog reda i pri tome su zadovoljeni uslovi*

1. $EX(t) = m$, *tj. matematičko očekivanje, odnosno srednja vrednost, je konstantno za svaki* $t \in T$ *i*

2. *autokovarijansna funkcija je funkcija jedne promenljive,* $K(t,s) = B(t-s)$, $t, s \in T$, *odnosno, zavisi samo od razlike svojih argumenata.* $\diamondsuit$

Pored ove stacionarnosti definiše se i stroga stacionarnost.

DEFINICIJA 5.6. *Slučajni proces* $\{X(t),\ t \in T\}$ *je* **strogo stacionaran** *ako važi*

$$F_{X_{t_1},\ldots,X_{t_n}}(x_1,\ldots,x_n) = F_{X_{t_1+k},\ldots,X_{t_n+k}}(x_1,\ldots,x_n),\ (x_1,\ldots,x_n) \in R^n$$

za proizvoljnu n–torku $(t_1, t_2, \ldots, t_n) \in T^n$ *i proizvoljno k iz skupa celih brojeva.* $\diamondsuit$

Naglasimo da je u literaturi opšteprihvaćeno da kada se kaže "stacionaran", misli se na slabu stacionarnost.

Posebnu pogodnost u postupku ocenjivanja pružaju *ergodični slučajni procesi* koji imaju svojstvo da im se parametri mogu ocenjivati samo na osnovu jedne realizacije.

DEFINICIJA 5.7. *Neka je* $\{X(t),\ t \in T\}$ *slabo stacionaran slučajni proces takav da je* $EX(t) = m < \infty$ *i* $E\left[X(t) - m\right]^2 = \sigma^2 < \infty$ *za svako* $t \in T$. *Neka je* $(X_{t_1}, \ldots, X_{t_n})$, $t_1, t_2, \ldots, t_n \in T$, *deo realizacije tog procesa i*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_{t_i}$$

sredina tog uzorka. Ako $\overline{X}$ *konvergira u verovatnoći ka* m *kada* $n \to \infty$*, kaže se da je proces* **ergodičan po očekivanju**.$\diamondsuit$

Definicija 5.8. *Neka je* $\{X(t),\, t \in T\}$ *slabo stacionaran slučajni proces takav da je* $EX(t) = m < \infty$ *i* $E\left[X(t) - m\right]^2 = \sigma^2 < \infty$ *za svako* $t \in T$*. Neka je* $(X_{t_1}, \ldots, X_{t_n})$*,* $t_1, t_2, \ldots, t_n \in T$*, deo realizacije tog procesa i*

$$\widehat{B}(h) = \frac{1}{n-h} \sum_{i=h+1}^{n} \left(X_{t_i} - m\right)\left(X_{t_{i-h}} - m\right) .$$

Kažemo da je proces **ergodičan po drugom momentu** *ako* $\widehat{B}(h)$ *konvergira u verovatnoći ka* $B(h)$ *kad* $n \to \infty$*,* $h < n$.$\diamondsuit$

Ergodičnost se fokusira na asimptotsku nezavisnost elemenata niza, odnosno proizvoljnog procesa, dok se stacionarnost bavi vremenskom invarijantnošću elemenata slučajnog procesa.

Ovi i drugi specijalni slučajevi biće naglašavani nadalje.

Uočimo da su parametri slučajnih procesa u opštem slučaju funkcije od vremena $t \in T$. Nadalje ćemo se kratko baviti ocenama ovih funkcija.

5.1 Slučajni procesi

Zadržimo se prvo na procesima neprekidnog vremena.

5.1.1 Ocene srednje vrednosti

Bavićemo se najpre opštim slučajem.

Neka je $\{X(t), t \in R\}$ realan slučajni proces čija je srednja vrednost

$$EX(t) = m(t)$$

nepoznata (funkcija jednog realnog argumenta). Neka je poznato n **nezavisnih** realizacija ovog slučajnog procesa:

$$x_1(t), x_2(t), \ldots, x_n(t)$$

u nekom fiksiranom vremenskom intervalu pravom podskupu od R. Bez smanjenja opštosti možemo pretpostaviti da je u pitanju interval $[0, T_0]$, tj. $t \in [0, T_0]$. Izaberimo proizvoljan momenat $t_0 \in [0, T_0]$ i posmatrajmo zasek procesa $X(\omega, t_0)$, $\omega \in \Omega$ u oznaci $X(t_0)$. Na realizacije ordinata u tom momentu možemo gledati kao na n realizovanih vrednosti slučajne veličine $X(t_0)$. Pod svim ovim uslovima se ocena vrednosti $m(t_0)$ može dobiti kao ocena nepoznatog matematičkog očekivanja slučajne promenljive $X(t_0)$ na osnovu prostog slučajnog uzorka. Dakle, uzećemo

$$\overline{x}(t_0) = \frac{1}{n} \sum_{j=1}^{n} x_j(t_0).$$

U tom slučaju za ocenu srednje vrednosti $m(t_0)$ imamo statistiku

$$\widetilde{m}(t_0) = \frac{1}{n} \sum_{j=1}^{n} X_j(t_0).$$

Jasno da je

$$E\widetilde{m}(t_0) = m(t_0),$$

odnosno da je statistika $\widetilde{m}(t_0)$ nepristrasna ocena srednje vrednosti $m(t_0)$. Srednjekvadratna greška ove ocene je

$$E[\widetilde{m}(t_0) - m(t_0)]^2 = D[\widetilde{m}(t_0)] = E\left\{ \frac{1}{n} \sum_{j=1}^{n} [X_j(t_0) - m(t_0)] \right\}^2 = \frac{1}{n} D(t_0),$$

jer smo pretpostavili da su realizacije nezavisne među sobom. Sa $D(t_0)$ je označena disperzija slučajne promenljive $X(t_0)$, gde je $t_0 \in [0, T_0]$ fiksirano.

Ukoliko je disperzija $D(t_0)$ konačna, statistika $\widetilde{m}(t_0)$ je postojana ocena srednje vrednosti.

Navedeni postupak bi trebalo ponoviti za svaku tačku $t_0 \in [0, T_0]$. To je, međutim, po pravilu nemoguće, bilo zato što nam je u praksi najčešće dostupan samo mali broj merenja, bilo zato što u određenom vremenskom intervalu možemo da dođemo do samo jedne realizacije posmatranog slučajnog procesa. U takvim slučajevima se nameće pitanje da li možemo vršiti usrednjenje samo jedne realizacije po vremenu, umesto usrednjenja više realizacija. Ta pitanja

su u tesnoj vezi sa pitanjima ergodičnosti. Naime, ergodični slučajni procesi imaju upravo to svojstvo da se može vršiti ocenjivanje njihovih parametara na osnovu samo jedne realizacije. Nadalje ćemo posmatrati samo ergodične slučajne procese.

Razmotrimo sada slabo stacionaran slučajni proces $\{X(t), t \in R\}$ sa nepoznatom srednjom vrednošću $EX(t) = m$ i nepoznatom autokorelacionom funkcijom $B(t)$. Neka su nam poznate vrednosti jedne realizacije procesa na intervalu $[0, T_0]$. Razbijmo taj interval na n jednakih delova dužine $\Delta = \frac{T_0}{n}$. Ako za ocenu srednje vrednosti uzmemo

$$\widehat{\widehat{m}} = \frac{1}{n+1} \sum_{j=0}^{n} x(j \cdot \Delta)\,,$$

dobićemo integralnu sumu funkcije $x(t)$ na intervalu $[0, T_0]$,

$$\widehat{\widehat{m}} = \frac{1}{(n+1) \cdot \Delta} \sum_{j=0}^{n} x(j \cdot \Delta) \cdot \Delta = \frac{1}{T_0 + \Delta} \sum_{j=0}^{n} x(j \cdot \Delta) \cdot \Delta\,.$$

Dakle, za $n \to \infty$, odnosno, $\Delta \to 0$ dobijamo ocenu

$$\widehat{m} = \frac{1}{T_0} \int_0^{T_0} X(t)dt. \tag{5.1}$$

U opštem slučaju, za interval $[a, b]$, $a < b$,

$$\widehat{m} = \frac{1}{b-a} \int_a^b X(t)dt\,.$$

Ocena $\widehat{m}$ je nepristrasna, što se lako može pokazati, dok je srednjekvadratna greška funkcija kovarijansne funkcije procesa.

Nastavimo, bez smanjenja opštosti, da govorimo o ocenama isključivo na intervalu $[0, T_0]$.

Osim ocene (5.1), u praksi se često koriste i druge, složenije ocene srednje vrednosti, gde se pomenuta ocena javlja samo kao specijalan slučaj. Najčešće se koristi ocena

$$\widehat{m} = \int_0^{T_0} a(t)X(t)dt\,,$$

gde je $a(t)$ težinska funkcija (jezgro) koja zadovoljava uslove:

$$\int_0^{T_0} a(t)dt = 1 \tag{5.2}$$

i

$$a(t) = 0 \quad \text{za} \quad t \notin [0, T_0]\,. \tag{5.3}$$

Jezgro može biti i prekidna funkcija, a za $a(t) = \frac{1}{T_0}$ za $t \in [0, T_0]$ dobija se ocena (5.1).

Ovako definisana ocena srednje vrednosti je nepristrasna:

$$E\widehat{m} = E\int_0^{T_0} a(t)X(t)dt = \int_0^{T_0} a(t)EX(t)dt = m\int_0^{T_0} a(t)dt = m\,,$$

a njena srednjekvadratna greška zavisi od jezgra.

5.1.2 Ocena disperzije

O oceni za disperziju govorićemo samo kod slabo stacionarnih procesa. Takođe ćemo, bez smanjenja opštosti, pretpostaviti da je srednja vrednost procesa jednaka 0. To otuda što se slabo stacionaran proces sa očekivanjem $EX(t) = m$ može translacijom $Y(t) = X(t) - m$ prevesti u takođe slabo stacionaran proces, ali sa očekivanjem 0.

Ocena za nepoznatu disperziju, $K(t,t) = B(0) = \sigma^2$, $t \in R$, o kojoj će ovde biti reči, je statistika

$$\widehat{\sigma^2} = \int_0^{T_0} a(t)X^2(t)dt\,,$$

gde je, kao i u prethodnom slučaju, $a(t)$, $t \in R$, jezgro. Da bi ova ocena bila nepristrasna, potrebno je i dovoljno da jezgro zadovoljava iste uslove kao i ranije, tj. uslove (5.2) i (5.3). Zaista,

$$E\widehat{\sigma^2} = E\left[\int_0^{T_0} a(t)X^2(t)dt\right] = \int_0^{T_0} a(t)EX^2(t)dt = \sigma^2\int_0^{T_0} a(t)dt = \sigma^2\,.$$

Srednjekvadratna greška ove ocene je

$$E(\widehat{\sigma^2} - \sigma^2)^2 = D\widehat{\sigma^2} = E(\widehat{\sigma^2})^2 - \sigma^4 = \int_0^{T_0} \int_0^{T_0} a(t)a(s)K_{X^2}(t,s)dtds\,,$$

gde je $K_{X^2}(t,s) = Cov\left(X^2(t), X^2(s)\right)$, $t, s \in [0,\, T_0]$.

Dakle, da bi se odredila greška ocene, potrebno je poznavati kovarijansnu funkciju procesa $\{X^2(t), t \in R\}$, odnosno momente četvrtog reda posmatranog slučajnog procesa. To znači da bi za precizan odgovor na pitanje o srednjekvadratnoj grešci ocene za disperziju procesa bilo neophodno poznavanje četvorodimenzione raspodele procesa $\{X(t)\}$.

5.1.3 Ocene autokovarijansne funkcije

Autokovarijansna funkcija nam daje sliku o zavisnosti "u nizanju vrednosti" slučajnog procesa. To je jedan od važnih razloga zbog kojeg se iskazuje interesovanje za kovarijansnu funkciju. Osim toga, iz kovarijansne funkcije se na posredan način određuje disperzija procesa.

Razmatraćemo proizvoljan realan slučajni proces $\{X(t), t \in R\}$ sa autokovarijansnom funkcijom

$$K(t,s) = E\left[(X(t) - m(t))(X(s) - m(s))\right]\,.$$

Pretpostavimo da nam je poznato n nezavisnih realizacija ovog procesa na intervalu $[0, T_0]$. Za ocenu autokovarijansne funkcije, za $t, s \in [0, T_0]$, može se uzeti statistika

$$\widehat{K}(t,s) = \frac{1}{n}\sum_{j=1}^{n}[X_j(t) - m(t)][X_j(s) - m(s)]\,,$$

ukoliko je poznata srednja vrednost procesa i proces ergodičan po drugom momentu. Ukoliko pak srednja vrednost procesa nije poznata, koristi se statistika

$$\widetilde{K}(t,s) = \frac{1}{n-1}\sum_{j=1}^{n}[X_j(t) - \widetilde{m}(t)][X_j(s) - \widetilde{m}(s)]\,,$$

gde je umesto nepoznate srednje vrednosti korišćena njena ocena

$$\widetilde{m}(t) = \frac{1}{n}\sum_{j=1}^{n} X_j(t)\,.$$

Lako je pokazati da su i $\widehat{K}$ i $\widetilde{K}$ nepristrasne ocene autokovarijansne funkcije posmatranog procesa.

Kod slabo stacionarnog procesa je $m(t) = m(s) = m$, pa je autokovarijansna funkcija oblika

$$K(t+\tau, t) = B(\tau) = E[(X(t+\tau) - m)(X(t) - m)]\,,$$

te se pod uslovom da je proces ergodičan, a po uzoru na (5.1), prelaskom na integralnu sumu na intervalu $[0, T_0]$, dobija ocena

$$\widehat{B}(\tau) = \frac{1}{T_0 - \tau}\int_0^{T_0-\tau} (X(t+\tau) - m)(X(t) - m)dt$$

na osnovu dela jedne realizacije.

Posmatrajmo slučajni proces $\{Y(t), t \in R\}$ nastao translacijom slabo stacionarnog slučajnog procesa $\{X(t), t \in R\}$ za srednju vrednost m, $Y(t) = X(t) - m$. Ovaj se proces može koristiti za definisanje statistike za ocenu autokovarijansne funkcije procesa $\{X(t), t \in R\}$:

$$\widehat{B}(\tau) = \int_0^{T_0-\tau} a(t,\tau)Y(t+\tau)Y(t)dt\,,$$

gde je jezgro takvo da je

$$\int_0^{T_0-\tau} a(t,\tau)dt = 1 \tag{5.4}$$

i

$$a(t,\tau) = 0 \quad \text{za} \quad t \notin [0, T_0]\,, \tag{5.5}$$

što obezbeđuje nepristrasnost ocene.

Srednjekvadratna greška ovih ocena je i sama funkcija od autokovarijansne funkcije (koja je nepoznata), pa se u postupku ocenjivanja ne traže jezgra

koja bi dala optimalne ocene, već se za unapred izabrano jezgro određuje srednjekvadratna greška. Tako je često u upotrebi jezgro

$$a(t,\tau) = \begin{cases} \frac{1}{T_0-\tau}, & 0 \le t \le T_0 \\ 0\,, & \text{van} \end{cases} .$$

U tom slučaju bi ocena autokovarijacione funkcije bila:

$$\widehat{B}(\tau) = \frac{1}{T_0-\tau}\int_0^{T_0-\tau} Y(t+\tau)Y(t)dt.$$

Ocenjivanje koje je upravo vršeno, vršeno je pod pretpostavkom da je srednja vrednost procesa m poznata. Međutim, ako i nju treba oceniti na osnovu uzorka (dela realizacije ergodičnog slabo stacionarnog procesa), koristi se statistika

$$\widetilde{B}(\tau) = \frac{1}{T_0-\tau}\int_0^{T_0-\tau} [X(t+\tau)-\widetilde{m}][X(t)-\widetilde{m}]dt,$$

gde je $\widetilde{m}$ neka nepristrasna ocena srednje vrednosti procesa. Međutim, ovako definisana statistika nije nepristrasna ocena autokovarijansne funkcije, što se može lako pokazati.

5.2 Vremenski nizovi

Mnoge pojave, pre svega u prirodi, iako su po svojoj suštini slučajni procesi (na pr. meteo i klimatski uslovi, proticaj vode u reci i sl.), proučavamo svodeći ih na slučajne nizove. Mnoge društvene pojave, kao što su razni ekonomski indeksi, bruto domaći proizvod, cene akcija na berzi i sl. su po svojoj suštini baš slučajni nizovi.

Ovde ćemo razamatrati samo realne vremenske nizove, tj. one čiji je kodomen skup realnih brojeva.

U praksi nam je poznat samo deo slučajnog niza nad nekim $T \subset Z$, odnosno samo deo realizacije $\{X_t,\, t \in T\}$. Napomenimo da se kod vremenskih nizova najčešće umesto oznake $X(t)$ koristi oznaka X_t. Zadatak statističke analize vremenskih nizova je da se na osnovu konačnog broja posmatranja, najčešće samo jedne realizacije, ocene nepoznati parametri posmatranog vremenskog

niza ili proceni njegovo ponašanje "u prošlosti" ili "budućnosti", tj. za $t \in Z \setminus T$.

Definicije stroge i slabe stacionarnosti slučajnog procesa važe i za vremenske nizove.

Bavićemo se slabo stacionarnim realnim vremenskim nizovima, dakle, onim za koje važi

$$EX_t = m(= const.)$$

i

$$Cov(X_{t+k}, X_t) = E[(X_{t+k} - m)(X_t - m)] = R_k.$$

Kao što je naznačeno, autokovarijansna funkcija zavisi samo od razlike "indekasa" slučajnih promenljivih. Otuda se autokovarijansnom funkcijom stacionarnog niza $\{X_t\}$ smatra brojni niz $\{R_k\}_{k\in Z}$. Svojstva autokovarijansne funkcije $\{R_k\}$ su analogna svojstvima funkcije B proizvoljnog realnog slučajnog procesa. Uočićemo neka od svojstava o kojima do sada nije bilo reči.

Primetimo da je

$$R_0 = Cov(X_t, X_t) = E(X_t - m)^2 \equiv \sigma^2 = const.$$

DEFINICIJA 5.9. **Autokorelaciona funkcija** *vremenskog niza* $\{X_t, t \in Z\}$ *je funkcija*

$$Corr(X_{t+k}, X_t) = \frac{Cov(X_{t+k}, X_t)}{\sqrt{Cov(X_{t+k}, X_{t+k})Cov(X_t, X_t)}} .\diamondsuit$$

Kod slabo stacionarnih nizova je ovo funkcija jedne promenljive.

Iz definicije autokorelacione funkcije ovog vremenskog niza sledi

$$Corr(X_{t+k}, X_t) = \frac{R_k}{R_0}.$$

Otuda je očigledno da je

$$|R_k| \leq R_0$$

i

$$R_{-k} = R_k.$$

Autokovarijansna funkcija slabo stacionarnog vremenskog niza je pozitivno

semidefinitna. Zaista, neka je $\{c_k\}$ proizvoljan niz realnih brojeva. Tada je

$$\sum_{k,j=1}^{n} R_{k-j} c_k c_j = \sum_{k,j=1}^{n} c_k c_j Cov(X_k, X_j) = D(\sum_{k=1}^{n} c_k X_k) \geq 0.$$

Od posebnog su interesa stacionarni vremenski nizovi za koje red $\sum_{k=-\infty}^{+\infty} R_k$ apsolutno konvergira, tj. konvergira red

$$\sum_{k=-\infty}^{+\infty} |R_k| = R_0 + 2\sum_{k=1}^{\infty} |R_k|.$$

Uočimo Furijeovu transformaciju autokovarijansne funkcije $\{R_k\}$:

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} R_k \cos k\lambda = \frac{1}{2\pi} R_0 + \frac{1}{\pi} \sum_{k=1}^{+\infty} R_k \cos k\lambda \quad , \quad \text{za } \lambda \in [-\pi, \pi].$$

Ako red kovarijansne funkcije $\{R_k\}$ apsolutno konvergira, onda njena Furijeova transformacija uniformno konvergira ka neprekidnoj funkciji argumenta λ. Suma reda, $f(\lambda)$, zove se *spektralna gustina* niza $\{X_t\}$. Postoji uzajamna jednoznačnost između spektralne gustine i kovarijansne funkcije slučajnog niza koja omogućuje da se koeficijenti Furijeovog reda, R_k, određuju na osnovu njegove spektralne gustine formulom

$$R_k = \int_{-\pi}^{\pi} f(\lambda) \cos(k\lambda) d\lambda \,.$$

Ova uzajamna jednoznačnost u predstavljanju kovarijansne funkcije i spektralne gustine jedne preko druge, omogućava da se stacionarni vremenski niz može da opiše ravnopravno preko bilo koje od njih, tj. i u vremenskom i u frekventnom domenu.

5.2.1 Ocena srednje vrednosti

Zadržimo se sada na ocenjivanju srednje vrednosti i kovarijansne funkcije realnog slabo stacionarnog i ergodičnog vremenskog niza $\{X_t\}$.

Neka nam je dato n posmatranja jedne realizacije vremenskog niza $\{X_t\}$:

$$X_1, X_2, \ldots, X_n\,.$$

Tada će nepristrasna ocena srednje vrednosti m biti

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,,$$

a greška ove ocene u smislu srednjekvadratnog odstupanja

$$\begin{aligned} E(\overline{X}-m)^2 &= \frac{1}{n^2}E\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-m)(X_j-m) = \\ &= \frac{1}{n}\left[R_0 + 2\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)R_k\right] \to 0 \end{aligned} \tag{5.6}$$

kada $n \to +\infty$.

Do ovog zaključka se dolazi na osnovu apsolutne konvergencije reda autokovarijansne funkcije. Otuda sledi da je $\overline{X}$ i postojana ocena srednje vrednosti vremenskog niza.

Na osnovu definicije spektralne gustine i zaključka (5.6) sledi

$$D(\overline{X}) = E(\overline{X}-m)^2 \sim \frac{2\pi}{n}f(0)\,.$$

U programskom jeziku `R` vrednost $\overline{X}$ dobijamo pozivom funkcije `mean(X)` gde je `X` objekat, odnosno vektor čiji su elementi $X_1, \ldots, X_n$. Slično, vrednost $D(\overline{X})$ izračunaćemo kao `mean((X-m)^2)`. Ako želimo da za $D(\overline{X})$ iskoristimo funkciju `var` (koja već postoji u programskom jeziku `R`), onda treba da definišemo novu funkciju `var.p` kao

```
var.p <- function(x){
   return(var(x)*(length(x)-1)/length(x))
}
```

jer funkcija `var` računa popravljenu dispeziju uzorka.

5.2.2 Ocene autokovarijansne funkcije

S obzirom na parnost autokovarijansne funkcije, dovoljno je naći ocene ove funkcije samo za $k \geq 0$. I ovoga puta ćemo morati da razlikujemo dva slučaja i to, kada je srednja vrednost vremenskog niza poznata i kada to nije. Ocenjivanje ćemo vršiti na osnovu n posmatranja samo jedne realizacije.

Kada je srednja vrednost vremenskog niza, m, poznata, statistika

$$\widehat{R_k}(n) = \frac{1}{n-k} \sum_{t=1}^{n-k} (X_t - m)(X_{t+k} - m)$$

je nepristrasna ocena nepoznate autokovarijansne funkcije, što je lako pokazati. Pri tome je neophodno da imamo veći broj posmatranja nego što je korak kovarijanse, tj. ocenjivanje je moguće samo za $0 \leq k < n$.

Razmotrimo sada slučaj nepoznate srednje vrednosti. U tom slučaju se srednja vrednost mora oceniti, a takođe je neophodno da imamo veći broj posmatranja n nego što je korak k. Statistika kojom se ocenjuje autokovarijansna funkcija je

$$\widetilde{R_k}(n) = \frac{1}{n-k} \sum_{t=1}^{n-k} (X_t - \overline{X})(X_{t+k} - \overline{X}) \,.$$

Ova ocena je asimptotski nepristrasana.

Ispitivanje srednjekvadratnih grešaka u oba slučaja ocenjivanja srednje vrednosti zahteva složeniju tehniku i ovde neće biti sprovedeno.

Da bismo izračunali $\widetilde{R_k}(n)$ u programskom jeziku `R` definisaćemo funkciju

```
cov.p <- function(x,k){
   len<-length(x)
   if(k>=len) return(NA_character_)
   return(cov(x[1:len-k], x[k:len)]*(len-1)/len)
}
```

gde je `x` vektor čiju kovarijansu ispitujemo, a `k` korak kovarijanse. Ovde smo iskoristili funkciju `cov` koja postoji u programskom jeziku `R` i daje popravljenu kovarijansu.

5.2.3 Predviđanje vrednosti vremenskog niza

Prilikom praktične primene analize vremenskih nizova često je potrebno sagledati ponašanje niza van vremenskog perioda u kome su vršena posmatranja. Nekada je to i osnovni cilj statističke analize. Vremenski trenuci u kojima nisu dostupne vrednosti realizacija vremenskog niza se mogu odnositi, kako na "budućnost" i "prošlost", tako i na trenutke "unutar" perioda posmatranja. Otkrivanje nepoznate "prošlosti", kao i predviđanje "budućih" vrednosti realizacija vrši se ekstrapolacijom vremenskog niza. Otkrivanje nepoznatih vrednosti unutar perioda posmatranja ostvaruje se interpolacijom. Ovde će biti reči o jednoj vrsti ekstrapolacije vrednosti vremenskog niza, predviđanju. Naime, predviđanje je ekstrapolacija sa korakom unapred. Ilustrovaćemo problem i metod za njegovo rešavanje samo u najjednostavnijem slučaju, tj. izložićemo samo predviđanje za jedan korak unapred.

Pretpostavimo da su nam poznate vrednosti konačnog dela jedne realizacije niza $\{X_t\}$ u "prošlosti", tj. u vremenskim trenucima $t = -n, -n+1, \ldots, -1, 0$. Zadatak je predvideti vrednost procesa u trenutku $t = 1$, tj. vrednost X_1 u "budućem trenutku". Jedan od mogućih načina za rešavanje ovog problema je nalaženje najboljeg linearnog predviđanja u smislu kriterijuma minimalnog srednjekvadratnog odstupanja od prave vrednosti.

Definišimo, dakle, linearni prediktor za X_1 koristeći se poznatom prošlošću vremenskog niza

$$X_{1n} = \sum_{t=-n}^{0} \beta_{tn} X_t \,,$$

gde koeficijente linearnog modela, β_{tn}, treba odrediti tako da

$$E(X_1 - X_{1n})^2 = E\left(X_1 - \sum_{t=-n}^{0} \beta_{tn} X_t\right)^2$$

bude minimalno. Prema našem znanju o najboljem linearnom prediktoru (modeli regresije prve vrste), za slučaj vremenskog niza, dobijamo ocene β^*_{tn} nepoznatih koeficijenata β_{tn} u obliku

$$\beta^*_{tn} = \sum_{j=-n}^{0} R^{tj} R_j \,, \quad t = -n, -n+1, \ldots, -1, 0 \,, \quad \|R^{tj}\|^0_{t,j=-n} = \|R_{|t-j|}\|^{-1} \,,$$

gde je matrica $\|R_{|t-j|}\|_{t,j=-n}^{0}$ matrica odgovarajućih autokovarijansi.

Rešenje ovako postavljenog problema je predviđanje

$$X_{1n}^{*} = \sum_{t=-n}^{0} \beta_{tn}^{*} X_t \,.$$

Srednjekvadratna greška ovog predviđanja je

$$E(X_1 - X_{1n}^{*})^2 = R_0 - \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} R_j R^{ji} R_i$$

zbog parnosti autokovarijansne funkcije. Problem se, naravno, usložnjava kada autokovarijansna funkcija nije poznata pa je treba oceniti. O tome ovom prilikom neće biti reči.

Iz programskog jezika `R` izračunaćemo predviđanje X_{1n}^{*} pozivom

```
> t(b) %*% Xn
```

gde je `b` vektor kolone čiji su elementi β_{tn}^{*}, $t = \overline{-n, 0}$, dok je `Xn` vektor kolone sa elementima X_t, $t = \overline{-n, 0}$.

5.2.4 Vremenski nizovi sa trendom i sezonskom komponentom

Poseban problem u statističkoj analizi vremenskih nizova predstavlja uočavanje i otklanjanje neslučajnih (determinističkih) komponenata vremenskog niza.

Na osnovu dijagrama procesa, grafičkog prikaza poznate realizacije, utvrđuje se da li ima promena u seriji kao što je nagla promena nivoa ili ponavljanje približno istih vrednosti realizacije procesa i sl. Preporučljivo je podeliti niz na homogene segmente i analizirati da li postoje posmatranja koja odudaraju od ostatka serije pri čemu su moguća dva slučaja: prvo, da ta posmatranja potiču od nekog drugog niza i da su greškom registrovana i drugo, da proces ima u sebi neslučajne komponente koje zavise od vremena, u kom slučaju je možda moguće reprezentovati proces modelom

$$X_t = m_t + s_t + \varepsilon_t \tag{5.7}$$

gde je:

-m_t sporo promenljiva funkcija od vremena tzv. **trend**,

-s_t periodična funkcija od vremena sa poznatom periodom d koju zovemo **sezonska komponenta**,

-ε_t je slučajni šum koji je stacionaran u širokom smislu.

Idealno je da niz $\{\varepsilon_t\}$ bude *beli šum*, tj. takav niz čije komponente imaju očekivanje 0 i nekorelirane su među sobom sa konstantnom disperzijom.

Linearni model (5.7) je najjednostavniji model vremenskog niza sa neslučajnim komponentama. Složeniji su, recimo, multiplikativni model, zatim modeli tipa ARMA, ARIMA, ARCH modeli itd., ali se njima ovde nećemo baviti.

Kod nekih vremenskih nizova se uočavaju i dva tipa periodičnosti. Prvi tip se obično vezuje za jednogodišnji period i za njega se najčešće vezuje termin *sezonska komponenta*, dok drugi tip periodičnosti pretpostavlja period kraći od godinu dana, ili višegodišnji period. Za drugi oblik periodičnosti se obično koristi termin *ciklična komponenta* i označava sa c_t.

Cilj statističke analize vremenskih nizova sa neslučajnim komponentama je da se ocene i eliminišu determinističke komponente sa nadom da će se slučajni ostatak, pokazati stacionarnim i samim tim omogućiti primenu teorije stacionarnih slučajnih procesa. Međutim, bilo da je posmatrani vremenski niz stacionaran ili ne, posle eliminacije neslučajnih komponenata moguće je utvrditi svojstva osnovnog procesa $\{X_t\}$ koji bi se nadalje, uz pomoć procesa $\{\varepsilon_t\}$ i neslučajnih komponenata mogao predvideti i kontrolisati. Naglasimo, međutim, da postoje razrađene statističke tehnike i za nestacionaran šum, ali se mi njima nećemo baviti.

Nadalje ćemo se baviti isključivo modelom (5.7) sa stacionarnim šumom.

Da bi se uopšte utvrdilo da posmatrani vremenski niz zadovoljava jednakost (5.7), sprovodi se postupak testiranja statističkih hipoteza nad elementima dela realizacije slučajnog niza. Navešćemo neke od testova.

5.2.5 Otkrivanje neslučajnih komponenata (analiza slučajnosti niza)

Dakle, potrebno je ispitati da li ima i neslučajnih komponenata u vremenskom nizu iz koga je uzet uzorak $X_1, X_2, \ldots, X_n$, na osnovu samo jedne

realizacije tog niza.

Analiza slučajnosti vremenskog niza obavlja se testiranjem nulte hipoteze o slučajnosti:

H_0: Vremenski niz iz koga je uzorak uzet je slučajan;
protiv alternativne hipoteze:

H_1: U vremenskom nizu iz koga je uzorak uzet postoji trend ili sezonska (ciklična) komponenta.

U okviru testiranja statističkih hipoteza govori se i o testiranju slučajnosti uzorka. U tom slučaju je nulta hipoteza, H_0, da je $(X_1, \ldots, X_n)$ prost slučajni uzorak, odnosno, da su sve slučajne promenljive X_t, $t = 1, 2, \ldots n$ nezavisne i sa istom raspodelom čija je funkcija raspodele $F(x)$, $x \in R$,

$$H_0 : F_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = F(x_1)F(x_2)\ldots F(x_n)$$

protiv alternativne

$$H_1 : F_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) \neq F(x_1)F(x_2)\ldots F(x_n)\,.$$

U slučaju da se prihvati alternativna hipoteza H_1, naknadno se proverava o kom se faktoru neslučajnosti radi.

Kada se već zna da kod vremenskog niza osim slučajne komponente mogu da se jave i neslučajne kao što su trend, ciklična i sezonska komponenta, onda je od interesa nultu hipotezu o slučajnosti hronološkog niza testirati protiv neke od alternativa, kao što je postojanje trenda, ciklične ili sezonske komponente.

Nadalje su navedeni neki od testova kojima se vrši testiranje ovih hipoteza. Naglasimo da su navedeni testovi namenjeni testiranju slučajnosti kod obeležja apsolutno neprekidnog tipa.

Test tačaka zaokreta

Posmatra se **hronološki** uzorak $(X_1, X_2, \ldots, X_n)$. Na osnovu njega se definišu slučajne veličine $Y_2, Y_3, \ldots, Y_{n-1}$:

$$Y_j = \begin{cases} 1, & \text{ako je } X_j \text{ veće od oba suseda : } X_{j-1} \text{ i } X_{j+1}, \\ & \text{ili ako je } X_j \text{ manje od oba suseda} \\ 0, & \text{u ostalim slučajevima} \end{cases} .$$

Očigledno da je $Y_j = 1$ kada uzorak ima (lokalnu) ekstremnu tačku. Test statistika Z_n se definiše kao ukupan broj ekstremnih tačaka u uzorku, tj. $Z_n = Y_2 + Y_3 + \ldots + Y_{n-1}$.

Ako je hipoteza o slučajnosti tačna, statistika Z_n ima sledeće numeričke karakteristike

$$E(Z_n) = \frac{2(n-2)}{3}, \quad D(Z_n) = \frac{16n-29}{90}.$$

Ako se u realizovanom uzorku obima n dobije sredina uzorka koja bitno odstupa od $E(Z_n)$, nulta hipoteza se odbacuje.

Za veliki obim uzorka ($n \geq 50$) koristi se test statistika

$$Z_n^* = \frac{Z_n - \frac{2(n-2)}{3}}{\sqrt{\frac{16n-29}{90}}}$$

koja ima približno normalnu normiranu raspodelu pomoću koje se nulta hipoteza testira na uobičajeni način:

H_0	H_1	C
Niz je slučajan	Postoji neslučajna komponenta	$\lvert z_n^* \rvert \geq z_{0,5-\alpha/2}$

,

gde konstanta $z_{0,5-\alpha/2}$ zadovoljava uslov

$$P\left\{|Z_0| \leq z_{0,5-\alpha/2}\right\} = P\left\{|Z_0| \leq z_{\frac{1-\alpha}{2}}\right\} = 1-\alpha$$

pri čemu slučajna promenljiva Z_0 ima $\mathcal{N}(0,1)$ raspodelu. Navedeni test tačaka zaokreta poznat je još i kao *test povratnih* ili *ekstremnih tačaka*. Ovaj test daje bolje rezultate kada je alternativna hipoteza postojanje periodične (sezonske, ciklične) komponente u vremenskom nizu, nego kada je alternativa postojanje trenda.

Test tačaka zaokreta se ne nalazi u listi osnovnih funkcija programskog jezika `R`, pa je za njegov poziv potrebno uključiti neke dodatne pakete. Umesto toga, sami ćemo implementirati funkciju koja računa vrednost testa, kao i njegovu p-vrednost.

```
test.tacaka.zaokreta <- function(x){
   len=length(x)
```

```
    diffNext = sign(diff(x)[2:(len-1)])
    diffPrev = sign(-diff(x)[1:(len-2)])
    y=diffNext * diffPrev
    z=sum(y[y>0])
    e=2*(len-2)/3
    d=(16*len-29)/90
    test=(z-e)/sqrt(d)
    tmp = pnorm(test)
    pValue = 2*min(tmp,1-tmp)
    result = list(test.statistika = test, p.vrednost = pValue)
    return(result)
}
```

Ako je broj tačaka zaokreta u posmatranom nizu, tj. realizaciji od n elemenata, blizak broju $n-2$, tj. maksimalnom mogućem broju, takva pojava će ukazivati na postojanje cikličnih varijacija u nizu. Ako je pak broj tačaka zaokreta 0, suočavamo se sa postojanjem trenda (rastućeg ili opadajućeg). Naravno da će se u oba slučaja hipoteza o slučajnosti odbacivati.

Primer 5.1. Neka je dat hronološki niz koji predstavlja srednje godišnje proticaje jedne reke (u milimetrima) za period od 61 godine na jednoj vodomernoj stanici:

1	2	3	4	5	6	7	8	9	10	11
3456	2622	2701	2558	3033	3457	4197	3200	3379	3782	4169
12	13	14	15	16	17	18	19	20	21	22
3136	2272	3834	3292	2453	2240	3341	2331	2393	3068	2942
23	24	25	26	27	28	29	30	31	32	33
2445	2924	3467	3134	2889	3093	2743	2917	2581	2945	2590
34	35	36	37	38	39	40	41	42	43	44
2447	4312	3915	3442	2613	2289	3748	2197	2576	2382	3098
45	46	47	48	49	50	51	52	53	54	55
3326	2378	2912	2584	3157	3098	3068	2852	2477	2268	2558
56	57	58	59	60	61					
2468	3145	2970	2559	2195	2381					

Testirati hipotezu da je niz slučajan sa pragom značajnosti $\alpha = 0,05$.

Obim uzorka je $n = 61$, te je ispunjen uslov za normalnu aproksimaciju. Dobijamo da je

$$E(Z_n) = 39,33\,, \quad D(Z_n) = 10,52\,, \quad \sigma(Z_n) = 3,24,$$

a realizovana vrednost test statistike je $z_n = 37$, a $z_n^* = -0.72$. S obzirom na to da je kritična oblast za zadati prag značajnosti $C = (-\infty,\, -1,96) \cup (1,96\,,\, \infty)$, možemo zaključiti da nema razloga da odbacimo hipotezu o slučajnosti, odnosno da možemo smatrati da je niz slučajan.

U programskom jeziku `R`, rezultat ćemo dobiti pozivom gore navedene funkcije `test.tacaka.zaokreta`. Najpre definišemo vektor `tok` čiji su elementi vrednosti iz tabele. Nakon poziva funkcije, dobićemo

```
> test.tacaka.zaokreta(tok)
$‘test.statistika‘
[1] -0.7193215
$ p.vrednost
[1] 0.4719428
```

pa na osnovu p-vrednosti zaključujemo da prihvatamo nultu hipotezu. $\triangle$

Ficov test

Ovaj test je još poznat i kao *test promena znakova ispod i iznad medijane.*

Neka je Me medijana u nizu $X_1, X_2, \ldots, X_n$. Definišimo slučajnu promenljivu

$$M_i = \begin{cases} 1, & \text{ako je} \quad (X_i > Me \wedge X_{i-1} < Me) \vee (X_i < Me \wedge X_{i-1} > Me) \\ 0, & \text{inače} \end{cases} .$$

Definišimo dalje slučajnu promenljivu $S_n = M_1 + M_2 + \cdots + M_n$, ukupan broj preskoka medijane u (hronološkom) uzorku.

Naglasimo da se realizovani uzorak **ne sme sređivati** za potrebe ovog testiranja sem u momentu pronalaženja medijane.

Statistika S_n, za veliko n, ima približno normalnu raspodelu sa parametrima

$$E(S_n) = \frac{n-1}{2}, \quad D(S_n) = \frac{n-1}{4},$$

pa

$$S_n^* = \frac{S_n - \frac{n-1}{2}}{\sqrt{\frac{n-1}{4}}}$$

ima približno normalnu normiranu raspodelu, te se ponovo mogu da koriste kvantili ove raspodele ili

H_0	H_1	C
Niz je slučajan	Postoji neslučajna komponenta	$\lvert s_n^* \rvert \geq z_{0,5-\alpha/2}$

.

Ficov test implementiraćemo u programskom jeziku R na sledeći način

```
test.ficov<-function(x){
   len=length(x)
   m = median(x)
   s=0
   for(i in 2:len){
      if((x[i]>m && x[i-1]<m) || (x[i]<m && x[i-1]>m))
         s=s+1
   }
   e=(len-1)/2
   d=(len-1)/4
   test=(s-e)/sqrt(d)
   tmp = pnorm(test)
   pValue = 2*min(tmp,1-tmp)
   result = list(test.statistika = test, p.vrednost = pValue, sn=s)
   return(result)
}
```

Primer 5.2. Na podatke o srednjim godišnjim proticajima iz prethodnog primera i sa istim pragom značajnosti, primeniti Ficov test.

Medijana realizovanog uzorka je $m_e = 2917$, $n = 61$, $(n-1)/2 = 30$, $(n-1)/4 = 15$, $s_n = 25$ i $s_n^* = -1,29$, te se ponovo može zaključiti da je niz slučajan.

Pozivom funkcije `test.ficov` za vektor `tok`, koji je formiran kao u prethodnom primeru, dobijamo sledeći rezultat

```
> test.ficov(tok)
$‘test.statistika‘
[1] -1.290994
$p.vrednost
[1] 0.1967056
$sn
[1] 25
```

△

Test rasta

Nasuprot testu tačaka zaokreta, postoji test pogodan za otkrivanje trenda u vremenskom nizu. Test statistika se tada definiše preko tačaka rasta.

Tačkom rasta naziva se element X_j uzorka $(X_1, X_2, \ldots, X_n)$, za koji važi da je $X_j < X_{j+1}$, za $j = 1, 2, \ldots, n-1$.

Uz pomoć tačaka rasta definišu se slučajne veličine

$$Y_j = \begin{cases} 1, & \text{ako je } X_j < X_{j+1} \\ 0, & \text{u ostalim slučajevima} \end{cases} .$$

Test statistika R_n je broj tačaka rasta u uzorku obima n :
$R_n = Y_1 + Y_2 + \ldots + Y_{n-1}$.

Ova statistika ima sledeće numeričke karakteristike

$$E(R_n) = \frac{n-1}{2}, \quad D(R_n) = \frac{n+1}{12},$$

odakle sledi da se nulta hipoteza odbacuje ako je broj tačaka rasta u realizovanom uzorku bitno različit od broja $E(R_n)$.

Pogodnost statistike R_n je u tome što već za obim uzorka $n > 12$ statistika

$$R_n^* = \frac{R_n - \frac{n-1}{2}}{\sqrt{\frac{n+1}{12}}}$$

ima približno normalnu raspodelu $\mathcal{N}(0,1)$, pa se za testiranje nulte hipoteze koristi tablica normalne normirane raspodele i uobičajena kritična oblast:

H_0	H_1	C
Niz je slučajan	Postoji neslučajna komponenta (trend)	$\lvert r_n^* \rvert \geq z_{0,5-\alpha/2}$

Ako je u nizu rastući trend, onda u nizu ima $(n-1)$ tačka rasta, a ako je u pitanju opadajući trend onda i nema tačaka rasta. Ukoliko je registrovano naizmenično rašćenje i opadanje trajektorije, svaka druga vrednost u nizu će biti tačka rasta i biće ih (približno) ukupno $0,5n$. Iz toga se i može videti da je kriterijum broja tačaka rasta efikasan kod testiranja slučajnosti samo u slučaju da je alternativna hipoteza, odnosno u pitanju, rastići ili opadajući trend.

Ovde još treba napomenuti da je moguće, odnosno da se do istih zaključaka dolazi ako se umesto tačaka rasta koriste tačke opadanja.

Za test rasta implementiraćemo funkciju `test.rasta` u programskom jeziku `R` koja računa vrednost testa, kao i njegovu p-vrednost.

```
test.rasta <- function(x){
   len=length(x)
   diffNext = sign(diff(x))
   z=sum(diffNext[diffNext>0])
   e=(len-1)/2
   d=(len+1)/12
   test=(z-e)/sqrt(d)
   tmp = pnorm(test)
   pValue = 2*min(tmp,1-tmp)
   result = list(test.statistika = test, p.vrednost = pValue)
   return(result)
}
```

Primer 5.3. Za podatke iz Primera 5.1 izvršiti testiranje o slučajnosti primenom testa rasta. Testiranje izvršiti sa pragom značajnosti $\alpha = 0,05$.

Za posmatrani realizovani uzorak je $r_n = 26$, $E(R_n) = 30$, $D(R_n) = 5,167$ i $r_n^* = -1,76$. Dakle, i primenom ovog testa, testa rasta, koji bolje "prepoznaje" postojanje trenda nego, recimo, test tačaka zaokreta, ne možemo da odbacimo hipotezu o slučajnosti. Zaključujemo, niz je slučajan. Primetimo, međutim, da je ovog puta realizovana vrednost test statistike bila dosta blizu granice

kritične oblasti!

Pozivom funkcije `test.rasta` koju smo implementirali u programskom jeziku `R` dobićemo rezultat

```
> test.rasta(tok)
$ ‘test.statistika‘
[1] -1.759765
$ p.vrednost
[1] 0.0784476
```

gde je `tok` vektor čiji su elementi vrednosti iz tabele date u Primeru 5.1. △

Test kvadrata uzastopnih razlika

Test kvadrata uzastopnih razlika primenjuje se kada je raspodela slučajnih promenljivih koje čine niz $\{X_n\}$ normalna i alternativna hipoteza postojanje trenda.

Ovaj test bazira se na statistici

$$D = \frac{\Delta^2}{2\widetilde{S}_n^2},$$

gde je

$$\Delta^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}(X_{i+1}-X_i)^2, \quad \widetilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2, \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n}X_i.$$

Statistika D ima očekivanje 1 i disperziju $\dfrac{n-2}{n^2-1}$, pa se za dovoljno veliki obim uzorka ($n \geq 20$) raspodela statistike

$$D^* = \frac{D-1}{\sqrt{\dfrac{n-2}{n^2-1}}}$$

može da aproksimira normalnom normiranom raspodelom, te je najbolja kritična oblast veličine α:

H_0	H_1	C
Niz je slučajan	Postoji trend	$\lvert d^*\rvert \geq z_{0,5-\alpha/2}$

.

Slično kao i kod prethodna dva testa, implementiraćemo i test kvadrata uzastopnih razlika u programskom jeziku R. Funkcija `test.kvadrata.razlika` računa realizovanu vrednost test statistike kao i p-vrednost testa.

```
test.kvadrata.razlika <- function(x){
    len = length(x)
    delta = sum(diff(x)^2)/(len-1)
    mu = mean(x)
    S = var(x)
    D = delta/(2*S)
    e=1
    d=(len-2)/(len^2-1)
    test=(D-e)/sqrt(d)
    tmp = pnorm(test)
    pValue = 2*min(tmp,1-tmp)
    result = list(test.statistika = test, p.vrednost = pValue)
    return(result)
}
```

Test serijskih korelacija

Test serijskih korelacija odnosi se na proveru uzoračkih serijskih korelacija, tj. bazira se na statistici

$$r_k = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k}\left[\left(X_i - \overline{X}_{n-k}\right)\left(X_{i+k} - \overline{X}_{k+(n-k)}\right)\right]}{\sqrt{\frac{1}{n-k}\sum_{i=1}^{n-k}\left(X_i - \overline{X}_{n-k}\right)^2 \frac{1}{n-k}\sum_{i=1}^{n-k}\left(X_{i+k} - \overline{X}_{k+(n-k)}\right)^2}},$$

gde je

$$\overline{X}_{n-k} = \frac{1}{n-k}\sum_{i=1}^{n-k} X_i\,, \quad \overline{X}_{k+(n-k)} = \frac{1}{n-k}\sum_{i=1}^{n-k} X_{k+i}\,.$$

U praksi se , međutim, statistika r_k zamenjuje statistikom koja se najčešće

isto označava:

$$r_k = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k}\left[\left(X_i - \overline{X}_{n-k}\right)\left(X_{i+k} - \overline{X}_{k+(n-k)}\right)\right]}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2},$$

ili čak sa:

$$r_k = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k}\left[\left(X_i - \overline{X}_n\right)\left(X_{i+k} - \overline{X}_n\right)\right]}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2},$$

za veliko n i malo k u odnosu na n, gde je $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (mada za ovo nema čvršćeg teorijskog opravdanja).

Testiranje slučajnosti počinje izračunavanjem uzoračkih serijskih korelacija $r_1, r_2, \ldots$ i njihovom analizom. Kod slučajnog uzorka (kod koga nema trenda) je $r_1 = r_2 = \ldots = 0, (r_0 = 1)$ te je za prihvatanje nulte hipoteze potrebno da ove vrednosti za $k = 1, 2, \ldots$ kod realizovanog uzorka budu jednake (tj. veoma bliske) nuli. Ako je to ispunjeno, što znači da je nulta hipoteza tačna, tada je

$$E(r_1) = -\frac{1}{n-1}, \quad \text{a} \quad D(r_1) = \frac{(n-2)^2}{(n-1)^3}$$

pod pretpostavkom da je uzorak iz normalne raspodele.

U tom, vrlo uprošćenom slučaju, testiranje slučajnosti se sprovodi statistikom

$$r_1^* = \frac{r_1 + \frac{1}{n-1}}{\frac{n-2}{\sqrt{(n-1)^3}}}$$

koja ima približno normalnu normiranu raspodelu.

Određivanje najbolje kritične oblasti veličine α je uobičajeno:

H_0	H_1	C
Niz je slučajan	Postoji trend	$\lvert r_1^* \rvert \geq z_{0,5-\alpha/2}$

.

U praksi se dosta često koristi i statistika

$$r_k^{'} = \frac{\sum_{i=1}^{n-k} \left(X_i - \overline{X}_n\right)\left(X_{i+k} - \overline{X}_n\right)}{\sum_{i=1}^{n-k} \left(X_i - \overline{X}_n\right)^2} \tag{5.8}$$

za koju, ako je hipoteza o slučajnosti niza tačna, važi da približno ima normalnu raspodelu $\mathcal{N}\left(\frac{-1}{n-k-1}, \frac{n-k-2}{(n-k-1)^2}\right)$, što znači da je pri velikom n i malom k raspodela za $r_k^{'}$ približno $\mathcal{N}\left(0,\, 1/n\right)$.

Kao posledica raspodele statistike (5.8), statistika

$$r_k^{'*} = \frac{r_k^{'} + \dfrac{1}{n-k-1}}{\sqrt{\dfrac{n-k-2}{(n-k-1)^2}}} \tag{5.9}$$

ima približno normalnu normiranu raspodelu.

Kao i kod prethodnih testova, implementiraćemo i test serijskih korelacija u programskom jeziku R, a funkcija za ovaj test glasi

```
test.serijskih.korelacija <- function(x,k){
   len = length(x)
   if(k>=len)
      return(NA_character_)
   mu1 = mean(x[1:(len-k)])
   mu2 = mean(x[(k+1):len])
   x1 = (x-mu1)[1:(len-k)]
   x2 = (x-mu2)[(k+1):len]
   var1 = varp(x[1:(len-k)])
   var2 = varp(x[(k+1):len])
   r=sum(x1*x2)/(len-k)/sqrt(var1*var2)
   e=-1/(len-k-1)
   d=(len-k-2)/(len-k-1)^2
   test=(r-e)/sqrt(d)
   tmp = pnorm(test)
   pValue = 2*min(tmp,1-tmp)
   result = list(test.statistika = test, p.vrednost = pValue, rk=r)
   return(result)}
```

U kodu smo koristili funkciju `varp` koja računa dispeziju populacije, a implementirali smo je na sledeći način

```
varp <- function(x) {
   return(mean((x-mean(x))^2))
}
```

Takođe, implementiraćemo pojednostavljenu statistiku koja je data jednačinom (5.9).

```
test.serijskih.korelacija.simple <- function(x,k){
   len = length(x)
   if(k>=len)
      return(NA_character_)
   mu = mean(x)
   x1 = (x-mu)[1:(len-k)]
   x2 = (x-mu)[(k+1):len]
   r=sum(x1*x2)/sum(x1^2)
   e=-1/(len-k-1)
   d=(len-k-2)/(len-k-1)^2
   test=(r-e)/sqrt(d)
   tmp = pnorm(test)
   pValue = 2*min(tmp,1-tmp)
   result = list(test.statistika = test, p.vrednost = pValue, rk=r)
   return(result)
}
```

Primer 5.4. Na podacima iz Primera 5.1 testirati hipotezu o slučajnosti sa pragom značajnosti $\alpha = 0,05$ primenom statistike (5.9).

Dobijamo sledeće rezultate sumirane u narednoj tabeli

Korak (k)	0	1	2	3	4	5	6	7	8	9	10
$r_k^{'}$	1	0,25	-0,03	0,05	0,19	0,04	-0,21	0,08	-0,04	-0,1	0,06
$r_k^{'*}$	7,94	2,05	-0,12	0,54	1,6	0,47	-1,48	0,74	-0,18	-0,6	0,58
p-vrednost	2e-15	0,04	0,91	0,59	0,11	0,63	0,14	0,46	0,85	0,54	0,56

Analizirajući ove rezultate, možemo zaključiti da su realizovane vrednosti za $r_k^{'*}$ sa velikim verovatnoćama bliske nuli i da su sa malim verovatnoćama velike. To, ponovo, ukazuje da je (sa velikom verovatnoćom) niz slučajan. Dakle, na

osnovu podataka iz ove tabele, možemo da zaključimo da je samo vrednost za $r_1^{'*}$ značajna, odnosno samo se ona značajno razlikuje od nule, dok su ostale p–vrednosti mnogo veće od zadatog praga značajnosti. Štaviše, ako bismo prag značajnosti spustili na 0,01 mogli bismo bez rezerve da zaključimo da se radi o slučajnom nizu. △

Metod pokretnih sredina

Metod pokretnih sredina ili kliznih proseka, primenjuje se u situacijama kada je alternativna hipoteza slučajnosti niza postojanje trenda. Metod pokretnih sredina ne predstavlja uobičajeni postupak testiranja hipoteza. Ovaj metod samo pomaže da se trend lakše uoči ako zaista postoji. Najčešće, uz njega se crta i dijagram.

Za seriju od n podataka vrši se izravnavanje (izglađivanje) pomoću usrednjavanja $m = 2k + 1$ članova. Formiraju se proseci

$$\begin{aligned} Y_1 &= \tfrac{X_1+X_2+\ldots+X_{k-1}+X_k+\ldots+X_{2k+1}}{2k+1}\,, \\ Y_2 &= \tfrac{X_2+X_3+\ldots+X_k+X_{k+1}+\ldots+X_{2k+2}}{2k+1}\,, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\,, \\ Y_{n-2k} &= \tfrac{X_{n-2k}+\ldots+X_{n-k}+\ldots+X_n}{2k+1}\,. \end{aligned}$$

Formirani niz $Y_1, \ldots, Y_{n-2k}$ ima manja kolebanja nego originalni niz. To je razlog da se trend lakše uočava i odstranjuje. Treba uočiti da su Y_i i Y_j nekorelirani među sobom ako je $|j - i| > m$.

Opšti način izravnavanja u okviru ovog metoda je pomoću pozitivnih pondera koji ispunjavaju uslov

$$\sum_{j=-k}^{k} c_j = 1\,, \quad c_j > 0\,, \quad j = -k, -k+1, \ldots, k\,.$$

U ovom slučaju, niz se formira na sledeći način:

$$\widetilde{Y}_t = \sum_{j=-k}^{k} c_j X_{t+j}\,, \quad t = k+1, k+2, \ldots, n-k\,,$$

a opet je u pitanju izravnavanje sa po $m = 2k + 1$ tačaka.

U oba slučaja izbor broja k zaslužuje posebnu pažnju i neće biti predmet našeg detaljnijeg razmatranja. Ipak napomenimo da se polazeći od $k = 1$ pa nadalje proveravaju realizovane vrednosti uzoračkih disperzija statistika Y_i i uočava momenat njihove stabilizacije oko neke konstante. Taj momenat određuje vrednost broja k koju treba prihvatiti.

Niz pokretnih sredina $Y_1, \ldots, Y_{n-2k}$ dobićemo u programskom jeziku `R` pokretanjem sledećeg koda

```
pokretne.sredine<-function(x,k){
   y=c()
   len = length(x)
   for(i in 1:(len-2*k)){
      y[i]=mean(x[i:(2*k+i)])
   }    return(y)
}
```

5.2.6 Otklanjanje neslučajnih komponenata

Pošto se gore navedenim testovima utvrdi postojanje neslučajnih komponenata u vremenskom nizu, pristupa se njihovom otklanjanju. Navodimo neke od postupaka za tu namenu.

5.2.7 Eliminacija trenda kod procesa bez sezonske komponente

U odsustvu sezonske komponente, proces (5.7) postaje

$$X_t = m_t + \varepsilon_t\,, \quad t = 1, \ldots, n\,, \tag{5.10}$$

gde bez smanjenja opštosti možemo smatrati da je $E\varepsilon_t = 0$. Za eliminaciju trenda u ovom slučaju primenićemo nekoliko sledećih postupaka.

Ocena najmanjih kvadrata za m_t

Ovaj metod pretpostavlja fitovanje neslučajne komponente parametarskom familijom funkcija, na primer

$$m_t = a_0 + a_1 t + a_2 t^2 \,,$$

određujući parametre a_0, a_1 i a_2 po metodu najmanjih kvadrata iz uslova da $\sum_t (x_t - m_t)^2$ bude minimalno. Ocenjene vrednosti za ε_t su razlike x_t i $\hat{m}_t = \hat{a}_0 + \hat{a}_1 t + \hat{a}_2 t^2$.

Izglađivanje pokretnom sredinom

Neka je q nenegativan ceo broj, i posmatrajmo tzv. **dvostranu pokretnu sredinu** niza $\{X_t\}$

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j} \,.$$

Sledi da je

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} + \frac{1}{2q+1} \sum_{j=-q}^{q} \varepsilon_{t+j} \approx m_t \quad \text{za} \quad q+1 \le t \le n-q \,,$$

ukoliko je tačna pretpostavka da je m_t približno linearna na intervalu $[t-q, t+q]$ i da je $E\varepsilon_t \approx 0$ na ovom intervalu. Ocena trenda je

$$\hat{m}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j} \,, \quad q+1 \le t \le n-q \,.$$

Pošto nemamo opservacije X_t za $t \le 0$ i $t > n$, ne možemo ovu aproksimaciju da primenimo za $t \le q$ ili $t > n-q$.

Umesto navedenih, moguće je koristiti jednostrane pokretne sredine

$$\hat{m}_t = \sum_{j=0}^{n-t} \alpha(1-\alpha)^j X_{t+j} \,, \quad t = 1, \ldots, q$$

i

$$\hat{m}_t = \sum_{j=0}^{t-1} \alpha(1-\alpha)^j X_{t-j} \,, \quad t = n-q+1, \ldots, n \,,$$

gde je $0 < \alpha < 1$ realan broj.

Ove ocene nisu previše osetljive na promenu α. Empirijski je dokazano da se najbolje ocene trenda dobijaju za α između $0,1$ i $0,3$.

Primena operatora razlike za generisanje podataka

Operator prve razlike za posmatrani vremenski niz, u oznaci ∇, je definisan kao

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t \,, \tag{5.11}$$

gde je B **operator pomeranja unazad** (za jedan korak),

$$BX_t = X_{t-1} \,.$$

Stepeni operatora B i ∇ se definišu na sledeći način

$$B^j(X_t) = X_{t-j} \,, \quad \nabla^j(X_t) = \nabla(\nabla^{j-1}(X_t)) \,, \quad j \geq 1 \,,$$

$$\nabla^0(X_t) = X_t \,.$$

Sa polinomima od B i ∇ se računa istovetno kao i sa realnim polinomima. Na primer,

$$\begin{aligned} \nabla^2 X_t = \nabla(\nabla X_t) &= \nabla(X_t - X_{t-1}) = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = \\ &= X_t - 2X_{t-1} + X_{t-2} \,. \end{aligned}$$

Suština je u tome da ako se ∇ primeni na linearni trend $m_t = at + b$, dobijamo konstantnu funkciju $\nabla m_t = a$. Na taj način se bilo koji polinomni trend stepena k može redukovati na konstantu primenom ∇^k.

Dakle, za $X_t = m_t + \varepsilon_t$ gde je $m_t = \sum_{j=0}^{k} a_j t^j$ i gde je $E\varepsilon_t = 0$ primenom operatora razlike dobijamo stacionaran proces (sa očekivanjem $k!a_k$):

$$\nabla^k X_t = k!a_k + \nabla^k \varepsilon_t \,.$$

To omogućava da, polazeći od proizvoljnog zadatog niza $\{X_t\}$ podataka, više-

strukom uzastopnom primenom operatora ∇, dođemo do niza $\{\nabla^k X_t\}$ koji se može modelirati kao da je jedna realizacija stacionarnog procesa. U praksi se pokazalo da k najčešće nije veliko, tj. da je jedan ili dva.

5.2.8 Istovremena eliminacija trenda i sezonske komponente

Posmatrajmo ponovo model (5.7)

$$X_t = m_t + s_t + \varepsilon_t\,, \quad E\varepsilon_t = 0\,,$$

kod koga je sezonska komponenta sa periodom d, tj.

$$s_{t+d} = s_t\,, \quad \sum_{t=1}^{d} s_t = 0\,.$$

Kod pojave cikličnih komponenata zgodno je podatke indeksirati godinom i mesecom, tj. označićemo sa $X_{j,k}$ podatak koji odgovara j-toj godini, $j = 1,\ldots,n$ i k-tom mesecu $k = 1,\ldots,d$.

Ponovo nam je cilj uklanjanje neslučajnih komponenata, pri čemu sada imamo dve neslučajne komponente. Predstavićemo metode kojima se istovremeno uklanjaju obe.

Metod malog trenda

Činjenica da je trend mali ukazuje na približno konstantnu vrednost za m_t u okviru svakog pojedinog ciklusa, pri svakom t. Izložićemo u kratkim crtama metod koji se može da primeni na vremenski niz kod koga se uočava mali trend.

Kako je $\sum_{k=1}^{d} s_k = 0$, to vodi do uobičajene nepristrasne ocene za m_j kao matematičkog očekivanja za X_j:

$$\widehat{m}_j = \frac{1}{d}\sum_{k=1}^{d} X_{j,k}\,,$$

a ocena za s_k, $k = 1, \ldots, d$, je

$$\widehat{s}_k = \frac{1}{n} \sum_{j=1}^{n} (X_{j,k} - \widehat{m}_j) .$$

Dakle, ocena slučajne komponente (greške, šuma) za k-ti mesec j-te godine je:

$$\widehat{\varepsilon}_{j,k} = X_{j,k} - \widehat{m}_j - \widehat{s}_k , \quad j = 1, \ldots, n , \quad k = 1, \ldots, d .$$

Ocena pokretnim sredinama

Ovaj metod je primenljiviji od predhodnog, jer ne pretpostavlja da je trend u toku jednog ciklusa približno konstantan.

Pretpostavimo da imamo posmatranja $\{X_1, \ldots, X_n\}$. U prvom koraku se ocenjuje trend primenom pokretnih sredina koje odstranjuju sezonsku komponentu i prigušuju šum.

Ako je period d paran, $d = 2q$, tada koristimo

$$\hat{m}_t = \frac{1}{d} \left(\frac{1}{2} X_{t-q} + X_{t-q+1} + \cdots + X_{t+q-1} + \frac{1}{2} X_{t+q} \right) , \quad q + 1 \leq t \leq n - q .$$

Ako je period d neparan, $d = 2q + 1$, koristimo običnu pokretnu sredinu:

$$\hat{m}_t = \frac{1}{d} \sum_{j=-q}^{q} X_{t+j} , \quad q + 1 \leq t \leq n - q .$$

U drugom koraku ocenjujemo sezonsku komponentu. Za svako $k = 1, \ldots, d$ izračunavamo sredinu w_k odstupanja $\{ (X_{k+jd} - \hat{m}_{k+jd}) , \quad q < k + jd \leq n - q \}$:

$$w_k = \frac{1}{n - 2q} \sum_{k+jd=q+1}^{n-q} (X_{k+jd} - \hat{m}_{k+jd}) .$$

Pošto ove aritmetičke sredine ne daju obavezno u zbiru nulu, sezonsku komponentu ocenjujemo sa

$$\widehat{s}_k = w_k - \frac{1}{d} \sum_{i=1}^{d} w_i \quad \text{za} \quad k = 1, \ldots, d \quad \text{i} \quad \widehat{s}_k = \widehat{s}_{k-d} \quad \text{i} \quad k > d .$$

Uklanjanjem sezonskog uticaja iz podataka, dobijamo novi niz podataka

koji nema sezonsku komponentu:

$$d_t = X_t - \widehat{s}_t\,, \quad t = 1, \ldots, n\,.$$

Konačno, ponovo ocenimo trend na osnovu niza $\{d_t\}$ nekim od već pomenutih metoda i dobijamo ocenu šuma:

$$\widehat{\varepsilon}_t = X_t - \widehat{\widehat{m}}_t - \widehat{s}_t\,,$$

gde je $\widehat{\widehat{m}}_t$ novodobijena ocena trenda.

Primena razlike sa korakom d

Kod ovog metoda se primenjuje **operator razlike sa korakom** d u oznaci ∇_d definisan kao:

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t\,,$$

gde je B, kao i malopre, operator pomeranja unazad za jedan korak, a B^d je njegov stepen. Pri tome treba razlikovati operator ∇_d od d–tog stepena ∇^d operatora prve razlike definisanog u (5.11). Primenom ovog operatora na model $X_t = m_t + s_t + \varepsilon_t$, gde $\{s_t\}$ ima period d, dobijamo

$$\nabla_d X_t = m_t - m_{t-d} + \varepsilon_t - \varepsilon_{t-d}$$

koji dovodi do toga da imamo niz $\{\nabla_d X_t\}$ koji ima samo trend oblika $m_t - m_{t-d}$ i šum $\varepsilon_t - \varepsilon_{t-d}$. Sada se trend $m_t - m_{t-d}$ može da eliminiše nekim već ranije pomenutim metodom.

5.2.9 Modeli autoregresije i pokretnih sredina vremenskih nizova

Pre nego što se upoznamo sa ova dva osnovna modela kojima se u teoriji modeliraju elementi vremenskih nizova, navešćemo još neke pojmove koji se koriste pri modeliranju elemenata vremenskih nizova.

Rekli smo da je $R_k = Cov(X_{t+k}, X_t)$, $t, k \in Z$ autokovarijansna funkcija stacionarnog vremenskog niza $\{X_t\}$. Označićemo sa $\{\rho_k\}$ njegovu *autokorela-*

cionu funkciju definisanu sa

$$\rho_k = \frac{R_k}{R_0}\,.$$

Grafik ove funkcije se naziva *korelogram*. Primetimo da je zbog $R_k = R_{-k}$, korelogram simetričan u odnosu na ordinatnu osu pa se grafički prikazuje samo deo za $k > 0$ (za $k = 0$ je $\rho_0 = 1$).

Ako je $\{\varepsilon_t\}$ beli šum, neka je σ^2 disperzija njegovih elemenata. Dakle, za elemente belog šuma važi

$$R_k = \begin{cases} \sigma^2, & k = 0 \\ 0, & k \neq 0 \end{cases} \quad \text{i} \quad \rho_k = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases} \quad .$$

Pri modeliranju vremenskih nizova dva su osnovna modela koji se koriste: *model pokretnih sredina* ili kako se još naziva *model pokretnih proseka* i *model autoregresije*. Svi ostali, složeniji, modeli izvedeni su iz ova dva modela.

Model pokretnih sredina

Ključnu ulogu u modeliranju vremenskih nizova ima Uoldova[1] teorema dekompozicije koju navodimo bez dokaza, a da bismo je naveli potrebna nam je i sledeća definicija:

DEFINICIJA 5.10. *Slabo stacionaran proces* $\{X_t,\, t \in Z\}$ *je* **(linearno) deterministički** *ili* **čisto deterministički** *ako zadovoljava relaciju*

$$P\{(X_t | X_{t-1}, X_{t-2}, \ldots) = X_t\} = 1\,.\diamondsuit$$

U vezi sa nedeterminističkim procesima navodimo sledeću teoremu bez dokaza:

Teorema 5.1 *(Uoldova teorema dekompozicije) Bilo koji slabo stacionaran proces* $\{X_t\}$ *sa očekivanjem nula a koji nije čisto deterministički može da se napiše u obliku zbira*

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + d_t\,, \tag{5.12}$$

[1] Wold

gde je

1. $\psi_0 = 1$, $\sum_{j=1}^{\infty} \psi_j^2 < \infty$,

2. $\{\varepsilon_t\}$ *je beli šum sa očekivanjem nula i disperzijom* σ^2,

3. $\{\psi_j\}$ *i* $\{\varepsilon_t\}$ *su jedinstveni,*

4. $\{d_t, t \in Z\}$ *je deterministički,*

5. ε_t *je granica linearne kombinacije za* X_s, $s \leq t$ *i*

6. $E(d_t \varepsilon_s) = 0$ *za svako* t *i* s.□

Posmatraćemo beskonačnu sumu iz jednakosti (5.12). U Teoremi 5.1 smo koristili proces $\{X_t\}$ sa srednjom vrednošću nula bez smanjenja opštosti. Teoremu smo mogli da iskažemo koristeći proizvoljan realan broj μ za srednju vrednost niza $\{X_t\}$. Ako je stacionaran niz $\{X_t\}$ sa očekivanjem μ čisto nedeterministički, iz Teoreme 5.1 sledi da se on može zapisati u obliku

$$X_t = \mu + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}\,, \quad \psi_0 = 1\,, \tag{5.13}$$

gde je $\{\varepsilon_t\}$ beli šum i ψ_j, $j = 1, 2, \ldots$ su (jedinstveni) realni brojevi. Niz definisan izrazom (5.13) naziva se *linearan proces*. Linearan proces se još naziva *pokretna sredina* ili *pokretni prosek* i obeležava se skraćeno sa MA[2] proces (niz). Koristi se, pogotovu u tehnici, i naziv *linearni filter* kojim se, zapravo, želi da naglasi da se jedan, ulazni proces, u posmatranom sistemu transformiše ili filtrira u drugi, izlazni proces.

Primetimo da X_s ne zavisi od ε_t ukoliko je $s < t$.

Koristeći operator pomeranja unazad, linarni proces (5.13) možemo da napišemo kraće kao

$$X_t - \mu = (1 + \psi_1 B + \psi_2 B^2 + \cdots)\varepsilon_t = \Psi(B)\varepsilon_t\,, \tag{5.14}$$

gde je $\Psi(B)$ polinom po B, $\Psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$. S obzirom na to da je

[2]Skraćenica potiče od engleskog imena ove reprezentacije – Moving Average

$E(X_t) = \mu$, što jasno sledi iz činjenice da je $E(\varepsilon_t) = 0$,

$$\begin{aligned} R_k &= E\,(X_t - \mu)\,(X_{t-k} - \mu) = \\ &= E(\varepsilon_t + \psi_1\varepsilon_{t-1} + \cdots + \psi_k\varepsilon_{t-k} + \psi_{k+1}\varepsilon_{t-k-1} + \cdots) \times \\ &\times \ (\varepsilon_{t-k} + \psi_1\varepsilon_{t-k-1} + \cdots) = \\ &= \sigma^2(\psi_k + \psi_1\psi_{k+1} + \psi_2\psi_{k+2} + \cdots) = \sigma^2 \sum_{j=0}^{\infty} \psi_j\psi_{j+k} \end{aligned} \tag{5.15}$$

i

$$\rho_k = \frac{\sum_{j=0}^{\infty} \psi_j\psi_{j+k}}{\sum_{j=0}^{\infty} \psi_j^2}\,.$$

Uslov stacionarnosti je kod ovih modela uvek ispunjen na osnovu uslova *1.* Uoldove teoreme dekompozicije. Zaista, s obzirom na to da je $|\rho_k| \leq 1$,

$$|R_k| \leq |R_0| = [D(X_t)D(X_{t-k})]^{\frac{1}{2}} = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 < \infty\,.$$

Iako model (5.13) ima to svojstvo da konvergira pravoj vrednosti elementa X_t, on ne može da se koristi u praksi jer ima beskonačno mnogo sabiraka, dakle i beskonačno mnogo (nepoznatih) parametara ψ_j, $j = 1, 2, \ldots$, koje bi trebalo oceniti na osnovu posmatranja, odnosno poznatog dela realizacije, i to najčešće samo jedne, vremenskog niza $\{X_t\}$, a takođe i σ^2 ako je nepoznato. Zbog toga se koriste modeli sa konačno mnogo sabiraka, recimo q. Takav model je MA(q), pokretna sredina reda q. Bez smanjenja opštosti govorićemo nadalje samo o pokretnim sredinama sa srednjom vrednošću 0, odnosno za koje je $\mu = 0$.

Neka je $\psi_1 = -\theta_1$, $\psi_2 = -\theta_2$, ..., $\psi_q = -\theta_q$, $\psi_j = 0$ za $j > q$, tada je pokretna sredina reda q oblika

$$X_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}\,, \tag{5.16}$$

gde je $\{\varepsilon_t\}$ beli šum sa očekivanjem nula i disperzijom σ^2. Zaključujemo da je svaki MA(q) proces stacionaran jer je samo konačno mnogo vrednosti ψ različito od nule. Koristeći operator pomeranja unazad, model (5.16) može da

se zapiše u obliku

$$X_t = \theta(B)\varepsilon_t\,,$$

gde je $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ polinom pokretne sredine reda q, polinom po B stepena q.

Primer 5.5. Kao primer ćemo posmatrati pokretnu sredinu reda 1, MA(1). Iz modela (5.16) dobijamo ovakvu pokretnu sredinu stavljajući $q = 1$. Tada dobijamo

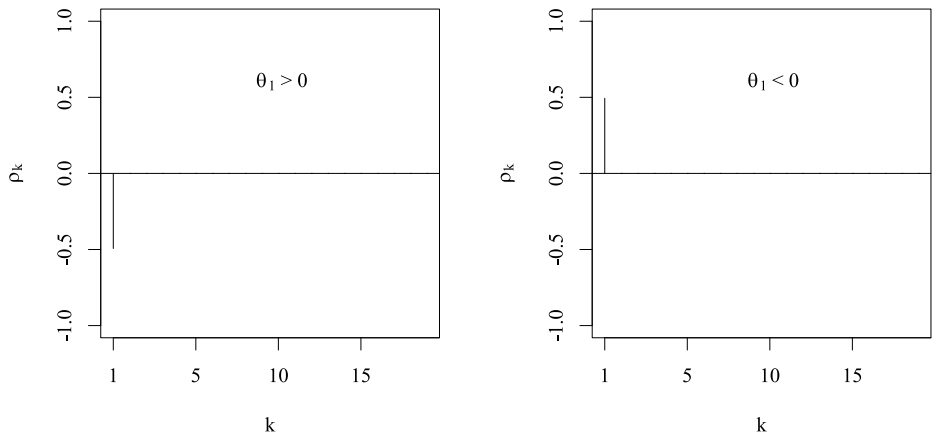
$$X_t = \varepsilon_t - \theta_1\varepsilon_{t-1} = (1 - \theta_1 B)\varepsilon_t\,.$$

Njegova autokovarijansna i autokorelaciona funkcija su, redom,

$$R_k = \begin{cases} \sigma^2(1+\theta_1^2), & k = 0 \\ -\sigma^2\theta_1, & k = 1 \\ 0, & k > 1 \end{cases} \quad \text{i} \quad \rho_k = \begin{cases} 1, & k = 0 \\ \frac{-\theta_1}{1+\theta_1^2}, & k = 1 \\ 0, & k > 1 \end{cases}\,.$$

U skladu sa ovim rezultatom, kaže se da memorija procesa iznosi samo jedan korak.

Korelogram MA(1) procesa je prikazan na slici 5.1.



Slika 5.1: Korelogram MA(1) procesa.

Ukažimo još na dve osobine MA(1) procesa:

a) Za ma koju vrednost θ_1, procesi $X_t = (1 - \theta_1 B)\varepsilon_t$ i $X_t = (1 - 1/\theta_1 B)\varepsilon_t$ će imati istu autokorelacionu funkciju.

b) Za MA(1) proces, vrednosti autokorelacione funkcije su između $-0,5$ i $0,5$, tj. $|\rho_k| < 0,5$ za $k \geq 1$.

Zaista, s obzirom na to da je $\rho_1 = \frac{-\theta_1}{1+\theta_1^2}$, zamenom θ_1 sa $\frac{1}{\theta_1}$, dobijamo isti izraz. Iz istog izraza zaključujemo i da je $|\rho_1| < 0,5$ jer θ_1 mora da bude realno. S obzirom na to da je $\rho_k = 0$ za $k > 1$, to izvodimo i zaključak pod b).

Primetimo da zbog osobine pod a) nećemo biti u mogućnosti da na jednoznačan način identifikujemo MA(1) niz samo na osnovu autokorelacione funkcije.△

U praktičnim primenamma imamo posla sa *uzoračkom autokorelacionom funkcijom*, koja se na osnovu uzorka $(X_1, \ldots, X_n)$, dela jedne realizacije, izračunava kao

$$\widehat{\rho}_k = \frac{\sum_{j=1}^{n-k}(X_j - \overline{X})(X_{j+k} - \overline{X})}{\sum_{j=1}^{n}(X_j - \overline{X})^2}, \quad k = 0, 1, 2, \ldots, \tag{5.17}$$

gde je $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$.

Primer 5.6. Za MA(1) model definisan sa $X_t = \varepsilon_t - 0,5\varepsilon_{t-1}$, gde je $\{\varepsilon_t\}$ Gausov $\mathcal{N}(0,1)$ beli šum, na osnovu 250 simulacija odrediti uzoračku autokorelacionu funkciju.

Rezultati do koraka 10 su prikazani tabelom

k	1	2	3	4	5	6	7	8	9	10
$\widehat{\rho}_k$	-0,45	0,06	-0,04	0,01	-0,06	0,10	-0,08	-0,01	-0,04	0,07

,

što svakako odstupa od teorijskih vrednosti, ali je vidljivo da je samo za $k = 1$ odstupanje od nule značajno. Dakle, uočava se karakteristika MA(1).

Funkcija `arima.sim` programskog jezika `R` računa simuliranu MA seriju. Implemetirali smo funkciju `acfOfsimARMA` koja računa autokorelacione koeficijente za simuliranu ARMA seriju.

```
acfOfsimARMA<-function(arCoefficients, maCoefficients, numberOfSim){
   testMA1 <- arima.sim(n=numberOfSim,
```

```
      list(ar=arCoefficients,ma=maCoefficients))
   acf.obj<-acf(testMA1)
   return(acf.obj[1:10])
}
```

Kako nama treba MA(1) model, parametar `arCoefficients` biće prazan niz, a parametar `maCoefficients` niz sa jednim elementom. Pa bi poziv funkcije za ovaj primer bio

```
> acfOfsimARMA(c(),c(-0.5), 250)
```

Svaki put kad pokrenemo ovu funkciju dobićemo drugačiji rezultat, jer se ovde radi o simuliranim vrednostima. △

Model autoregresije

Za modeliranje vrednosti vremenskog niza koristi se i regresija na prethodne vrednosti istog vremenskog niza plus *slučajni šok*, označimo ga sa ε. Dakle,

$$X_t = \pi_1 X_{t-1} + \pi_2 X_{t-2} + \cdots + \varepsilon_t \tag{5.18}$$

ili koristeći operator pomeranja unazad,

$$\pi(B) X_t = \varepsilon_t \,,$$

gde je $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$ autoregresivni polinom za koga je

$$1 + \sum_{j=1}^{\infty} |\pi_j| < \infty \,. \tag{5.19}$$

Ako se vremenski niz može da prikaže u obliku (5.18) sa ispunjenim uslovom (5.19), kaže se da je on *invertibilan*.

Primetimo da i kod ovog modela X_s ne zavisi od ε_t kada je $s < t$.

Niz $\{\varepsilon_t\}$, u slučaju modela autoregresije se naziva *inovacioni niz*, s obzirom na to da se njegovim elementima obuhvataju novine procesa $\{X_t\}$ u odgovarajućem trenutku koje nisu objašnjive prošlim vrednostima niza $\{X_t\}$. Kao inovacioni niz u modelu se koristi niz čiji su elementi nekorelirani među sobom i svi imaju istu raspodelu sa disperzijom σ^2. Što se tiče srednje vrednosti elemenata niza $\{\varepsilon_t\}$, ona ne mora da bude 0, ali je u svakom slučaju ista za

sve elemente niza.

I u modelu (5.18) imamo beskonačno mnogo sabiraka, pa samim tim i beskonačno mnogo koeficijenata koje treba oceniti. To je, naravno, nemoguće u praktičnim primenama, pa se definiše model autoregresije konačnog reda. Stavljajući da je $\pi_1 = \phi_1$, $\pi_2 = \phi_2$,..., $\pi_p = \phi_p$ i $\pi_j = 0$ za $j > p$, dobijamo autoregresiju reda p, AR(p).

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \tag{5.20}$$

ili korišćenjem operatora B,

$$\phi(B) X_t = \varepsilon_t \,,$$

gde je $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 + \cdots + \phi_p B^p$ autoregresivni polinom po B.

Znači da je tekuća vrednost X_t vremenskog niza $\{X_t\}$ linearna kombinacija p neposredno prošlih vrednosti plus slučajni poremećaj ε_t.

S obzirom na to da je $\sum_{j=1}^{\infty} |\pi_j| = \sum_{j=1}^{p} |\phi_j| < \infty$, AR($p$) proces je uvek invertibilan.

Primer 5.7. Kao primer ćemo posmatrati autoregresiju reda 1, AR(1). Iz modela (5.20) dobijamo ovakav autoregresivni niz stavljajući $p = 1$. Pretpostavićemo bez smanjenja opštosti da je $E(X_t) = 0$ (za svako t). Dakle,

$$X_t = \phi_1 X_{t-1} + \varepsilon_t \,, \tag{5.21}$$

odnosno korišćenjem operatora B

$$(1 - \phi_1 B) X_t = \varepsilon_t \,.$$

S obzirom na to da je

$$X_t = (1 - \phi_1 B)^{-1} \varepsilon_t = (1 + \phi_1 B + \phi_1^2 B^2 + \cdots)\varepsilon_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \cdots \,,$$

ako je $E(\varepsilon_t) = 0$, onda je i $E(X_t) = 0$. Nadalje pretpostavimo da je to ispunjeno. Što se tiče disperzije

$$D(X_t) = \sigma^2 (1 + \phi_1 + \phi_1^2 + \phi_1^4 + \cdots) \,,$$

pa zaključujemo da će proces $\{X_t\}$ imati disperziju ako konvergira red $\sum_{j=0}^{\infty} |\phi_1|^j$, a on konvergira za $|\phi_1| < 1$. To je istovremeno i uslov stacionarnosti procesa AR(1).

Da bismo odredili autokovarijansnu funkciju procesa AR(1), pomnožimo i levu i desnu stranu jednakosti (5.21) sa X_{t-k}, $k \geq 0$ i nađimo očekivanje:

$$E(X_t X_{t-k}) = \phi_1 E(X_{t-1} X_{t-k}) + E(\varepsilon_t X_{t-k}) \,,$$

odakle sledi

$$R_k - \phi_1 R_{k-1} = E(\varepsilon_t X_{t-k}) \,.$$

Vrednost

$$E(\varepsilon_t X_{t-k}) = \begin{cases} \sigma^2, & k = 0 \\ 0, & k > 0 \end{cases} \,,$$

pa se dobija sistem jednačina poznat pod nazivom Jul–Uokerove[3] jednačine:

$$R_0 - \phi_1 R_{-1} = \sigma^2 = R_0 - \phi_1 R_1 \tag{5.22}$$

$$R_k - \phi_1 R_{k-1} = 0 \,, \; k = 1, 2, \ldots \tag{5.23}$$

Zamenom $R_1 = \phi_1 R_0$ u (5.23), dobijamo

$$R_0 = \frac{\sigma^2}{1 - \phi_1^2} \,,$$

a deljenjem (5.23) sa R_0, dobijamo autokorelacionu funkciju procesa AR(1):

$$\rho_k = \phi_1 \rho_{k-1} \,. \tag{5.24}$$

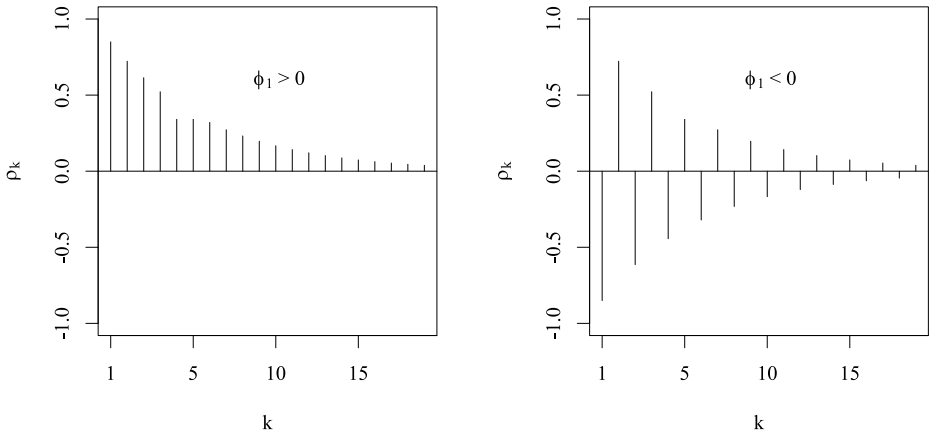
Rešavanjem homogene diferencne jednačine (5.24), dobijamo

$$\rho_k = \phi_1^k \rho_0 = \phi_1^k \,, \quad k = 0, 1, 2, \ldots \,.$$

Ako je $\phi_1 > 0$, vrednosti autokorelacione funkcije opadaju eksponencijalnom brzinom ka nuli. Ako je $\phi_1 < 0$, vrednosti autokorelacione funkcije alterniraju,

[3] Yule–Walker

opadajući po apsolutnoj vrednosti ka nuli (slika 5.2),



Slika 5.2: Korelogram AR(1) procesa.

odnosno predstavljaju kombinaciju dve opadajuće eksponencijalne krive. U oba slučaja opadanje je sporije što je parametar autoregresije ϕ_1 bliži granicama nestacionarnosti, vrednostima 1, odnosno -1.$\triangle$

Uzoračku autokorelacionu funkciju ponovo predstavlja statistika (5.17) na osnovu dela realizacije niza $(X_1, \ldots, X_n)$.

Primer 5.8. Za AR(1) proces $X_t = 1 + 0,9X_{t-1} + \varepsilon_t$ gde je $\{\varepsilon_t\}$ Gausov $\mathcal{N}(0,1)$ beli šum, na osnovu 250 simulacija odrediti uzoračku autokorelacionu funkciju.

Rezultati do koraka 10 su prikazani tabelom

k	1	2	3	4	5	6	7	8	9	10
$\widehat{\rho}_k$	0,87	0,78	0,69	0,57	0,48	0,39	0,31	0,24	0,18	0,16

.

Vidljivo je opadanje uzoračke autokorelacione funkcije, što je i bilo očekivano.

U Primeru 5.6 definisali smo funkciju `acfOfsimARMA` koju ćemo i ovde iskoristiti, a pozvaćemo je na sledeći način

```
> acfOfsimARMA(c(0.9),c(), 250)
```

Vrednosti koje funkcija daje razlikuju se prilikom svakog poziva jer je reč o simuliranju nizova čije autokorelacione koeficijente računamo. △

Dodatak

Osnove programskog jezika R

`R` je besplatan programski jezik razvijen pretežno za upotrebu u statistici i naukama koje se bave velikim skupovima podataka (data science, machine learning i slične). Za lakši rad u programskom jeziku `R` potrebno je imati i neko razvojno okruzenje, kako na primer RStudio. RStudio je takođe besplatan softver koji nam omogućava bolji pregled koda, lakše kretanje kroz debug-er, a takođe sadrži i alate za preuzimanje podataka iz fajlova kreiranih u prgramima Excel, SAS, SPSS i Stata. Razvojna okruženja koja mogu biti alternativa za RStudio su Visual Studio for R, Eclipse, TINN-R i drugi.

Programski jezik `R` izvršava funkcije pozivom iz komandne linije. U RStudio-u taj poziv se izvršava iz konzole. U konzolu možemo uneti funkciju po želji, a pritiskom na *Enter* dobićemo rezultat

```
> 1+1
[1] 2
```

Suština programskog jezika `R` je da se iskoriste njegove složenije funkcije za čiju implementaciju u drugim programskim jezicima bi nam trebalo dosta vremena. Na primer, devedeset peto procentni kvantil standardne normalne raspodele možemo dobiti kao

```
> qnorm(0.95)
[1] 1.644854
```

dok bi za ovaj rezultat trebalo dosta kodiranja u jezicima kao što su C++, C#, Java i slični. Ako neka funkcija nije dostupna u programskom jeziku `R`, možemo je sami napisati.

Radno okruženje (workspace)

Radno okruženje predstavlja skup objekata koje kreiramo u toku rada. Dakle, sve javne promenljive i funkcije pamte se u radnom okruženju. Ako u komandnoj liniji izvršimo kod

```
> x<-5
```

promenljiva `x`, čija je vrednost 5, pojaviće se u radnom okruženju i možemo je nadalje koristiti. Jedna napomena, u programskom jeziku `R` za operator dodele koristi se "`<-`" ili "`=`". Postoje male razlike između ova dva operatora, ali se sada time nećemo baviti.

Na slici 5.3 dat je izgled radnog okruženja u programu RStudio. Radno okruženje je podeljeno u četiri glavne celine. U gornjem levom polju imamo editor gde u .R fajlove možemo pisati naš kod. Od standardnog text editora (kao što je Notepad++) razlikuje ga to što podržava komandu autocomplete prilikom kucanja teksta. Auto-complete dobijamo pritiskom tastera *ctrl+space*.

Kod koji pokrenemo komandom Code → Run Selected Line(s) (*ctrl+enter*) biva zapamćen u vidu promenljivih ili funkcija u gornjem desnom polju, Global Environment. Tu možemo uočiti tri celine. U prvoj se pamte objekti kao što su model linearne regresije, ili učitana tabela iz .csv fajla. Druga celina, Values, sadrži promenljive, nizove, matrice i slično. Treća celina, Functions, sadrži listu svih funkcija koje smo implementirali u .R fajlovima. Ako hoćemo da izbrišemo neki objekat iz radnog okruženja upotrebićemo komandu `rm(x)` (koja briše promenljivu `x`).

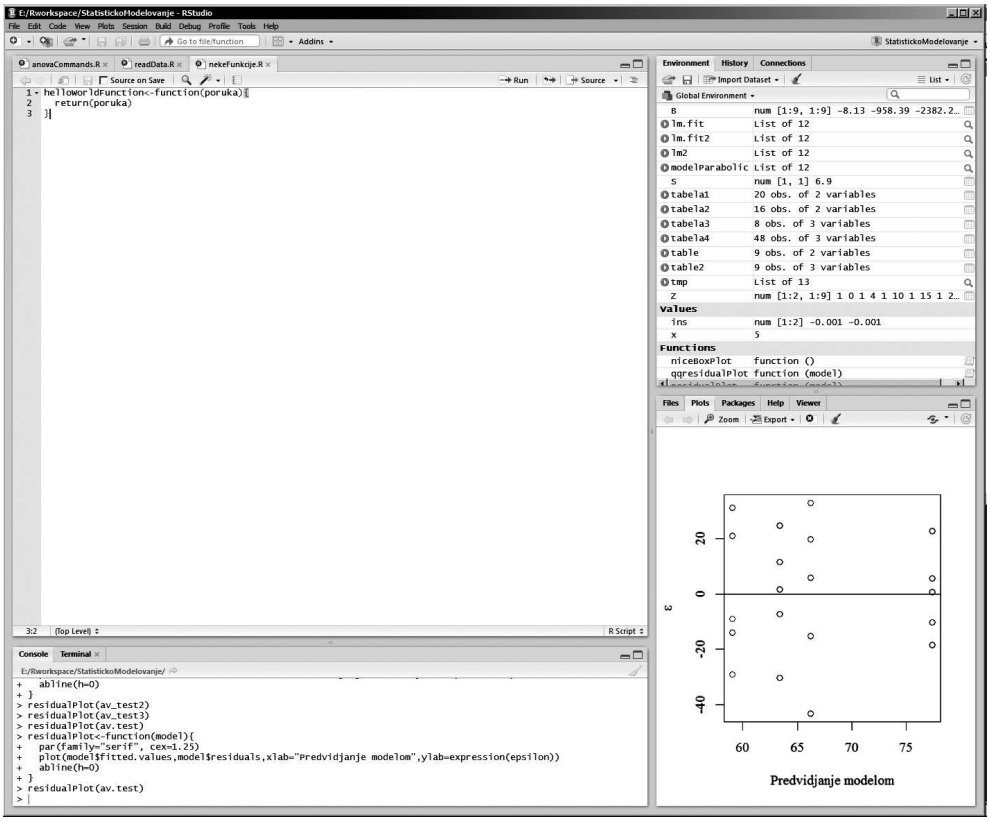
Konzola za izvršavanje koda nalazi se u donjem desnom polju. Funkciju ili bilo koji kod koji želimo da izvršimo pokrećemo pritiskom na taster *Enter*. Takođe, konzola sadrži spisak funkcija koje smo prethodno pozvali. Prilikom debagiranja koda u konzoli nam se pojavljuje korisnički interfejs koji omogućava lakše kretanje kroz debug-er.

Polje u donjem desnom uglu sadrži pet kartica. Prva kartica, Files, sadrži spisak svih fajlova koji su deo radnog okruženja. Kartica Plots, sadži listu svih grafika koje smo generisali u toku rada. Kartica Packages nam omogućava lakšu instalaciju biblioteka (paketa) koje želimo da koristimo u toku rada. Zatim imamo karticu Help gde možemo naći opis željene funkcije i njenih ulaznih parametara. Na kraju imamo i karticu Viewer koja nam prikazuje izgled web stranice koju kodiramo.

Spisak komandi koje smo izvršili iz komandne linije u toku rada možemo zapamtiti u radnom okruženju pozivom funkcije `savehistory`, u suprotnom taj niz komandi gubi se nakon gašenja RStudija. Kada ponovo pokrenemo RStudio spisak izvršenih komandi povratićemo pozivom funkcije `loadhistory`.

Definisanje promenljivih i rad sa njima

Promenljive u progremskom jeziku `R` se ne difinišu preko tipa (dakle, promenljiva `x` nije definisana kao integer ili string ili nešto treće). Ako imamo



Slika 5.3: Radno okruženje u programu RStudio

promenljivu `x` i pozovemo `x<-5`, `x` ima vrednost 5. Dalje, ako pozovemo `x<-"PMF"`, `x` je sada string sa dodeljenom vrednošću PMF. Dakle, svaka promenljiva može se menjati po potrebi.

Vektorska promenljiva definiše se na sledeći način

```
x<-c(1,2,3,9,100)
```

dok bi vektor sa celim brojevima od 3 do 9 (uključujući 3 i 9), redom, definisali sa

```
x<-c(3:9)
```

Slično je i za vektor koji nije numeričkog tipa

```
x<-c("Toyota","Mazda","Honda")
```

Prilikom definisanja ovog vektora, svejedno je da li koristimo karakter " ili '.

Vektor sa logičkim veličinama TRUE i FALSE definišemo pozivom

```
x<-c(T,F,T)
```

Matrice definišemo pozivom funkcije `matrix`, pa bi tako jedna matrica bila

```
x<-matrix(c(1, 2, 3, 10, 20, 30),nrow=2, ncol=3, byrow = TRUE)
        [,1]  [,2]  [,3]
  [1,]    1     2     3
  [2,]   10    20    30
```

Elementi ove matrice upisuju se po vrsti. Da li će se elementi upisivati po vrsti ili koloni određuje parametar "byrow", čija je pretpostavljena vrednost FALSE.

Kolonama i vrstama možemo da dodelimo imena pozivom funkcija `colnames` odnosno `rownames`. To nam omogućava da koloni ili vrsti ne pamtimo redni broj već naziv, pa da preko tog naziva pristupamo elementima.

```
> rownames(x)<-c("Jednocifreni", "Dvocifreni")
> x
               [,1]  [,2]  [,3]
 Jednocifreni    1     2     3
 Dvocifreni     10    20    30
```

Velika prednost programskog jezika `R` je rad sa višedimenzionalnim promenljivama. Recimo da imamo dve kvadratne matrice jednakih dimenzija `m1` i `m2`, njihov zbir bi našli jednostavnom pozivom `m1+m2` (slično i za razliku). Proizvod matrica dobili bismo sa `m1 %*% m2`, dok bi poziv `m1 * m2` računao proizvod između elemenata na istim pozicijama. Funkcija `mean(v1)` izračunala bi srednju vrednost svih elemenata matrice. Ovo je samo mali deo onoga što programski jezik `R` nudi u radu sa vektorima i matricama.

Posebno zanimljive strukture za čuvanje vrednosti jesu okviri tj. data frames. To je lista vektora iste dužine koji imaju jedinstvene nazive kolona, pri čemu emelenti na istim pozicijama u tim vektorima određuju osobine neke jedinke. Na primer, imamo promenljive `marka<-c("Toyota", "Mazda", "Volkswagen")` i `zemlja<-c("Japan", "Japan", "Nemacka")`. Kreiraćemo jedan okvir sa ove dve liste

```
autoIndustrija <- data.frame(marka, zemlja)
```

Odgovarajućim kolonama iz okvira pristupamo pomoću simbola `$`, pa imamo

```
> autoIndustrija$marka
```

```
[1] Toyota  Mazda  Volkswagen
```

Indeksiranje

Elementima vektora, matrica ili okvira može se pristupiti preko indeksa. Recimo iz malopre definisanog okvira podataka `autoIndustrija` svakom elementu možemo pristupiti sa, na primer

```
> autoIndustrija[1,2]
[1] Japan
```

U matrici `x`, koju smo kreirali u prethodnom poglavlju, elementima prve vrste pristupamo sa

```
> x[1,]
[1]  1  2  3
```

ili preko naziva kolone

```
> x["Jednocifreni",]
[1]  1  2  3
```

Moguće je i uslovno indeksiranje. Recimo, hoćemo da izdvojimo sve elemente matrice `x` koji su veći od 2

```
> x[x>2]
[1]  4  5  3  6
```

ili elemente od prvog do trećeg (gledajući po koloni)

```
> x[1:3]
[1]  1  4  2
```

Ako su nam potrebni svi elementi matice `x` sem trećeg i četvrtog (gledajući po koloni) pozvali bismo komandu

```
> x[-c(3,4)]
[1]  1  4  3  6
```

Analogno, bez znaka minus dobili bismo elemente na tačno željenim pozicijama. Kao i kod okvira, i kod matrica pojedinačno pristupamo elementima navodeći pozicije elemenata.

Učitavenje podataka iz fajla

Ono što programski jezik `R` izdvaja od ostalih je mogućnost rada sa velikim skupovima podataka. Ovde neće biti reči o razmeni podataka između `R`–a i baze, kako to i nije predmet izučavajnja ove knjige, već ćemo se fokusirati na učitavanje podataka iz .txt i .csv fajlova.

Funkcijom `read.table` učitavamo podatke iz tekstualnih fajlova. Da bi ovo učitavanje bilo uspešno potrebno je da fajl ispunjava određene kriterijume. Podaci u fajlu moraju biti uneti u ASCII formatu. Ako drugačije nije definisano, podaci su podeljeni praznim karakterom (space). Poželjno je da prva linije u fajlu sadrži imena kolona. Ako nekoj koloni nedostaje vrednost, biće upisan karakter `NA`, što se može desiti ako nam nedostaje neko merenje ili smo greškom dva puta pritisnuli *space bar*.

Pored `read.table` koriste se i funkcije `read.csv` i `read.csv2`. Kod funkcije `read.csv` podrazumeva se da su elementi razdvojeni zarezom, a kod `read.csv2` tačka-zarezom (tj. simbolom ;). Dakle, pozivom komande

```
> ohlc <- read.csv("C:/Users/Admin/Downloads/IBM.csv",header=T)
> ohlc
          Date      Open    High     Low    Close   Adj.Close   Volume
   1   2018-08-30  147.03  147.30  145.25  145.93    145.93    3340400
   2   2018-08-31  145.72  146.78  145.54  146.48    146.48    3488500
   3   2018-09-04  145.98  146.19  144.81  145.68    145.68    3326200
   4   2018-09-05  145.19  146.75  145.05  146.66    146.66    3126500
   5   2018-09-06  146.88  147.66  145.54  146.39    146.39    4248800
  ...
```

odbićemo okvir (data frame) koji sadrži cene akcija kompanije IBM. Kako su podaci dati u fajlu koji sadrži zaglavlje i mi smo pravilno učitali to zaglavlje, možemo cenama na zatvaranju berze da pristupimo pozivom

```
ohlc$Close
```

Fajlovi se još mogu učitati pomoći funkcija `read.delim` i `read.delim2` gde sami zadajemo po kom karakteru su podeljeni elementi, a podrazumevani karakter je *tab*.

Definisanje funkcija

Funkcije, a i promenljive, definišemo u .R script fajlovima. Napisani kod treba pokrenuti (*run code*) kako bi se on zapamtio u vidu novih funkcija ili promenljivih u našem radnom okruženju. Na primer, funkcija koja štampa poruku glasi

```
helloWorldFunction<-function(poruka){
    return(poruka)
}
```

Poziv i odgovor ove funkcije bio bi

```
> helloWorldFunction("Hello world!")
[1] "Hello world!"
```

Treba voditi računa da naziv funkcije bude jedinstven. U suprotnom, u radnom okruženju biće zapamćena samo ona funkcija koja je poslednja pokrenuta (nad kojom je izvršena komanda *run code*).

Crtanje grafika

Postoji više paketa za crtanje grafika u programskom jeziku `R`, ali mi ćemo se baviti samo funkcijama iz osnovnog paketa. Pođimo od funkcije `plot` i uzmimo objekat `ohlc` koji smo kreirali u prethodnoj diskusiji. Grafik cena na zatvaranju berze dobićemo pozivom `plot(ohlc$Close)`, a grafik je predstavljen na slici 5.4. Da bismo dobili grafik sa slike 5.5, potrebno je pozvati čitav niz komandi.

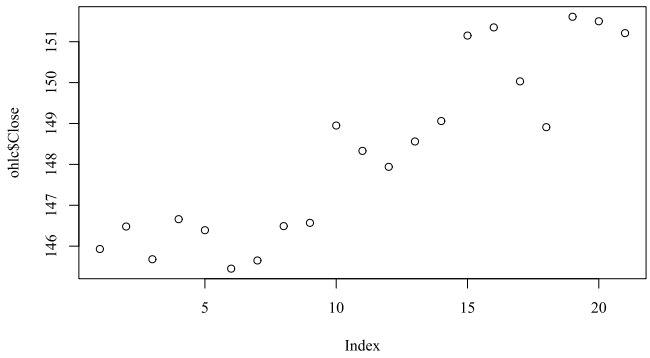
Najpre definišimo granice za vrednosti na y-osi, kao i gornju granicu na x-osi i korak sa kojim ćemo ispisivati vrednosti na x-osi

```
> ymax= trunc(max(ohlc$Close))+1
> ymin= trunc(min(ohlc$Close))
> xmax = length(ohlc$Close)
> stepLabel = 1
```

Grafik iscrtavamo deo po deo. Za početak podesimo font na Times New Roman i veličinu oznaka komandom

```
> par(family="serif", cex=0.85)
```

Zatim iscrtavamo samo vrednosti serije bez koordinatnih osa



Slika 5.4: Grafik cene akcija kompanije IBM.

```
> plot(ohlc$Close, axes=F, ann=F, type="o", col=boja,
    ylim=c(ymin,ymax), lwd=2)
```

gde smo izabrali tip i boju grafika, a uneli smo i granice za y-osu. Na y-osi ne želimo nikakve izmene pa možemo odmah da je nacrtamo komandom

```
> axis(2)
```

Kod x-ose hoćemo da odredimo gustinu podeoka, debljinu ose, debljinu podeonih linija, ali još nećemo da ispišemo vrednosti.

```
> axis(1, at=seq(1 , xmax, by=stepLabel),lwd=1, lwd.ticks = 2,
    labels = F)
```

Vrednosti na x-osi su datumi, pa ih čitamo iz kolone `ohlc$Date`. Određujemo im položaj i rotiramo za 60 stepeni komandom

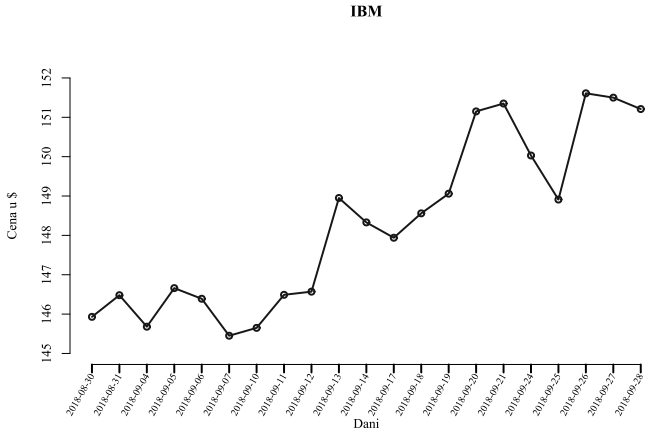
```
> text(seq(1,xmax,by=stepLabel), par("usr")[3]-0.25, srt = 60,
    adj= 1, xpd = TRUE, labels = ohlc$Date[seq(1, xmax, stepLabel)],
    cex=0.75)
```

Na kraju nam ostaje da ispišemo imena koordinatnih osa i grafika.

```
title(line=3.35, ylab="Cena u $", xlab="Dani", main = "IBM")
```

Ovaj niz komandi treba smestiti unutar jedne funkcije kako bismo olakšali iscrtavanje grafika i omogućili primenu istog dizajna na neku drugu seriju podataka.

Pored funkcije `plot` često se koristi i funkcija `hist` koja iscrtava histogram.



Slika 5.5: Grafik cene akcija kompanije IBM.

Podešavanje opcija se radi slično kao i kod prethodne funkcije pa se nećemo detaljnije baviti funkcijom `hist`.

Paketi

Prilikom instaliranja programskog jezika `R` učitava se skup paketa koji sadrži osnovne funkcije. Skup tih osnovnih funkcija je poprilično velik, tako da u ovoj knjizi nećemo koristiti funkcije iz nekih dodatnih paketa. Ukoliko bismo želeli da uključimo dodatni paket to bismo uradili pozivom funkcije `install.packages`, a zatim `library`. Recimo da nam treba funkcija `qplot` za crtanje složenih dijagrama. Ta funkcija se nalazi u paketu `ggplot2` pa bismo iz komandne linije pozvali

```
> install.packages("ggplot2")
> library(ggplot2)
```

kako bismo mogli da pristupimo željenoj funkciji. Nakon ponovnog pokretanja RStudija moraćemo ponovo da učitamo željenu biblioteku, bez prethodnog poziva za njenu instalaciju. Ako u toku rada ne želimo više da radimo sa nekim paketom, paket možemo isključiti pozivom

```
> detach("package:ggplot2")
```

AIC i BIC kriterijum

Kada se koristi neki statistički model da bi se njime predstavili (prikupljeni) podaci, model uglavnom nikada neće tačno odslikavati podatke. Bilo koji model da izaberemo, izgubićemo deo informacija o posmatranom obeležju a koje su sadržane u uzorku. Postoji više kriterijuma kojima se u statistici vrši izbor najboljeg modela među njih konačno mnogo. Dakle, kriterijumi koje ćemo ovde definisati nisu testovi kojima se može da testira hipoteza o "dobrom" prilagođavanju[4] modela, već se, ponovo naglašavamo, njima vrši samo izbor najboljeg među konačno mnogo posmatranih modela za dati skup podataka (dati realizovani uzorak).

Akaikeov[5] informacioni kriterijum (AIC) je ocena relativnog kvaliteta izabranog statističkog modela u odnosu na druge posmatrane modele. Bazira se na teoriji informacija odakle mu i naziv. On sada već predstavlja bazu statističkog zaključivanja.

Neka je dat uzorak $\mathbf{X} = (X_1, \ldots, X_n)$ i model M kojim modeliramo podatke dobijene iz uzorka $\mathbf{X}$. Označimo sa $L(\boldsymbol{\theta}; M)$ funkciju verodostojnosti $L(\boldsymbol{\theta}; M) = p(\mathbf{x}|\boldsymbol{\theta}, M)$ koja odgovara modelu M, a $\boldsymbol{\theta}$ parametar odgovarajuće dimenzije modela M. Neka je $\widehat{\boldsymbol{\theta}}$ vrednost parametra koja maksimalizuje vrednost funkcije verodostojnostiu modela M i neka je $\widehat{L}$ taj maksimum. Neka je k dimenzija parametra $\boldsymbol{\theta}$, odnosno broj ocenjenih parametra modela M. Tada

$$AIC = 2k - 2\ln(\widehat{L})$$

daje vrednost na osnovu koje biramo model. Izabraćemo onaj među posmatranim modelima za koji vrednost AIC bude najmanja.

U bliskoj vezi sa AIC kriterijumom je Bajesov informacioni kriterijum (BIC). On se još naziva i Švarcov[6] kriterijum (SBC ili SBIC). On glasi

$$BIC = \ln(n)k - 2\ln(\widehat{L})\,,$$

[4] goodness–of–fit

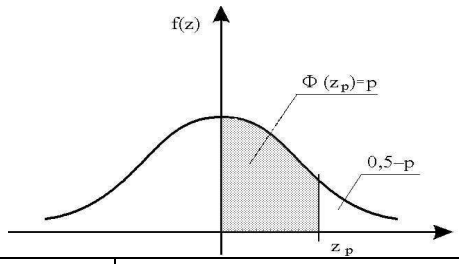
[5] Hirotugu Akaike, naučnik koji je formulisao navedeni kriterijum i najavio ga 1971. godine, a objavio njegovu preciznu definiciju u svom radu objavljenom 1974. godine.

[6] Gideon E. Schwarz, naučnik koji je formulisao navedeni kriterijum i objavio ga u svom radu 1978.godine.

gde oznake imaju isto značenje kao i kod AIC, a n je obim uzorka. Opet biramo onaj model među razmatranim koji ima najmanju vrednost BIC.

Statističke tablice

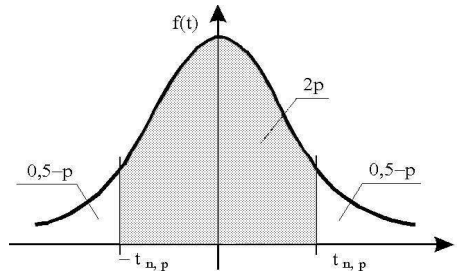
1. Normalna raspodela



$\Phi(z_p) = p$

z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

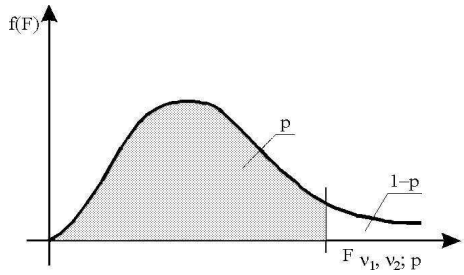
2. Studentova raspodela



$$P\{|t_n| < t_{n,p}\} = 2p$$

$n \setminus p$	0.100	0.200	0.300	0.400	0.450	0.475	0.490	0.495
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921
17	.257	.534	.863	1.133	1.740	2.110	2.567	2.898
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807
24	.256	.531	.857	1.318	1.711	2.064	2.492	2.797
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763
29	.256	.530	.854	1.311	1.699	2.045	2.045	2.462
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576

3. Fišerova raspodela



$P\{F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2;\ p}\} = p$

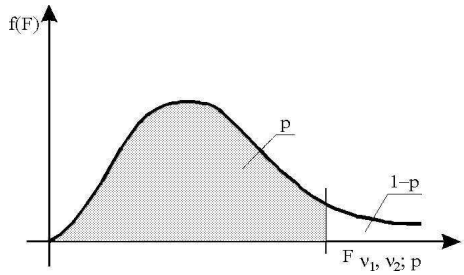
3. a) $p = 0,990$

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10
1	4052	5000	5403	5625	5764	5859	5928	5981	6023	6056
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.28	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

nastavak 3. a)

12	15	20	24	30	40	60	120	ν_1/ν_2
6106	6157	6209	6235	6261	6287	6313	6339	1
99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	2
27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	3
14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	4
9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	5
7.72	7.56	7.40	4.31	7.23	7.14	7.06	6.97	6
6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	7
5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	8
5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	9
4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	10
4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	11
4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	12
3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	13
3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	14
3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	15
3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	16
3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	17
3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	18
3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	19
3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	20
3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	21
3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	22
3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	23
3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	24
2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	25
2.96	2.82	2.66	2.58	2.50	2.42	2.33	2.23	26
2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	27
2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	28
2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	29
2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	30
2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	40
2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	60
2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	120
2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	∞

Fišerova raspodela



$P\{F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2;\, p}\} = p$

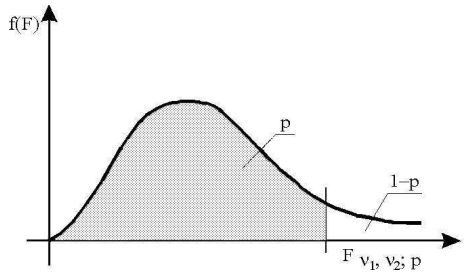
3. b) $p = 0,975$

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10
1	648	799	864	900	922	937	948	957	963	969
2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4
3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.5
4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05

nastavak 3. b)

12	15	20	24	30	40	60	120	ν_1/ν_2
977	985	993	997	1001	1006	1010	1014	1
39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5	2
14.3	14.3	14.2	14.1	14.1	14.0	14.0	13.9	3
8.75	8.65	8.56	8.51	8.46	8.41	8.36	8.31	4
6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	5
5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	6
4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	7
4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.71	8
3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	9
3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	10
3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	11
3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	12
3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	13
3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	14
2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	15
2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	16
2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	17
2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	18
2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	19
2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	20
2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	21
2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	22
2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	23
2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	24
2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	25
2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	26
2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	27
2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	28
2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	29
2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	30
2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	40
2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	60
2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	120
1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	∞

Fišerova raspodela



$P\{F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2;\ p}\} = p$

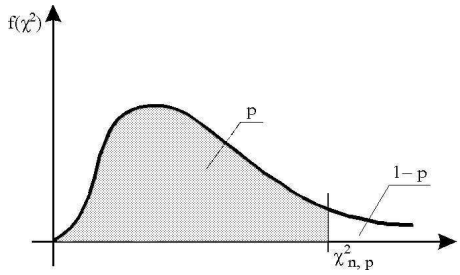
3. c) $p = 0,950$

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.48	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.87	1.83

nastavak 3. c)

12	15	20	24	30	40	60	120	ν_1/ν_2
244	246	248	249	250	251	252	253	1
19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	2
8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	3
5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	4
4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	5
4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	6
3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	7
3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	8
3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	9
2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	10
2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	11
2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	12
2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	13
2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	14
2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	15
2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	16
2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	17
2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	18
2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	19
2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	20
2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	21
2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	22
2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	23
2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	24
2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	25
2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	26
2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	27
2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	28
2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	29
2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	30
2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	40
1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	60
1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	120
1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	∞

4. χ^2 raspodela



$P\{\chi^2_n < \chi^2_{n;\ p}\} = p$

$n \setminus p$	0.005	0.010	0.025	0.050	0.95	0.975	0.990	0.995
1	.0000	.0002	.0010	.0039	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.103	5.99	7.38	9.21	10.6
3	.0717	.115	.216	.352	7.81	9.35	11.3	12.8
4	.207	.297	.484	.711	9.49	11.1	13.3	14.9
5	.412	.554	.831	1.15	11.1	12.8	15.1	16.7
6	.676	.872	1.24	1.64	12.6	14.4	16.8	18.5
7	.989	1.24	1.69	2.17	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	28.9	37.5	34.8	37.2
19	6.84	7.63	8.91	10.1	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	79.1	83.3	88.4	92.0
70	43.3	45.4	48.8	51.7	90.5	95.0	100	104
80	51.2	53.5	57.2	60.4	102	107	112	116
90	59.2	61.8	65.6	69.1	113	118	124	128
100	67.3	70.1	74.2	77.9	124	130	136	140

Literatura

1. Anděl J.: **Matematická statistika**, SNTZ/Alfa, Praha, 1985

2. Anderson T.: **The statistical analysis of time series**, John Wiley & Sons, New York, 1971

3. Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M: **Time series analysis, forecasting and control**, Wiley, Hoboken, 2016 (fifth edition)

4. Brockwell P., Davis R.: **Time series: Theory and methods**, Springer--Verlag, New York, 1987

5. Brownlee K.: **Statistical theory and methodology in science and engineering**, John Wiley & Sons, New York, 1977

6. Cochran W.: **Sampling techniques**, John Wiley & Sons, New York, 1963

7. DeGroot M.: **Optimal statistical decisions**, McGrow-Hill Co., New York, 1970

8. Faraway J. J.:**Linear models with R**, CRC Press, Taylor & Francis Group, 2015 (second edition)

9. Fuller W. A.: **Itroduction to statistical time series**, John Wiley & Sons, New York, 1976

10. Hair J. F., Black W. C., Babin B. J., Anderson R. E., Tatham R. L.: **Multivariate data analysis**, Pearson, New Jersey, 2006 (sixth edition)

11. Hadžić O.: **Numeričke i statističke metode u obradi eksperimentalnih podataka**, Univerzitet u Novom Sadu, Institut za matematiku, Novi Sad, 1992

12. Hogg R., McKean J., Craig A.:**Introduction to mathematical statistics**, Pearson, 2005

13. Ivčenko G. I., Medvedev J. I.: **Matematičeskaja statistika**, Višaja škola, Moskva, 1984

14. Kendall M. G., Steward A.: **The Advanced Theory of Statistics**, vol. 3: Design and Analysis, and Time Series, Griffin company, London, 1968

15. Larsen R. J., Marx M. L.: **An introduction to mathematical statistics and its applications**, Pearson, Boston, 2012, fifth edition

16. Mališić J.: **Vremenske serije**, Matematički fakultet, Beograd, 2002

17. Oehlert G. W.: **A first course in design and analysis of experiments**, Gary W. Oehlert, University of Minnesota, 2010

18. Popović B.: **Matematička statistika i statističko modelovanje**, Prirodno–matematički fakultet, Niš, 2003

19. Popović B., Blagojević B.: **Matematička statistika sa primenama u hidrotehnici**, Univerzitet u Nišu, Niš, 2003 (treće izdanje)

20. Rao S.R.: **Linejnije statističeskie metodi i ih primenenija**, Nauka, Moskva, 1968 (prevod sa engleskog)

21. Roussas G. G.: **A course in mathematical statistics**, Academic Press, San Diego, 1997 (second edition)

22. Rubin, D. B.: **Iteratively reweighted least squares**, In Encyclopedia of Statistical Sciences, Volume 4, Wiley, 1983

23. Thompson S.: **Sampling**, John Wiley & Sons, New York, 1992

24. Wackerly D., Mendenhall W., Scheaffer R.: **Mathematical statistics with applications**, Duxbury Press, Belmont, 1996 (fifth edition)

25. Wei W. W. S.:**Time series analysis, univariate and multivariate methods**, Pearson, 2006 (second edition)

Indeks pojmova