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

Suzana Vidić

SEGMENTIRANA REGRESIJA SA PRIMENOM

- master rad -

Mentor:
prof. dr Zorana Lužanin

Novi Sad, 2014.

Sadržaj

Predgovor

Tema ovog master rada je iz oblasti ekonometrije. Ekonometrija kao nauka je veoma značajna, zbog toga što nalazi široku primenu u stvarnom životu. To je relativno mlada nauka, nastala 30-ih godina dvadesetog veka. Ekonometrija na specifičan način povezuje ekonomiju, matematiku, statistiku i stvarne podatke.

Regresiona analiza je jedna od najčešće korišćenih alata u ekonometrijskom radu kako bi se opisale veze među pojavama. Često u regresionim modelima se pretpostavlja da regresiona funkcija ima jedan parametarski oblik tokom celog domena nezavisne promenljive. Međutim, u mnogim problemima je neophodno uzeti u obzir regresione modele koji imaju različite analitičke forme u različitim segmentima domena nezavisne promenljive. Važan specijalan slučaj je segmentirana regresija u kojoj je svaki segment regresione funkcije različita funkcija. Jedna klasa segmentiranog modela sastoji se od funkcija gde je svaki segment u formi linearnog modela.

Tema ovog rada se odnosi upravo na ovaj regresioni model, model segmetirane regresije. Model segmentirane regresije sa dva segmenta prvi je proučavao Kvant 1958. godine. Vremenom sve veći broj naučnika proučava ovaj model kako bi se povećala njegova efikasnost i učinkovitost. Fokus ovog master rada je na primeni ovog modela na podatke dobijene prilikom medicinskih istraživanja.

***

Ovom prilikom želela bih da se zahvalim svim profesorima i asistentima, sa kojima sam sarađivala tokom osnovnih i master akademskih studija.

Posebno bih se zahvalila svom profesoru i mentoru, dr Zorani Lužanin, na svim sugestijama i stručnom usmeravanju pri izradi ovog master rada, kao i na veoma zanimljivim predavanjima i prenetom znanju tokom studiranja.

Takođe, zahvalila bih se članovima komisije, dr Andreji Tepavčević i dr Dori Seleši.

Veliku zahvalnost dugujem svojoj porodici, posebno majci Svetlani, za podršku i razumevanje tokom celokupnog školovanja.

Suzana Vidić

1 Uvod

U zdravstvenim ustanovama se prikupljaju velike količine podataka, smeštenih u istorijama bolesti, praćenih dugi niz godina. Na ovakav način uskladišteni podaci teško mogu poslužiti za predviđanje ishoda bolesti ili ishoda lečenja novih pacijenata. Kao kvalitetan i savremen način prikupljanja, analize podataka i interpretacije rezultata, nudi se primena savremenih analitičko-statističkih metoda u svakodnevnom radu. Na osnovu istih mogu se saznati uzročno-posledične veze, sličnosti, razlike i zakonitosti, predikcija ishoda bolesti i planiranje adekvatnog tretmana, a samim tim i mogućnost pružanja kvalitetnijih usluga.

Kada se analizira učestalost pojave raka i stope smrtnosti, zdravstveni i medicinski istraživači su posebno zainteresovani da znaju da li je bilo promena u trendu tokom vremena, i ako je došlo do promena kada se to desilo. Ovakva pitanja igraju važnu ulogu u merenju napretka u borbi protiv raka i uticaja intervencije na ishod bolesti. U statističkim terminima, promena u trendu se može definisati kao promena nagiba u regresiji. Segmentirana regresija se može smatrati kao veoma značajan metod analize trendova i detektovanja tačke promene.

Model segmentirane regresije je model regresije, gde je veza između zavisne i jedne ili više nezavisnih promenljivih linearna po delovima, sa tačkom promene. Drugim rečima, zavisnost može biti predstavljena sa dve ili više pravih linija pridruženih odgovarajućim segmentima. U epidemološkim studijama, na primer, model segmentirane regresije se može koristiti kao prag model, gde se pretpostavlja da intervencija proizvodi efekat na zdravstveno stanje samo nakon prelaska nekog (često nepoznatog) praga. U medicini, na primer, može da se koristi za procenu efekata terapije. Pravilno upravljanje interakcijom lekova može da spreči neželjene događaje, a uspeh intervencije na osnovu promene u stopama kritičnih interakcija lekova se statistički procenjuje upotrebom segmentirane regresije.

Na početku rada ukratko definišemo okvir razmatranja koja slede. U drugom poglavlju su predstavljeni osnovni pojmovi vezani za linearnu regresiju i formulisan je model segmentirane regresije. U trećem poglavlju, sledi teorija ocena, gde je prikazano ocenjivanje parametara i testiranje hipoteza. Detaljno je objašnjen metod maksimalne verodostojnosti za ocenjivanje parametara, kao i testovi za postojanje tačke promene, Fišerov, Studentov i Dejvisov test. U četvrtom poglavlju su navedeni testovi koji detektuju tačku promene. To su test količnika verodostojnosti, zatim Švarcov informacioni kriterijum i neparametarski EL (*empirical*

likelihood) metod. U petom poglavlju, govori se o asimptotskom ponašanju parametara, tačnije o konzistentnosti, konvergenciji i asimptotskoj raspodeli. U šestom poglavlju definiše se koeficijent determinacije koji služi za procenu uspešnosti ocenjivanja modela. Ovo poglavlje razmatra i uticaj autlajera. Poslednje poglavlje je posvećeno primeni modela segmentirane regresije. Navedeno je nekoliko primera, a detaljno su predstavljena dva primera sa stvarnim podacima. Podaci su obrađivani upotrebom softverskog programa R. U dodatku su prikazani podaci korišćeni za kompjutersko izračunavanje kod primene modela.

2 Regresiona analiza

Pronalaženjem veza između pojava bavi se regresiona analiza. Regresiona analiza je od velikog značaja, kako u ekonomiji i privredi, tako i u prirodnim naukama, kao što su: hemija, fizika, biologija, farmakologija, toksikologija, biohemija, medicina i druge. Problem opisivanja ovakvih veza svodi se na pronalaženje modela koji povezuje jednu ili više *zavisnih* promenljivh Y sa jednom ili više *nezavisnihih*, objašnjavajućih, promenljivih X pomoću neke funkcionalne zavisnosti. Oblik ove funkcionalne zavisnosti je najčešće nepoznat, pa ostaje na istraživaču da izabere onu koja je po nekom kriterijumu najbolja. Veoma često se koriste polinomne funkcije, ali isto tako i eksponencijalne ili neke druge funkcije. Opšti problem nalaženja funkcije koja dobro aproksimira posmatrani skup podataka, često se naziva "*fitovanje"* krive ili određivanje *regresione linije*.

U medicinskim istraživanjima najčešće se sreće linearni model regresione analize, pa će se naša razmatranja odnositi na taj model. Segmentirana regresija je linearna regresija po delovima, stoga ćemo prvo navesti osnovne pojmove vezane za linearnu regresiju.

2.1 Linearna regresija

Veza između promenljivih može biti različitog oblika, a regresioni model kojim se opisuje linearna međuzavisnost između dve promenljive naziva se ***prosti linearni regresioni model*** [6], koji se definiše na sledeći način:

$$Y_i = \alpha + \beta X_i + \varepsilon_i\,, \qquad i = 1, \dots, n \tag{2.1.1}$$

gde su Y i ε slučajne promenljive, a X deterministička promenljiva. Nezavisna promenljiva X je kontrolisana, vrednosti zavisne promenljive Y se mogu meriti, dok se vrednosti promenljive ε, koja se naziva slučajna greška ili rezidual, ne mogu meriti, a α i β su nepoznati parametri regresije.

Klasične pretpostavke prostog linearnog regresionog modela su sledeće:

1° ***Sredina slučajne greške*** je jednaka nuli, što označavamo:

$$E(\varepsilon_i) = 0, \qquad i = 1, \dots, n. \tag{2.1.2}$$

2° ***Homoskedastičnost***: jednaka varijansa za sva opažanja, što označavamo:

$$Var(\varepsilon_i) = E[\varepsilon_i - E(\varepsilon_i)]^2 = \sigma^2, \qquad i = 1, \dots, n. \tag{2.1.3}$$

Ako varijansa slučajne greške nije ista za sva opažanja, već zavisi od neke nezavisne promenljive, govorimo o *heteroskedastičnosti*, koju označavamo sa $Var(\varepsilon_i) = {\sigma_i}^2$, $i = 1, \dots, n$, što nam govori da varijansa slučajnog odstupanja ε_i nije konstantna.

3° ***Odsustvo autokorelacije slučajnih odstupanja***: za dve fiksirane vrednosti x_i i x_j, za $i \neq j$, kovarijansa (korelacija) između dva slučajna odstupanja ε_i i ε_j, za bilo koje $i \neq j$ je nula, što označavamo:

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad \text{za } i \neq j, \quad i, j = 1, \dots, n. \tag{2.1.4}$$

4° ***Normalnost***: slučajna promenljiva $\varepsilon_i,\ i = 1, \dots, n$ ima normalnu raspodelu. (2.1.5)

5° ***Nestohastičnost promenljive X***: X je nestohastička promenljiva sa fiksnim vrednostima u ponovljenim uzorcima i takva da je, za bilo koji uzorak veličine n,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2.1.6}$$

različita od nula i da je njena granična vrednost konačan broj kada $n \to \infty$.

Iz pretpostavki (2.1.2), (2.1.3) i (2.1.5) sledi da je ε_i slučajna promenljiva koja ima normalnu raspodelu sa očekivanjem nula i disperzijom σ^2, što zapisujemo $\varepsilon_i: \mathcal{N}(0, \sigma^2)$.

Na osnovu specifikacije modela, opisanu regresionom jednačinom (2.1.1) i sa pet osnovnih pretpostavki ($(2.1.2) - (2.1.6)$), možemo odrediti raspodelu zavisne promenljive Y. Primenom matematičkog očekivanja na jednačinu regresije i koristeći pretpostavku modela da je $E(\varepsilon_i) = 0,\ i = 1, \dots, n$ dobijamo:

$$E(Y_i) = E(\alpha + \beta X_i + \varepsilon_i) = \alpha + \beta X_i, \qquad i = 1, \dots, n. \tag{2.1.7}$$

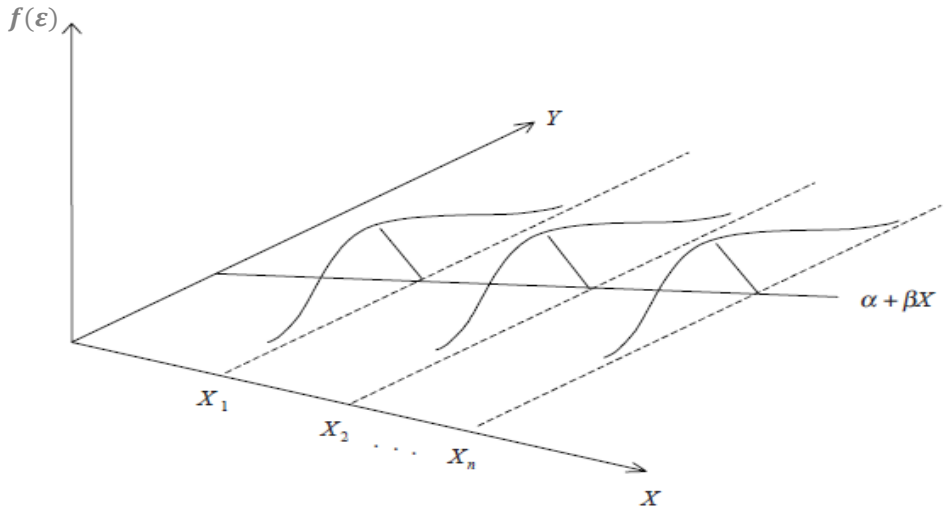
Pored toga, varijansa zavisne promenljive Y je:

$$Var(Y_i) = E[Y_i - E(Y_i)]^2$$

$$= E[\alpha + \beta X_i + \varepsilon_i - (\alpha + \beta X_i)]^2$$

$$= E(\varepsilon_i{}^2) = \sigma^2.$$

Pri izvođenju varijanse prvo smo iskoristili opštu definiciju varijanse, zatim smo uvrstili izraze za Y_i i $E(Y_i)$, jednačine (2.1.1) i (2.1.7), respektivno. Budući da je ε_i slučajna promenljiva sa normalnom raspodelom, a iz jednakosti (2.1.1) vidimo da je Y_i njena linearna transformacija, pa sledi da i Y_i ima normalnu raspodelu. Prema tome, zavisna promenljiva Y_i je slučajna promenljiva koja ima normalnu raspodelu sa očekivanjem $\alpha + \beta X_i$ i disperzijom σ^2, što zapisujemo Y_i: $\mathcal{N}(\alpha + \beta X_i, \sigma^2)$. To se može grafički prikazati na sledeći način:



Grafik 2.1.1: [1] *Raspodela slučajne promenljive Y*

Dalje imamo da je, na osnovu (2.1.2) i (2.1.4), za $i \neq j$

$$Cov(Y_i, Y_j) = E[(Y_i - E(Y_i))(Y_j - E(Y_j))] = E(\varepsilon_i \varepsilon_j) = 0,$$

$$i = 1, \dots, n, \qquad j = 1, \dots, n.$$

Slučajne promenljive $Y_1, Y_2, \dots, Y_n$ možemo posmatrati kao skup *n* normalno i nezavisno raspodeljenih promenljivih (nezavisne jer su ε_i i ε_j, $i, j = 1, \dots, n$ međusobno nezavisne).

Međutim, te promenljive nisu identično raspodeljene jer imaju različita očekivanja. Jednačina (2.1.7), koja daje očekivanu vrednost promenljive Y za svaku vrednost promenljive X je ***regresiona prava populacije***. Odsečak te linije, α, meri srednju vrednost promenljive Y koja odgovara vrednosti nula promenljive X. Nagib linije, β, meri promenu srednje vrednosti promenljive Y koja odgovara jedinici promene vrednosti promenljive X. Pošto su vrednosti tih parametara nepoznate, nepoznata je regresiona prava populacije. Kada se ocene vrednosti α i β, dobijamo ***regresionu pravu uzorka***, koja služi kao ocena regresione prave populacije. Neka su $\hat{\alpha}$ i $\hat{\beta}$ ocene za α i β, tada je regresiona prava uzorka

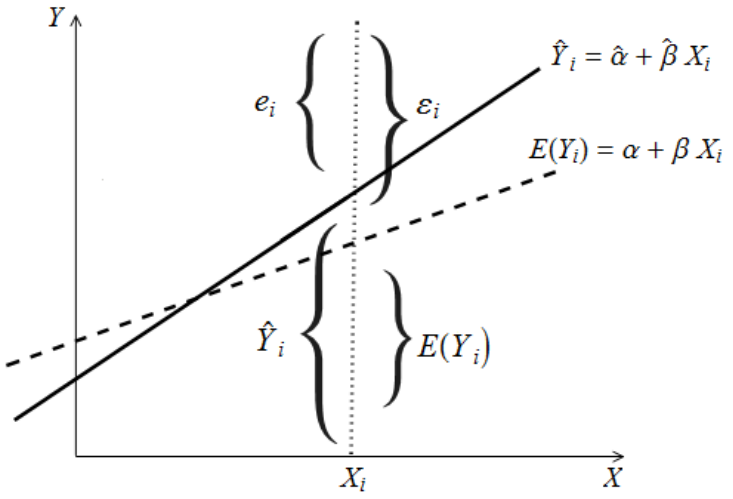
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i, \qquad i = 1, \dots, n$$

gde je $\hat{Y}_i$ prilagođena vrednost slučajne promenljive Y_i. Većina opaženih vrednosti promenljive Y neće ležati tačno na regresionoj liniji populacije, pa će se vrednosti Y_i i $\hat{Y}_i$ razlikovati. Ta razlika se naziva ostatak (rezidual) i označava se sa e_i. Stoga moramo razlikovati sledeće:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = 1, \dots, n \qquad \text{(populacija)}$$

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + e_i, \qquad i = 1, \dots, n \qquad \text{(uzorak)}$$

gde su e_i ocene za odstupanja ε_i. Ovo se može prikazati grafički na sledeći način:



Grafik 2.1.2: *Regresione prave za populaciju i uzorak*

Regresioni model kojim se opisuje linearna međuzavisnost između jedne zavisne i dve ili više nezavisnih promenljivih naziva se ***višestruka linearna regresija***. Model višestruke linearne regresije se definiše pomoću sledeće jednačine:

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_K X_{ik} + \varepsilon_i, \qquad i = 1, \dots, n \tag{2.1.8}$$

ili kraće

$$Y = \alpha + \sum_{j=1}^{k} \beta_j X_{ij}\,, \qquad i = 1, \dots, n,$$

gde je k broj nezavisnih promenljivih, a n veličina uzorka. Jednačina (2.1.8) se može zapisati u matričnom obliku kao:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

gde je

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}_{n\times(k+1)}, \quad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}.$$

Klasične pretpostavke modela višestruke linearne regresije su sledeće:

1° $E(\varepsilon_i) = 0,$ (2.1.9)

2° $Var(\varepsilon_i) = \sigma^2,$ (2.1.10)

3° $Cov(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j,\ i,j = 1, \dots, n,$ (2.1.11)

4° ε_i ima normalnu raspodelu, (2.1.12)

5° Sve nezavisne promenljive su determinističke, imaju fiksne vrednosti za različite uzorke i takve su bez obzira na veličinu uzorka,

$$\sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right)^2 \neq 0, \qquad j = 1, \dots, k \tag{2.1.13}$$

Granična vrednost ovog izraza je konačan broj, kada $n \to \infty$, za svako $j = 1, \dots, k,$

6° broj nezavisnih promenljivih mora biti manji od obima uzorka ($k<n$), (2.1.14)

7° ne postoji linearna veza između nezavisnih promenljivih. (2.1.15)

Prve četiri pretpostavke $\big((2.1.9)-(2.1.12)\big)$ su potpuno iste kao kod modela proste linearne regresije, a pretpostavka (2.1.13) je takođe ista kao i pretpostaka (2.1.6), samo je proširena na veći broj nezavisnih promenljivih. Pretpostavke od (2.1.9) do (2.1.12) mogu se zapisati u matričnom obliku kao $\boldsymbol{\varepsilon}: \mathcal{N}(\mathbf{0}, \Sigma)$, gde je $\mathbf{0}$ vektor nula, a $\Sigma = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)$ ili $\Sigma = \sigma^2 \boldsymbol{I}$, gde je $\boldsymbol{I}$ matrica identiteta dimenzije $n \times n$, sa jedinicama na dijagonali i nulama na svim ostalim mestima.

Pretpostavke od (2.1.13) do (2.1.15) se mogu zajedno izraziti na sledeći način: elementi matrice $\boldsymbol{X}$ su deterministički sa fiksnim vrednostima za različite uzorke, a matrica $\frac{1}{n}(\boldsymbol{X}^T\boldsymbol{X})$ je nesingularna i njeni elementi su konačni kada $n \to \infty$.

U slučaju malog uzorka (ispod 30 opažanja) poželjno je da ocene parametara linearne regresije imaju sledeće osobine:

1° **nepristrasnost**: Ocena je nepristrasna ako je očekivana vrednost ocene $\hat{\beta}$ jednaka stvarnoj vrednosti β, tj.

$$\lim_{n\to\infty} E(\hat{\beta}) = \beta.$$

2° **efikasnost**: Ocena $\hat{\beta}$ je efikasna ocena za parametar β, ako je:

- $\hat{\beta}$ nepristrasna ocena i
- $\hat{\beta}$ ima najmanju varijansu među svim ostalim nepristrasnim ocenama istog parametra, tj.

$$Var(\hat{\beta}) \leq Var(\tilde{\beta}), \text{ gde je } \tilde{\beta} \text{ bilo koja druga ocena za } \beta.$$

3° **BLUE - najbolja linearna nepristrasna ocena** (*eng. best linear unbiased estimator*): Ocena $\hat{\beta}$ ima ovu osobinu ako zadovoljava uslove da je ocena $\hat{\beta}$:

- linearna funkcija opažanja iz uzorka,
- nepristrasna i
- ima najmanju varijansu od svih ostalih linearnih nepristrasnih ocena za β.

Za ocene parametara poželjne su sledeće asimptotske osobine (kod velikih uzoraka):

1° **asimptotska nepristrasnost**: Asimptotska nepristrasnost podrazumeva da se povećanjem veličine uzorka dobija što bolja ocena koeficijenta, tj. očekivana vrednost ocene teži stvarnoj vrednosti parametra kako veličina uzorka raste, što zapisujemo

$$\lim_{n\to\infty} E(\hat{\beta}) = \beta.$$

2° **konzistentnost**: Ocena $\hat{\beta}$ je konzistentna ako konvergira u verovatnoći ka β, tj. ako za svako $\varepsilon > 0$ važi

$$\lim_{n\to\infty} P(|\hat{\beta} - \beta| < \varepsilon) = 1.$$

3° **asimptotska efikasnost**: Ocena $\hat{\beta}$ je asimptotski efikasna ukoliko zadovoljava uslove da je:

- konzistentna,
- ima najmanju asimptotsku disperziju i
- ima asimptotsku raspodelu sa konačnim očekivanjem i disperzijom.

2.2 Segmentirana regresija

Kada analiziramo vezu između zavisne promenljive, Y, i nezavisne promenljive, X, može biti očigledno da se za različite vrednosti promenljive X javljaju različite linearne veze. U tom slučaju, prosta linearna regresija ne može obezbediti adekvatan opis podataka i model segmentirane regresije je prikladniji. Segmentirana regresija je model regresije koji omogućava da za različite vrednosti X imamo više linearnih modela, pa se još naziva i po delovima linearna regresija. Tačke prekida ili prelomne tačke (*eng. breakpoint*) su vrednosti X u kojima se menja nagib linearne funkcije, nazivaju se još i tačke promene (*eng. changepoint*), prag vrednost ili čvor. Vrednost tačke promene može ili ne mora biti poznata pre analize, ali obično je nepoznata i potrebno je da se proceni. Regresiona funkcija može imati prekid u tački promene ili može biti neprekidna u svakoj tački, uključujući i tačku promene.

Model segmentirane regresije sadržane od dva segmenta, gde prvi segment ima odsečak α_1 i nagib β_1, a drugi odsečak α_2 i nagib $\beta_1 + \beta_2$, i tačke promene ξ, može se prikazati jednačinama za svaki segment:

Segment 1: $$y_i = \alpha_1 + \beta_1 x_i + \varepsilon_i, \quad \text{ako } x_i \leq \xi \tag{2.2.1}$$

Segment 2: $$y_i = \alpha_2 + \beta_2 x_i + \varepsilon_i, \quad \text{ako } x_i > \xi \tag{2.2.2}$$

gde je y_i, $i = 1, \ldots, n$ zavisna promenljiva, x_i, $i = 1, \ldots, n$ nezavisna promenljiva, α_1, α_2, β_1, β_2 i ξ su nepoznati parametri regresije i ε_i, $i = 1, \ldots, n$ su slučajne greške koje imaju normalnu raspodelu $\mathcal{N}(0, \sigma^2)$.

Uvodeći indikator promenljivu: $I = \begin{cases} 1, & x_i > \xi \\ 0, & x_i \leq \xi \end{cases}$, $i = 1, \ldots, n$

i kombinujući jednačine (2.2.1) i (2.2.2), model segmentirane regresije koji se sastoji od dva segmenta i samo jedne tačke promene, za $x = \xi$, može se definisati pomoću jedinstvene jednačine na sledeći način:

$$y_i = \alpha + \beta_1 x_i + \beta_2 (x_i - \xi) I + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{2.2.3}$$

Kao i kod proste linearne regresije, model segmentirane regresije zadovoljava klasične pretpostavke $(2.1.2) - (2.1.6)$.

Pod pretpostavkom da je $E(\varepsilon_i) = 0,\ i = 1, \ldots, n,$ sledi da je očekivanje zavisne promenljive do tačke promene ξ, za $x_i \leq \xi$

$$E(y_i) = \alpha + \beta_1 x_i, \qquad i = 1, \ldots, n$$

i da je očekivanje zavisne promenljive nakon tačke promene ξ, za $x_i > \xi$

$$\begin{aligned} E(y_i) &= \alpha + \beta_1 x_i + \beta_2 (x_i - \xi) \\ &= \alpha - \beta_2 \xi + (\beta_1 + \beta_2) x_i, \qquad i = 1, \ldots, n. \end{aligned}$$

Dakle, β_1 je nagib regresione linije u prvom segmentu, $\beta_1 + \beta_2$ je nagib regresione linije u drugom segmentu.

U slučaju kada imamo dve tačke promene ξ_1 i ξ_2, onosno tri segmenta, model se definiše na sledeći način [9]:

$$y_i = \alpha + \beta_1 x_i + \beta_2 (x_i - \xi_1) I_1 + \beta_3 (x_i - \xi_2) I_2 + \varepsilon_i, \qquad i = 1, \ldots, n,$$

gde je $I_k = \begin{cases} 1, & x_i > \xi_k \\ 0, & x_i \leq \xi_k \end{cases}$ indikator promenljiva, za $k = 1,2.$

Analogno slučaju sa jednom tačkom promene, sledi da je u ovom slučaju očekivanje zavisne promenljive po segmentima

$$E(y_i) = \begin{cases} \alpha + \beta_1 x_i, & x_i \leq \xi_1, \\ \alpha - \beta_2 \xi_1 + (\beta_1 + \beta_2) x_i, & \xi_1 < x_i \leq \xi_2, \\ \alpha - \beta_2 \xi_1 - \beta_3 \xi_2 + (\beta_1 + \beta_2 + \beta_3) x_i, & x_i > \xi_2, \end{cases} \quad i = 1, \ldots, n.$$

U opštem slučaju, sa K čvorova (tačaka promene) i $K + 2$ parametra, model segmentirane regresije se definiše na sledeći način:

$$y_i = \alpha + \beta_1 x_i + \sum_{k=1}^{K} \beta_{k+1} (x_i - \xi_k) I_k + \varepsilon_i, \qquad i = 1, \ldots, n,$$

gde je $I_k = \begin{cases} 1, & x_i > \xi_k \\ 0, & x_i \leq \xi_k \end{cases}$, za $k = 1, \ldots, K.$

Segmentirana regresija je primer opštije klase funkcija poznatijih kao splajn funkcije. Velika prednost segmentirane linearne regresije u odnosu na ostale regresione splajnove je u jednostavnijem konceptu i implementaciji. Kao što je već spomenuto, razlikujemo dva tipa ovog

modela: neprekidni i sa prekidima. Neprekidni slučaj znači da je regresiona funkcija neprekidna u tački promene $x = \xi$, tako da u slučaju sa jednom tačkom promene treba da bude zadovoljena jednakost:

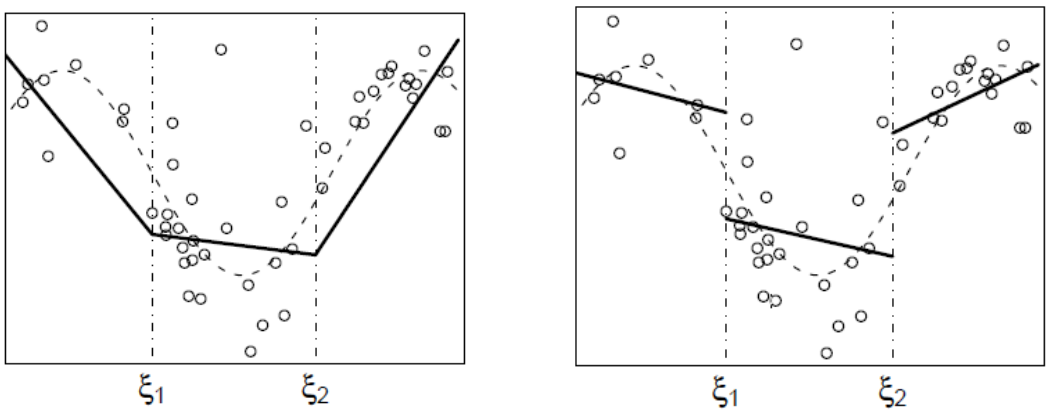
$$\alpha_1 + \beta_1 \xi = \alpha_2 + \beta_2 \xi. \tag{2.2.4}$$

Ukoliko jednakost (2.2.4) nije zadovoljena, model nije neprekidan.

Dakle, u slučaju sa K čvorova model je neprekidan ako je zadovoljena sledeća jednakost:

$$\alpha_k + \beta_k \xi_k = \alpha_{k+1} + \beta_{k+1} \xi_k, \qquad k = 1, \dots, K.$$

Kada postoje dve tačke promene i tri segmenta, grafički se može prikazati na sledeći način:



Grafik 2.2.1: [8] *Neprekidna i sa prekidima segmentirana regresija*

U ovom radu ograničavamo pažnju na neprekidan slučaj segmentirane regresije, gde su svi segmenti regresione funkcije u formi linearnog modela.

Model proste segmentirane linearne regresije se može uopštiti do višestruke segmentirane linearne regresije. ***Model višestruke segmentirane regresije*** sa dva segmenta se definiše na sledeći način:

$$y_i = \begin{cases} \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + \varepsilon_{1i}, & i = 1, \dots, k \\ \boldsymbol{x}_i^T \boldsymbol{\beta}_2 + \varepsilon_{2i}, & i = k+1, \dots, n \end{cases} \tag{2.2.5}$$

gde je $\boldsymbol{\beta}_d$, $d = 1, 2$, vektor dimenzije $p \times 1$.

Da bi $\boldsymbol{\beta}_1$ i $\boldsymbol{\beta}_2$ bili procenjivi (tj. broj posmatranja za svaki segment je najmanje p), ograničićemo k tako da je $k = p, p+1, \ldots, n-p$. Pored toga, pretpostavljamo da su slučajne greške ε_{di} nezavisne i $\varepsilon_{di} \sim \mathcal{N}(0, \sigma_d^2)$, za $d = 1,2$. Obično pretpostavljamo da je $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Za fiksirano k, model višestruke segmentirane regresije označen sa (2.2.5) se može zapisati u matričnom obliku na sledeći način:

$$\boldsymbol{Y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1,$$
$$\boldsymbol{Y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2, \qquad (2.2.6)$$

gde je $\boldsymbol{Y}_1$ vektor dimenzije $k \times 1$ i $\boldsymbol{Y}_2$ je vektor dimenzije $(n-k) \times 1$.

U opštem slučaju, sa K tačaka promene i $K+2$ parametra, p nezavisnih promenljivih, model višestruke segmentirane regresije se definiše na sledeći način:

$$Y_i = \alpha_k + \sum_{j=1}^{p} \beta_{jk}\, x_j\, \boldsymbol{I}_{jk}(\boldsymbol{x}) + \varepsilon_i, \qquad i = 1, \ldots, n, \qquad k = 1, \ldots, K.$$

gde je $\boldsymbol{I}_{jk}(\boldsymbol{x}) = \begin{cases} \mathbf{1}, & \text{ako } x_i \in [\xi_{k-1}, \xi_k) \\ \mathbf{0}, & \text{inače} \end{cases}$, $k = 1, \ldots, K$.

3 Ocenjivanje parametara segmentirane regresije

Cilj ocenjivanja parametara je dobiti ocene koje će imati što je više moguće poželjnih osobina. Takve ocene se mogu potom upotrebiti za testiranje hipoteza koje se tiču regresionog modela. Ocenjivanje parametara se vrši pomoću metode najmanjih kvadrata, metodom momenata, metodom maksimalne verodostojnosti i metodom najboljih linearnih nepristrasnih ocenjivača (BLUE metod). Mi ćemo predstaviti metod maksimalne verodostojnosti [10].

Pretpostavka o homoskedastičnosti, potrebna za dobijanje ocena metodom maksimalne verodostojnosti u opštem slučaju modela, ovde nije neophodna, jer se u slučaju segmentirane regresije poklapaju ocene dobijene metodom najmanjih kvadrata i metodom maksimalne verodostojnosti, a metoda najmanjih kvadrata ne zahteva pretpostavku o homoskedastičnosti. Kao takve, ocene dobije metodama metodom najmanjih kvadrata i metodom maksimalne verodostojnosti ostaju nepristrasne, čak i ako se varijansa menja tokom segmenata.

3.1 Metoda maksimalne verodostojnosti

Da bismo našli ocene metodom maksimalne verodostojnosti prvo treba da odredimo funkciju verodostojnosti za opažanja u uzorku i potom je maksimizirati po nepoznatim parametrima. U slučaju našeg regresionog modela, uzorak sadrži n opažanja.

Funkciju verodostojnosti možemo prikazati kao

$$\ell = \varphi(y_1)\varphi(y_2)\cdots\varphi(y_n).$$

Budući da su vrednosti koje maksimiziraju ℓ iste kao i vrednosti koje maksimiziraju njen logaritam, mi ćemo maksimizirati

$$L = \log \ell = \sum_{i=1}^{n} \log \varphi(y_i).$$

Posmatraćemo segmentiranu regresiju sa dva segmenta, zbog jednostavnosti, a celokupni postupak se može primeniti i na slučaj segmentirane regresije sa više od dva segmenata. Neka su

Segment 1: $$y_1 = \alpha_1 + \beta_1 x_i + \varepsilon_1, \quad i = 1, \dots, k \tag{3.1.1}$$

Segment 2: $$y_2 = \alpha_2 + \beta_2 x_i + \varepsilon_2, \quad i = k+1, \dots, n \tag{3.1.2}$$

gde su ε_1 i ε_2 nezavisne i normalno raspodeljene slučajne promenljive sa očekivanjem nula i standardnim devijacijama σ_1 i σ_2. Neka su ova dva segmenta generisana za ukupno n posmatranja i neka je ξ tačka promene. Pretpostavimo da su prvih k posmatranja generisana sa (3.1.1), a preostalih $n-k$ sa (3.1.2).

Budući sa su y_i normalno raspodeljene slučajne promenljive sa očekivanjem $(\alpha_i + \beta_i x_i)$ i disperzijom σ_i^2 za segment i, sledi da su funkcije gustine za y_1 u tački i i za y_2 u tački j:

$$\varphi(y_1) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_1^2}(y_i - \alpha_1 - \beta_1 x_i)^2\right\}$$

i

$$\varphi(y_2) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_2^2}\left(y_j - \alpha_2 - \beta_2 x_j\right)^2\right\}$$

Funkcije verodostojnosti za uzorak obima k iz (3.1.1) i uzorak obima $n-k$ iz (3.1.2) su

$$\ell_1 = \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^k \exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{k}(y_i - \alpha_1 - \beta_1 x_i)^2\right\}$$

i

$$\ell_2 = \left(\frac{1}{\sigma_2\sqrt{2\pi}}\right)^{n-k} \exp\left\{-\frac{1}{2\sigma_2^2}\sum_{j=k+1}^{n}\left(y_j - \alpha_2 - \beta_2 x_j\right)^2\right\}$$

i funkcija verodostojnosti za celokupan uzorak je

$$\ell = \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^k \left(\frac{1}{\sigma_2\sqrt{2\pi}}\right)^{n-k} \exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{k}(y_i - \alpha_1 - \beta_1 x_i)^2 - \frac{1}{2\sigma_2^2}\sum_{j=k+1}^{n}\left(y_j - \alpha_2 - \beta_2 x_j\right)^2\right\}.$$

Logaritam funkcije verodostojnosti je

$$L = -n\log\sqrt{2\pi} - k\log\sigma_1 - (n-k)\log\sigma_2 - \frac{1}{2\sigma_1^2}\sum_{i=1}^{k}(y_i - \alpha_1 - \beta_1 x_i)^2 - \frac{1}{2\sigma_2^2}\sum_{j=k+1}^{n}\left(y_j - \alpha_2 - \beta_2 x_j\right)^2. \qquad (3.1.3)$$

Računajući parcijalne izvode za (3.1.3) po α_1, α_2, β_1, β_2 i izjednačavajući ih sa nula dobijamo sledeće ocene za ove parametre:

$$\hat{\beta}_1 = \frac{k\sum_{i=1}^{k} x_i y_i - \sum_{i=1}^{k} x_i \sum_{i=1}^{k} y_i}{k\sum_{i=1}^{k} {x_i}^2 - \left(\sum_{i=1}^{k} x_i\right)^2}, \qquad \hat{\alpha}_1 = \frac{\sum_{i=1}^{k} y_i}{k} - \hat{\beta}_1\frac{\sum_{i=1}^{k} x_i}{k},$$

$$\hat{\beta}_2 = \frac{(n-k)\sum_{i=k+1}^{n} x_i y_i - \sum_{i=k+1}^{n} x_i \sum_{i=k+1}^{n} y_i}{(n-k)\sum_{i=k+1}^{n} {x_i}^2 - (\sum_{i=k+1}^{n} x_i)^2}, \qquad \hat{\alpha}_2 = \frac{\sum_{i=k+1}^{n} y_i}{n-k} - \hat{\beta}_2\frac{\sum_{i=k+1}^{n} x_i}{n-k}.$$

Računajući parcijalne izvode za (3.1.3) po σ_1 i σ_2 i izjednačavajući ih sa nula i uvrštavajući dobijene ocene za α_1, α_2, β_1, β_2 dobijamo sledeće ocene:

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{k}\left(y_i - \hat{\alpha}_1 - \hat{\beta}_1 x_i\right)^2}{k}$$

$$\hat{\sigma}_2^2 = \frac{\sum_{i=k+1}^{n}\left(y_j - \hat{\alpha}_2 - \hat{\beta}_2 x_j\right)^2}{n-k}.$$

Zamenom ovih ocena u formulu (3.1.3) dobijamo:

$$L(k) = -n\log\sqrt{2\pi} - k\log\hat{\sigma}_1 - (n-k)\log\hat{\sigma}_2 - \frac{n}{2} \qquad (\sigma_1^2 \neq \sigma_2^2) \qquad (3.1.4)$$

što predstavlja maksimum logaritma funkcije verodostojnosti za date vrednosti n i to je funkcija koja zavisi samo od k. U slučaju homoskedastičnosti imamo da je

$$L(k) = -n\log\sqrt{2\pi} - n\log\hat{\sigma} - \frac{n}{2}. \qquad (\sigma_1^2 = \sigma_2^2 = \hat{\sigma}) \qquad (3.1.5)$$

Kod višestruke segmentirane regresije, ocene dobijene metodom maksimalne verodostojnosti su:

$$\widehat{\boldsymbol{\beta}}_d = (\boldsymbol{X}_d^T\boldsymbol{X}_d)^{-1}\boldsymbol{X}_d^T\boldsymbol{Y}_d, \qquad d = 1, 2,$$

$$\hat{\sigma}_1^2 = \frac{\hat{s}_1}{k}, \qquad \hat{\sigma}_2^2 = \frac{\hat{s}_2}{n-k},$$

gde je

$$\hat{s}_d = \left(\boldsymbol{Y}_d - \boldsymbol{X}_d\widehat{\boldsymbol{\beta}}_d\right)^T\left(\boldsymbol{Y}_d - \boldsymbol{X}_d\widehat{\boldsymbol{\beta}}_d\right)$$

suma kvadrata reziduala za d-ti segment, $d = 1, 2$.

Ako je $\sigma_1^2 = \sigma_2^2 = \sigma^2$, tada

$$\hat{\sigma}^2 = \frac{\hat{s}_1 + \hat{s}_2}{n}.$$

Ocena metodom maksimalne verodostojnosti za k, $\hat{k}$, dobija se kada se maksimizira funkcija $L(k)$ zadata formulom (3.1.4), po $k = p, p+1, \ldots, n-p$. Kada tražimo vrednost k koja maksimizira (3.1.4), obično bismo jednom diferencirali $L(k)$ po k i izjednačili taj izvod sa nula. Međutim, ta procedura je neodgovarajuća, pošto k nije neprekidna promenljiva. Niti je to pouzdana tehnika za traženje vrednosti $\hat{k}$ za koje važi:

$$L(\hat{k}-1) < L(\hat{k}) \qquad \text{i} \qquad L(\hat{k}+1) < L(\hat{k}),$$

pošto možda postoji nekoliko maksimuma, a ta tehnika je nesposobna da napravi razliku između njih. Stoga, preporučuje se sledeći postupak: izračunati vrednosti funkcije verodostojnosti iz (3.1.4) za sve moguće vrednosti od k i izabrati kao ocenu onu vrednost k koja odgovara najvećem maksimumu.

Postavlja se pitanje da li je moguće osmisliti test za hipotezu da se nije dogodila promena ili prekid tokom perioda posmatranja. Test količnika funkcija verodostojnosti (*eng. likelihood*

ratio test), tzv. količnik verodostojnosti se može pokazati korisnim u testiranju hipoteze da nema tačke promene protiv alternativne hipoteze da postoji promena.

3.2 Višestruka tačka promene

Model segmentirane regresije može imati jednu ili više tačaka promene. Višestruka tačka promene $\boldsymbol{\xi} = (\xi_1, \dots, \xi_K)^T$ kod iste promenljive X može nastati na najmanje dva načina:

1) segmentirana veza je različita među nivoima $w_1, w_2, \dots, w_K$ za neku kategorijalnu promenljivu W i takva da postoji jedna tačka promene ξ_k za svaku grupu $k = 1, \dots, K$,

2) odnos između zavisne i nezavisne promenljive X doživljava nekoliko promena u odnosu na K tačaka promene. Ovo može biti protumačeno kao specijalni slučaj prethodnog slučaja (pod 1).

Pretpostavimo K nivoa, koristićemo parametrizaciju datu sa [13]:

$$\boldsymbol{\xi} = \xi_1 W_1 + \xi_2 W_2 + \cdots + \xi_K W_K, \qquad k = 1, \dots, K,$$

gde je $W_k = 1$ za posmatranja koja pripadaju grupi k, a u suprotnom $W_k = 0$, ξ_k je tačka promene u grupi k.

U pitanju je nelinearni izraz koji zavisi od interakcije dve promenljive X i W, zapravo imamo proizvod K izraza $(X \times W_k)$:

$$\sum_{k=1}^{K} \alpha_k (X \times W_k) + \sum_{k=1}^{K} \beta_k \big((X \times W_k) - \xi_k\big) I_k \, .$$

Razvijajući ovo u Tejlorov polinom prvog reda u okolini $\xi_k^{(s)}$ dobijamo $2K$ novih promenljivih u svakoj iteraciji s: $U_k = \left((X \times W_k) - \xi_k^{(s)}\right) I_k$ i $V_k = -I_k \left((X \times W_k) > \xi_k^{(s)}\right)$, za $k = 1, \dots, K$. Onda sledi da

$$\sum_{k=1}^{K} \alpha_k (X \times W_k) + \sum_{k=1}^{K} \beta_k U_k + \sum_{k=1}^{K} \gamma_k V_k \tag{3.2.1}$$

su linearni izrazi modelirane segmentirane regresije sa tačkama promene u zavisnosti od kategorijalne promenljive W. Naime model segmentirane regresije je sveden na iterativno fitovanje linearnog modela preko promenljivih U i V. Koeficijent β koji stoji uz U predstavlja razliku u nagibu među segmentima, a koeficijent γ koji stoji uz V može se posmatrati kao

reparametrizacija od ξ. U svakoj iteraciji s, koeficijent γ meri razliku između dve fitovane prave linije (pre i posle $\xi^{(s)}$) u $X = \xi^{(s)}$. Pošto posmatramo neprekidan slučaj, koeficijent γ je nula, pa kada algoritam konvergira očekuje se da $\hat{\gamma}$ bude oko nula. Sukcesivne aproksimacije za tačke promene su date sa:

$$\xi_k^{(s+1)} = \frac{\hat{\gamma}_k}{\hat{\beta}_k} + \xi_k^{(s)}, \qquad k = 1, \dots, K.$$

Poboljšanja u proceni tačke promene zavise od procena dobijenih metodom maksimalne verodostojnosti, tj. od

$$\hat{\xi} = \frac{\hat{\gamma}}{\hat{\beta}} + \xi^{(0)}.$$

Kada se algoritam zaustavi i $\hat{\gamma} \approx 0$, nema poboljšanja u proceni tačke promene i zbog toga je s-ta aproksimacija zapravo procena dobijena metodom maksimalne verodostojnosti, tj. $\xi^{(s)} = \hat{\xi}$.

Višestruka tačka promene koja se odnosi na istu segmentiranu vezu se nelinearno modelira pomoću

$$\alpha X + \sum_{k=1}^{K} \beta_k (X - \xi_k)\, I_k.$$

Prema ovoj parametrizaciji α je „prvi nagib“, tj. kada je $X \leq \xi_1$, a β_k je razlika u nagibima pre i posle tačke promene ξ_k, tj. razlika između k-tog i $(k+1)$-og nagiba. Sledi da je $\alpha + \sum_{k=1}^{\tilde{k}} \beta_k$ nagib za $\xi_{\tilde{k}} < X \leq \xi_{\tilde{k}+1}$.

Pretpostavljajući višedimenzionalnu tačku promene $\boldsymbol{\xi}$ i $W_k = 1$ za svako k, tada (3.2.1) postaje

$$\alpha X + \sum_{k=1}^{K} \beta_k U_k + \sum_{k=1}^{K} \gamma_k V_k$$

i služi za rukovanje višestrukim promenama u pojedinačnim segmentiranim odnosima.

Iako bi u principu trebalo da bude moguće da se procenjuje bilo koji višedimenzionalni parametar, nekoliko tačaka promena (najčešće jedna do tri) su verovatno dovoljne za rukovanje nekoliko praktičnih situacija, jer značenje „tačka promene“ postaje veliki znak pitanja kada se njihov broj povećava.

3.3 Testiranje hipoteze

Posmatramo hipotezu za prostu linearnu regresiju, tj.

$$H_0 : \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon}: \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n) \tag{3.3.1}$$

protiv alternativne hipoteze

$$H_1: \boldsymbol{Y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1,$$
$$\boldsymbol{Y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2.$$

Maksimum logaritma funkcije verodostojnosti za model (3.3.1) je

$$L(n) = n \log \sqrt{2\pi} - n \log \tilde{\sigma} - \frac{n}{2}, \tag{3.3.2}$$

gde je $\tilde{\sigma}^2 = \frac{\tilde{s}}{n}$, a $\tilde{s}$ je suma kvadrata reziduala za (3.3.1). Tada je $-2(LLR$[1]$)$ za testiranje hipoteze (koristeći (3.1.4) i (3.3.2)):

$$L_1 = [n \log \tilde{\sigma}^2 - k \log \hat{\sigma}_1^2 - (n-k) \log \hat{\sigma}_2^2]_{k=\hat{k}} .$$

Worsley smatra da je ovaj test dobar i za promenu u disperziji, kao i za promenu u regresiji. Pod pretpostavkom homoskedastičnosti, iz (3.1.5) i (3.3.2), $-2(LLR)$ je

$$L_2 = n \log \left[\frac{\tilde{s}}{\hat{s}_1 + \hat{s}_2}\right]_{k=\hat{k}}$$

$$= n \max_{p \le k \le n-p} \log \frac{\tilde{s}}{\hat{s}_1 + \hat{s}_2} . \tag{3.3.3}$$

Prema uobičajenoj asimptotskoj teoriji L_2 bi trebalo da ima asimptotsku χ^2_{p+1} raspodelu. Međutim, standardna teorija ne važi u slučaju segmentirane regresije, jer k uzima samo diskretne vrednosti i (3.3.1) važi i ukoliko je promena izvan domena podataka. Količnik verodostojnosti (LLR) nema ograničenu raspodelu, ali teži ka beskonačnosti, kada $n \to \infty$.

[1] *LLR* (*eng. log-likelihood ratio*) je tzv. količnik verodostojnosti, koji je objašnjen u Poglavlju 4

3.3.1 Fišerov (Fisher) test

Za poznato k uobičajena F-test statistika za hipotezu H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ u modelu (2.2.8) je, pod pretpostavkom homoskedastičnosti,

$$F_k = \frac{[\tilde{s} - (\hat{s}_1 + \hat{s}_2)]/p}{(\hat{s}_1 + \hat{s}_2)/(n-2p)}$$

$$= \left(\frac{\tilde{s}}{\hat{s}_1 + \hat{s}_2} - 1\right)\frac{n-2p}{p},$$

koja ima Fišerovu $F_{p,n-2p}$ raspodelu kada prihvatamo hipotezu H_0. Intuitivno privlačan postupak je da se zasnuje test na

$$F_{\max} = \max_{p \le k \le n-p} F_k. \qquad (3.3.1.1)$$

Štaviše, postupak dat sa (3.3.1.1) je jasan ekvivalent postupku količnika verodostojnosti (3.3.3). Backman, Cook i Worsley formiraju aproksimaciju za nula raspodelu od $F_{\max}$. Sada $F_{\max}$ je maksimum od $n - 2p + 1$ koreliranih F-statistika F_k, koje će biti veoma tesno povezane za susedne vrednosti k. To se može prikazati tako da ako je $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}$ F-statistika F_k za testiranje hipoteze H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ u modelu (2.2.6) ne zavisi od parametara $\boldsymbol{\beta}$ i σ^2, iako jasno zavisi od tačke promene k.

Tada nula raspodela za $F_{\max}$ je nezavisna od parametara $\boldsymbol{\beta}$ i σ^2 i zavisi samo od matrice dizajna za nezavisnu promenljivu $\boldsymbol{X}$. Za bilo koji određeni skup podataka može da se simulira raspodela za $F_{\max}$ koristeći proizvoljne vrednosti za $\boldsymbol{\beta}$ i σ^2 (na primer $\boldsymbol{\beta} = \mathbf{0}$, $\sigma^2 = 1$). Hipoteza će biti odbačena na nivou značajnosti α ako je $F_{\max}$ izvan simuliranih vrednosti.

Za model segmentirane regresije sa dva segmenta i jednom tačkom promene kada testiramo da li ima promene u regresionim parametrima kada je tačka promene poznata, Fišerova test statistika je

$$F = \frac{SSE_1 - SSE_2}{SSE_2/(n-3)}$$

gde su SSE_1 i SSE_2 sume kvadrata reziduala za prvi i drugi segment, redom, n je broj posmatranja. Test statistika ima Fišerovu raspodelu sa 1 i $n-3$ stepeni slobode.

Za testiranje promene u nagibu regresione linije, tj. $H_0: \beta_1 = 0$, koristi se sledeća test statistika:

$$F = \frac{(SSE_1 - SSE_2)/2}{SSE_2/(n-4)}$$

koja ima Fišerovu raspodelu sa 2 i $n-4$ stepena slobode.

Takođe za testiranje hipoteze da je koeficijent $\beta_j = 0$, koristi se standardizovani koeficijent ili z-score:

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}},$$

gde je v_j j-ti dijagonalni elemenat matrice $(X^T X)^{-1}$. Pod nultom hipotezom da je $\beta_j = 0$, z_j ima Studentovu t raspodelu sa $N-p-1$ stepeni slobode (gde je N veličina uzorka, a p broj nepoznatih parametara), i stoga će velike (apsolutne) vrednosti z_j dovesti do odbacivanja nulte hipoteze. Z-score veći od 2 u apsolutnoj vrednosti je statistički značajan na nivou značajnosti 5%.

Za testiranje iste hipoteze o promeni u nagibu regresione linije koristi se još i Studentov t-test. Odgovarajuća test statistika je:

$$t = \frac{\hat{\beta}}{SE_{\hat{\beta}}}$$

koja ima Studentovu t- raspodelu sa $n-2$ stepeni slobode ako je se prihvata nulta hipoteza. Standardna greška za koeficijent nagiba je:

$$SE_{\hat{\beta}} = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

3.3.2 Dejvisov (Davies) test

Ako tačka promene ne postoji, tada je parametar razlike u nagibu nula, pa je test za postojanje tačke promene ξ:

$$H_0: \ \beta_2(\xi) = 0$$

Naglasimo da je $\beta_2(\xi)$ parametar koji nas interesuje, β_2 zavisi od parametra (tačke promene) ξ, koji nestaje pod hipotezom H_0. Lako je pretpostaviti da nekoliko nepravilnosti poseduje takva hipoteza, i kao posledica toga, standardni statistički testovi (na primer, Wald-ov) možda nisu primenljivi. Tačnije, p-vrednost dobijena klasičnim testovima je u velikoj meri potcenjena, odnosno empirijski nivo p-vrednosti je tri do pet puta veći od nivoa značajnosti. Dejvisov test je pogodan za testiranje ove hipoteze [12]. On funkcioniše na sledeći način: neka je dato K fiksiranih uređenih vrednosti za tačke promene $\xi_1 < \xi_2 < \cdots < \xi_K$ u opsegu za X i relevantnih K vrednosti test statistike $\{S(\xi_k)\}_{k=1,\dots,K}$ imaju standardnu normalnu raspodelu za fiksirano ξ_k, Dejvis predviđa gornju granicu kao

$$p - vrednost \approx \Phi(-M) + V\, e^{\frac{-M^2}{2}} \sqrt{8\pi} \qquad (3.3.2.1)$$

gde je $M = \max\{S(\xi_k)\}_k$ od K test statistika, $\Phi(\cdot)$ je funkcija standardne normalne raspodele i

$$V = \sum_{k=1}^{K} |S(\xi_k) - S(\xi_{k-1})|$$

je ukupna varijacija od $\{S(\xi_k)\}_k$. Formula (3.3.2.1) je gornja granica, stoga je ta p-vrednost donekle potcenjena i test je pomalo konzervativan. Dejvisov test ne obezbeđuje smernice za izbor broja i lokacije fiksnih vrednosti $\{\xi_k\}_{k=1,\dots,K}$. Neki simulacioni eksperimenti pokazuju da je obično dovoljno da $K \in [5,10]$. Formula (3.3.2.1) se odnosi na jednostrano testiranje hipoteze, pa je alternativna hipoteza:

$$H_1: \ \beta_2(\xi) > 0.$$

p-vrednost za alternativnu hipotezu se dobija korišćenjem da je $M = \min\{S(\xi_k)\}_k$, dok u slučaju dvostranog testa se uzima da je $M = \max\{|S(\xi_k)|\}_k$ i dvostruki test formule (3.3.2.1).

Dejvisov test je pogodan za testiranje da li postoji tačaka promene, ali ne i za ispitivanje broja tačaka promene.

4 Detektovanje tačke promene

U poslednjih trideset godina, razvijene su značajne tehnike za testiranje hipoteza, ocenjivanje parametara i odgovarajući računarski programi, za detektovanje tačke promene u segmentiranoj regresiji.

Posmatramo slučaj segmentirane regresije sa dva segmenta i jednom tačkom promene:

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i, & x_i \leq \xi \\ \alpha_2 + \beta_2 x_i + \varepsilon_i, & x_i > \xi \end{cases}, \qquad i = 1, \ldots, n, \tag{4.1}$$

gde su $\{\varepsilon_i\}_{i=1}^n$ slučajne promenljive sa očekivanjem nula. Neka je $\{x_{(i)}\}_{i=1}^n$ statistički redosled za nezavisnu promenljivu $\{x_i\}_{i=1}^n$. Ako je k^* takvo da $x_{(k^*)} \leq \xi \leq x_{(k^*+1)}$, onda se k^* naziva trenutkom promene, a ξ tačkom promene u tom trenutku.

Pre primene modela opisanog sa formulom (4.1) neophodno je testirati postojanje tačke promene. Postoje dva tipa pristupa zasnovana na funkciji verodostojnosti: Švarcov informacioni kriterijum (*eng. Schwarz information criteria*), u oznaci SIC, i klasični parametarski metod maksimalne verodostojnosti [15].

U ovom delu, bavićemo se pomenutim problemom koristeći i nedavno razvijen neparametarski empirijski pristup verovatnoće (*eng. nonparametric empirical likelihood approach*). Empirijsku verovatnoću (EL) kao neparametarsku tehniku za upravljanje podacima je prvi predložio Owen. EL izračunava funkciju verodostojnosti bez prethodne pretpostavke o raspodeli podataka.

4.1 Test količnika verodostojnosti (Q-test)

Jednostavan segmentiran model, kada posmatramo niz tačaka (x_i, y_i), $i = 1, \dots, n$, može se zapisati na sledeći način:

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i, & x_i \leq \xi \\ \alpha_2 + \beta_2 x_i + \varepsilon_i, & x_i > \xi \end{cases} \tag{4.1.1}$$

gde su α_1, α_2, β_1, β_2, ξ nepoznati parametri, ε_i slučajne greške koje imaju normalnu raspodelu $\mathcal{N}(0, \sigma^2)$, a ξ je tačka promene.

Segmentirani model koji predlaže Kvant (*eng. Quandt*) je sličan modelu (4.1.1). Razlika je u tome što prethodna definicija modela pretpostavlja homoskedastičnost, dok Kvantov model pretpostavlja heteroskedastičnost. Za posmatrani niz tačaka (x_i, y_i), $i = 1, \dots, n$, Kvantov segmentirani model se definiše na sledeći način:

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i, & i = 1, \dots, k \\ \alpha_2 + \beta_2 x_i + \varepsilon_i, & i = k+1, \dots n \end{cases}$$

gde su ε_i nezavisne i normalno raspodeljene slučajne promenljive sa očekivanjem nula i standardnom devijacijom σ_1, ako je $i \leq k$ i σ_2, ako je $i > k$.

Postoje različiti testovi za detektovanje tačke promene, zasnovani na odnosu funkcija verodostojnosti [2]. Kvant je prvi predložio test količnika funkcija verodostojnosti za detektovanje tačke promene u jednostavnom linearnom regresionom modelu.

Test statistika količnika funkcija verodostojnosti je

$$\Lambda = \max_{3 \leq k \leq n-3} \{\lambda(k)\},$$

pri čemu je

$$\lambda = \frac{L(k)}{L(n)},$$

gde je $L(n)$ maksimum logaritma funkcije verodostojnosti za linearnu regresiju (samo jedan segment) i $L(k)$ je maksimum logaritma funkcije verodostojnosti za segmentiranu regresiju (prisustvo tačke promene). Uvrštavanjem formule (3.1.4) za $L(k)$ i formule (3.3.2) za $L(n)$ dobijamo da je

$$\lambda(k) = -2\log\left(\frac{\hat{\sigma}_1^k(k)\hat{\sigma}_2^{n-k}(k)}{\hat{\sigma}^n}\right),$$

gde je $\hat{\sigma}$ ocena standardne devijacije proste linarne regresije (obuhvata sva posmatranja), $\hat{\sigma}_1$ i $\hat{\sigma}_2$ su ocene za σ_1 i σ_2 za fiksirano k, respektivno, a k je izabrano tako da maksimizira λ. Velike vrednosti Λ impliciraju postojanje tačke promene.

Ovaj test se koristi za testiranje hipoteze da nije došlo do promene u parametrima protiv alternativne hipoteze da postoji promena parametara, tj.

$$H_0: \alpha_1 = \alpha_2,\ \beta_1 = \beta_2,\ \sigma_1 = \sigma_2$$

protiv

$$H_1: \alpha_1 \neq \alpha_2,\ \beta_1 \neq \beta_2\ ili\ \sigma_1 \neq \sigma_2.$$

Slučajne greške ε_i su nezavisne i normalno raspodeljene slučajne promenljive sa $\mathcal{N}(0, {\sigma_1}^2)$ za $i \leq k$ i $\mathcal{N}(0, {\sigma_2}^2)$ za $i > k$.

Kvant pretpostavlja da $\lambda(k)$ ima χ_4^2 raspodelu pod hipotezom da nema promene u parametrima (H_0), za svako k između 2 i $n-2$. Međutim, mnogi naučnici nisu mogli da se usaglase koja je zapravo asimptotska raspodela u pitanju, neki od njih, recimo Hinkley je tvrdio da ima χ_1^2, dok je Feder smatrao da uopšte nije u pitanju χ^2 raspodela [15].

4.2 Švarcov informacioni kriterijum

Švarcov informacioni kriterijum je nedavno predložen da se koristi kod detektovanja tačke promene u regresionim modelima [5].

Posmatramo niz podataka $(\boldsymbol{x}_1^T, Y_1), (\boldsymbol{x}_2^T, Y_2), \dots, (\boldsymbol{x}_n^T, Y_n)$. Cilj je da se testira hipoteza oblika:

$$H_0 : Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \dots, n \tag{4.2.1}$$

tj. regresioni koeficijenti se ne menjaju, protiv alternativne hipoteze

$$H_1 : Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + \varepsilon_i, \qquad i = 1, \dots, k \tag{4.2.2}$$

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_2 + \varepsilon_i, \qquad i = k+1, \dots, n,$$

gde $\boldsymbol{\beta}_1 = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$, $\boldsymbol{\beta}_2 = (\beta_0^*, \beta_1^*, \dots, \beta_{p-1}^*)^T$, to znači da postoji promena (u regresionim koeficijentima) u nepoznatom položaju k, označenom kao trenutak tačke promene.

Posmatrajmo linearni regresioni model

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \dots, n,$$

gde x_i, $i = 1, \dots, n$, odgovara i-toj komponenti nezavisne promenljive X, $\boldsymbol{X}$ matrica dimenzije $n \times p$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$ je vektor nepoznatih parametara, i ε_i označava slučajne greške. Pretpostavljamo da su ε_i nezavisne slučajne promenljive, svaka od njih ima normalnu raspodelu $\mathcal{N}(0, \sigma^2)$, gde je σ^2 nepoznati parametar ($\sigma^2 > 0$). U ovom slučaju, imamo da su zavisne promenljive, Y_i, $i = 1, \dots, n$, međusobno nezavisne slučajne promenljive sa normalnom raspodelom $\mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2)$.

Zbog alternativne hipoteze uvodimo sledeće oznake, pri čemu $k = p, \dots, n-p$, gde je p najmanji broj posmatranja za svaki segment, a n je veličina uzorka i zadovoljavaju uslov $p \leq \left\lfloor \frac{n}{2} \right\rfloor$,

$$\boldsymbol{Y}_1 = (Y_1, Y_2, \dots, Y_k)^T, \qquad \boldsymbol{Y}_2 = (Y_{k+1}, Y_{k+2}, \dots, Y_n)^T,$$

$$\boldsymbol{X}_1 = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_k^T \end{pmatrix}, \qquad \boldsymbol{X}_2 = \begin{pmatrix} \boldsymbol{x}_{k+1}^T \\ \boldsymbol{x}_{k+2}^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}.$$

Za testiranje izbora modela koristimo Švarcov informacioni kriterijum (*Schwarz Information Criterion*), u oznaci SIC, definisan sa:

$$SIC = -2L(\widehat{\boldsymbol{\theta}}) + s \log n,$$

gde je $L(\widehat{\boldsymbol{\theta}})$ maximum logaritma funkcije verodostojnosti, $\widehat{\boldsymbol{\theta}}$ je vektor ocenjenih parametara, s je broj parametara u modelu i n predstavlja veličinu uzorka. Maksimiziranje logaritma funkcije verodostojnosti je ekvivalentno minimiziranju Švarcovog informacionog kriterijuma (SIC).

Pod hipotezom H_0, postoji model takav da nema promena u regresionim koeficijentima, sa duge strane, pod hipotezom H_1 postoji grupa modela sa tačkom promene na poziciji p ili $p+1$ ili ... ili $n-p$. Dakle, cilj je izabrati model iz grupe modela.

Metodom maksimalne verodostojnosti ocene za $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$, pod hipotezom H_0, su date sa:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \qquad \hat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right).$$

Švarcov informacioni kriterijum pod hipotezom H_0, označen sa $SIC(n)$, je dat sa:

$$SIC(n) = -2L_0(\widehat{\boldsymbol{\theta}}) + (p+1)\log n$$

$$= n \log Q(\widehat{\boldsymbol{\beta}}) + n(\log 2\pi + 1) + (p + 1 - n)\log n,$$

gde $L_0(\widehat{\boldsymbol{\theta}})$ odgovara maximumu logaritma funkcije verodostojnosti pod hipotezom H_0 i $Q(\widehat{\boldsymbol{\beta}}) = \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$.

Sada posmatramo model pod alternativnom hipotezom H_1, tj. model sa tačkom promene u trenutku k, gde je $k = p, \ldots, n-p$. U ovom slučaju, $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma^2)^T$, i ocene dobijene metodom maksimalne verodostojnosti su:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} = (\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{1}})^{-1}\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{T}}\boldsymbol{Y}_{\boldsymbol{1}}, \qquad \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}} = (\boldsymbol{X}_{\boldsymbol{2}}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{2}})^{-1}\boldsymbol{X}_{\boldsymbol{2}}^{\boldsymbol{T}}\boldsymbol{Y}_{\boldsymbol{2}},$$

$$\hat{\sigma}^2 = \frac{1}{n}\left[\left(\boldsymbol{Y}_{\boldsymbol{1}} - \boldsymbol{X}_{\boldsymbol{1}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\right)^T\left(\boldsymbol{Y}_{\boldsymbol{1}} - \boldsymbol{X}_{\boldsymbol{1}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\right) + \left(\boldsymbol{Y}_{\boldsymbol{2}} - \boldsymbol{X}_{\boldsymbol{2}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}\right)^T\left(\boldsymbol{Y}_{\boldsymbol{2}} - \boldsymbol{X}_{\boldsymbol{2}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}\right)\right].$$

Tada je Švarcov informacioni kriterijum pod hipotezom H_1, označen sa $SIC(k)$, za $k = p, \dots, n-p$:

$$SIC(k) = -2L_k\left(\widehat{\boldsymbol{\theta}}\right) + (2p+1)\log n$$

$$= n\log\left[Q\left(\widehat{\boldsymbol{\beta}}_{\mathbf{1}}\right) + Q\left(\widehat{\boldsymbol{\beta}}_{\mathbf{2}}\right)\right] + n(\log 2\pi + 1) + (2p + 1 - n)\log n,$$

gde je $L_k\left(\widehat{\boldsymbol{\theta}}\right)$ maximum logaritma funkcije verodostojnosti pod hipotezom H_1.

Kriterijum izbora je izabrati model sa tačkom promene u trenutku k, ako za neko k važi:

$$SIC(n) > SIC(k).$$

Kada se odbacuje nulta hipoteza (H_0), ocena za trenutak promene u regresionim koeficijentima, dobijena metodom maksimalne verodostojnosti, označena sa $\hat{k}$, mora da zadovoljava:

$$SIC\left(\hat{k}\right) = \min\{SIC(k): p \le k \le n-p\},$$

$$= \max\{L_k(\theta): p \le k \le n-p\}.$$

4.3 EL (empirical likelihood) metod

U poslednjih nekoliko godina predmet interesovanja je detektovanje tačke promene kod modela segmentirane regresije. Glavni problem je kako otkriti tačku promene. Na primer, u ekonometriji je važan i još uvek težak problem odrediti što je ranije moguće polazne i krajnje vrednosti za tačku promene u sumnjivom delu posmatranja, u segmentiranom linearnom regresionom modelu. Postojeće procedure za detektovanje tačke promene su uglavnom izgrađene pod pretpostavkom homoskedastičnosti u parametarskim modelima ili preko klasične rang-test statistike kod neparametarskih modela. Zaključci na osnovu ovih postupaka su ponekad poništeni zbog heteroskedastičnosti. Pored toga, u postojećim postupcima je problem ako se vrednosti nezavisne promenljive X ne koriste na efikasan način da se konstruiše postupak za detektovanje tačke promene. U ovom poglavlju predstavićemo novi empirijski pristup verovatnoće (*eng. empirical likelihood*), u oznaci EL, za rešavanje ovih problema [14]. Ovaj metod je poboljšanje metoda segmentirane regresije, koji predlaže Kvant [15].

EL metod je neparametarska metoda za zaključivanje o funkcionalnim karakteristikama populacije, kao što su sredine i medijane. Jedna od najprimamljivijh osobina EL metoda je da ima svojsvo velikog uzorkovanja slično kao parametarska metoda maksimalne verodostojnosti.

Posmatramo sledeći model segmentirane regresije (sa jednom tačkom promene), definisan na sledeći način:

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i, & i = 1, \dots, k \\ \alpha_2 + \beta_2 x_i + \varepsilon_i, & i = k+1, \dots n \end{cases}. \tag{4.3.1}$$

U segmentiranom linearnom regresionom modelu (4.3.1) sa slučajnim greškama $\varepsilon_1, \dots, \varepsilon_n$ koje su identično i nezavisno raspodeljene sa očekivanjem nula i standardnom devijacijom σ_1, ako je $i \leq k$ i σ_2, ako je $i > k$, Dong predlaže EL-tip Wald-ove statistike za detektovanje trenutka promene k.

Neka je $\boldsymbol{y} = (y_1, \dots, y_n)^T$, $\boldsymbol{\theta}_1 = (\alpha_1, \beta_1)^T$, $\boldsymbol{\theta}_2 = (\alpha_2, \beta_2)^T$ i $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ i neka je

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta_1}^T \\ \boldsymbol{\theta_2}^T \end{pmatrix}, \qquad \boldsymbol{X}_k = \begin{pmatrix} X_{1k} & 0 \\ 0 & X_{2k} \end{pmatrix},$$

gde je

$$X_{1k} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_k \end{pmatrix}_{k\times 2} \quad \text{i} \quad X_{2k} = \begin{pmatrix} 1 & x_{k+1} \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{(n-k)\times 2}.$$

Tada se model (4.3.1) može predstaviti u matričnoj notaciji na sledeći način:

$$\boldsymbol{y} = \boldsymbol{X}_k\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{4.3.2}$$

sa $E(\boldsymbol{\varepsilon}) = 0$ i $Var(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, gde I_n jedinična matrica dimenzije $n \times n$.

Neka je

$$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1^T, \widehat{\boldsymbol{\theta}}_2^T)^T = \left(\boldsymbol{X}_k{}^T\boldsymbol{X}_k\right)^{-1}\boldsymbol{X}_k\boldsymbol{y}$$

ocena za $\boldsymbol{\theta}$ dobijena metodom najmanjih kvadrata. Dongov EL test baziran na Wald-ovoj test statistici je:

$$W = \left(\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_2\right)^T \left[\hat{\sigma}_1\left(\boldsymbol{X}_{1k}{}^T\boldsymbol{X}_{1k}\right)^{-1} + \hat{\sigma}_2\left(\boldsymbol{X}_{2k}{}^T\boldsymbol{X}_{2k}\right)^{-1}\right]^{-1}\left(\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_2\right), \tag{4.3.3}$$

gde je $\hat{\sigma}_i$ EL ocenjivač (ocena dobijena pomoću EL metoda) za standardnu devijaciju za i-ti segment regresionog modela, i =1, 2.

Nedavno, Liu i Qian predložili su interesantnu i računski jednostavnu proceduru za detektovanje tačke promene na osnovu EL odnosa. Ponovo pretpostavimo da su slučajne greške $\varepsilon_1, \dots, \varepsilon_n$ identično i nezavisno raspodeljene sa očekivanjem nula i standardnom devijacijom σ_1, ako je $i \le k$ i σ_2, ako je $i > k$. Za dato k, neka je

$$e_i = y_i - \begin{cases} \hat{\alpha}_1 + \hat{\beta}_1 x_i, & i = 1, \dots, k \\ \hat{\alpha}_2 + \hat{\beta}_2 x_i, & i = k+1, \dots, n \end{cases} \tag{4.3.4}$$

i

$$R(k) = sup\left\{\prod_{i=1}^{n} n\omega_i \,|\, \sum_{i=1}^{n} \omega_i e_i = 0, \omega_i \ge 0, \sum_{i=1}^{n} \omega_i = 1\right\}.$$

Definišemo test statistiku:

$$M^* = \max_{3<k<n-3}\{-2\log R(k)\}$$

i nultu hipotezu $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ odbacujemo kada je M^* značajno veliko.

Izmenjena test statistika sa $3 \leq t_1 < t_2 \leq n-3$ je

$$M = \max_{t_1 \leq k \leq t_2} \{-2 \log R(k)\}.$$

Motivacija koja dovodi do testiranja M^* jeste da je $E(e_i) = 0$ ako i samo ako je $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, tako da se očekuje da klasična dvostrana test statistika $-2 \log R(k)$ ima male vrednosti za svako k i stoga se očekuje da vrednosti M^* i M budu male.

Sada ćemo predstaviti postupak EL metoda.

Neka uslovna raspodela od y_i za dato x_i prati model (4.3.1), za $i = 1, \ldots, n$. Želimo da testiramo nultu hipotezu $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ protiv alternativne $H_1: \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, za neko $k \leq n$. Za pretpostavljeni trenutak promene k, neka su verovatnoće $p_i = P\{\varepsilon_i | x_i\}$ i $q_j = P\{\varepsilon_j | x_j\}$ za $i = 1, \ldots, k$ i $j = k+1, \ldots, n$ takve da zadovoljavaju

$$p_i > 0, \qquad q_j > 0, \qquad \sum_{i=1}^{k} p_i = 1, \qquad \sum_{j=k+1}^{n} q_j = 1.$$

Funkcija verodostojnosti je tada data sa

$$\ell(p, q|k) = \prod_{i=1}^{k} p_i \prod_{j=k+1}^{n} q_j.$$

Za dato k, funkcija verodostojnosti $\ell(p, q|k)$ dostiže svoj maksimum samo pod uslovima

$$p_i = \frac{1}{k} \quad \text{i} \quad q_j = \frac{1}{n-k}, \tag{4.3.5}$$

pa empirijski logaritam funkcije verodostojnosti postaje

$$r(p, q|k) = \prod_{i=1}^{k} k p_i \prod_{j=k+1}^{n} (n-k) q_j.$$

Kada je ε_i greška u merenju i η_i nespecifična greška kod x_i, sledi da je $E(x_i\eta_i) = 0$. Pošto je e_i ocena za η_i, razuman način da se maksimizira $r(p, q|k)$ je pod uslovima

$$p_i e_i(1, x_i)^T = 0 \quad \text{i} \quad q_j e_j(1, x_j)^T = 0,$$

uz ograničenja (4.3.5). Stoga, za dato k, $R(k)$ se definiše kao

$$\tilde{R}(k) = sup\left\{ r(p, q|k) | \sum_{i=1}^{k} p_i e_i(1, x_i)^T = 0, p_i \geq 0, \sum_{i=1}^{k} p_i = 1, \right.$$
$$\left. \sum_{j=k+1}^{n} q_j e_j\left(1, x_j\right)^T = 0, q_j \geq 0, \sum_{j=k+1}^{n} q_j = 1 \right\},$$

i test statistika je tada

$$\widetilde{M} = \max_{t_1 \leq k \leq t_2} \{-2 \log \tilde{R}(k)\}.$$

Napomenućemo da je računski zgodno izračunati $\log \tilde{R}(k)$ u dva koraka, gde se svaki od njih vrši u istom algoritmu. Neka

$$Q_1(k) = -2sup\left\{ \sum_{i=1}^{k} \log k p_i \,|\, \sum_{i=1}^{k} p_i e_i(1, x_i)^T = 0, p_i \geq 0, \sum_{i=1}^{k} p_i = 1 \right\},$$

i

$$Q_2(k) = -2sup\left\{ \sum_{j=k+1}^{n} \log(n-k) q_j \,|\, \sum_{j=k+1}^{n} q_j e_j\left(1, x_j\right)^T = 0, q_j \geq 0, \sum_{j=k+1}^{n} q_j = 1 \right\}.$$

Tada

$$\widetilde{M} = \max_{t_1 \leq k \leq t_2} \{Q_1(k) + Q_2(k)\}.$$

Dakle, tačka promene postoji kada je $\widetilde{M}$ veliko i tačka promene je onda ocenjena sa $\hat{k}$ tako da

$$\tilde{R}\left(\hat{k}\right) = \widetilde{M}.$$

5 Asimptotsko ponašanje

Ovo poglavlje se bavi teorijom o asimptotskoj raspodeli za ocene parametara u modelu segmentirane regresije, gde je svaki segment u formi linearnog modela. Prvo ćemo razmotriti konzistentnost ocena parametara, i potom ćemo proučiti asimtotsku raspodelu za te ocene, sa ograničenjem neprekidnosti u tačkama promene [17].

Posmatramo model segmentirane regresije sa r segmenata, gde je regresiona funkcija segmentiranog modela sledećeg oblika:

$$E(Y|x) = \mu(\boldsymbol{\theta}, x) = \begin{cases} f_1(\beta_1, x), & \xi_0 \le x \le \xi_1 \\ f_2(\beta_2, x), & \xi_1 < x \le \xi_2 \\ \vdots & \vdots \\ f_r(\beta_r, x), & \xi_{r-1} < x \le \xi_r. \end{cases}$$

Ovo može biti kraće zapisano kao:

$$\mu(\boldsymbol{\theta}, x) = \sum_{j=1}^{r} f_j(\beta_j, x) I_j(x),$$

gde je $I_j(x)$ indikator funkcija za interval $[\xi_{j-1}, \xi_j)$. Pretpostavljamo da je $\mu(\boldsymbol{\theta}, x)$ neprekidna u $x = \xi_j,\ j = 1, \dots, r-1$. U ovom modelu su ξ_0 i ξ_r poznate konstante, a $\beta_j,\ j = 1, \dots, r$ i $\xi_j, j = 1, \dots, r-1$ su nepoznati parametri. Bez gubljenja opštosti, pretpostavićemo da je $\xi_0 = 0$ i $\xi_r = 1$.

Neka je $\boldsymbol{\beta} = (\beta_1, \dots, \beta_r)^T$, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{r-1})^T$ i $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\xi})^T$. Neka $\mathcal{B}$ označava skup dopustivih vektora $\boldsymbol{\beta}$. To jest, $\mathcal{B}$ je kolekcija parametara $\boldsymbol{\beta}$ koji određuju funkciju $\mu(\boldsymbol{\theta}, x)$, zadovoljavajući uslov neprekidnosti. Za svako $\boldsymbol{\beta} \in \mathcal{B}$ posmatramo skup tačaka promene $\boldsymbol{\xi}$ (koje zavise od $\boldsymbol{\beta}$), koje određuju funkciju $\mu(\boldsymbol{\theta}, x)$, zadovoljavajući uslov neprekidnosti. Formiramo vektor $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\xi}(\boldsymbol{\beta}))^T \equiv (\boldsymbol{\beta}, \boldsymbol{\xi})^T$. Neka Θ označava skup takvih tačaka $\boldsymbol{\xi}$ i neka je $U = \{\mu(\boldsymbol{\theta}, x): \boldsymbol{\theta} \in \Theta\}$. Tokom ovog poglavlja uzimaćemo u obzir samo $\boldsymbol{\beta}$ iz $\mathcal{B}$ i $\boldsymbol{\theta}$ iz Θ.

Neka je $\varphi \equiv (\boldsymbol{\theta}, \sigma^2)$ i neka $\varphi^{(0)} = \left(\boldsymbol{\beta}_1^{(0)}, \dots, \boldsymbol{\beta}_r^{(0)}, \xi_1^{(0)}, \dots, \xi_{r-1}^{(0)}, \sigma_0^2\right)$ označava stvarne vrednosti parametara.

Pokazaćemo da je ocena $\hat{\varphi}$ konzistentna pod odgovarajućim pretpostavkama. Posmatraćemo asimptotsko ponašanje za $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})^T$ kada je funkcija $\mu(\boldsymbol{\theta}, \boldsymbol{x}) = E(Y|\boldsymbol{x})$ neprekidna u svakoj tački promene.

Za dato n, pretpostavljamo da n posmatranja Y_1, ..., Y_n su takvi da

$$Y_i = \mu(\boldsymbol{\theta}, x_i) + \varepsilon_i = \sum_{j=1}^{r} \boldsymbol{\beta}_j^T x_i \, \boldsymbol{I}_j(x_i) + \varepsilon_i, \qquad i = 1, \dots, n,$$

gde su ε_i nezavisne i identično raspoređene slučajne promenljive sa očekivanjem nula i disperzijom σ^2. Neka je $\widehat{\boldsymbol{\theta}}$ ocena za $\boldsymbol{\theta}$ dobijena metodom najmanjih kvadrata, koja minimizira

$$\sum_{i=1}^{n} \big(y_i - \mu(\boldsymbol{\theta}, x_i)\big)^2,$$

i

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(y_i - \mu\left(\widehat{\boldsymbol{\theta}}, x_i\right)\right)^2}{n - q},$$

gde je q broj nepoznatih parametara.

Uvodimo sledeću definiciju, koja nam je potrebna tokom ovog poglavlja.

Definicija 5.1. ($\boldsymbol{O_p}$ i $\boldsymbol{o_p}$): Za niz slučajnih promenljivih $\{Y_n\}$ se kaže da konvergira sa redom

(i) $O_p(1)$ ako za svako $\varepsilon > 0$ postoje konstante $D(\varepsilon)$ i $N(\varepsilon)$ tako da $n > N(\varepsilon)$ implicira da je

$$P\big(|Y_n| < D(\varepsilon)\big) \geq 1 - \varepsilon,$$

(ii) $o_p(1)$ ako za svako $\varepsilon > 0$, $\delta > 0$ postoji konstanta $N(\varepsilon, \delta)$ tako da $n > N(\varepsilon, \delta)$ implicira da je

$$P(|Y_n| < \delta) \geq 1 - \varepsilon.$$

(iii) Za niz slučajnih promenljivih $\{Y_n\}$ se kaže da konvergira sa redom $O_p(r_n)$ $\big(o_p(r_n)\big)$ ako niz $\{Y_n/r_n\}$ konvergira sa redom $O_p(1)$ $\big(o_p(1)\big)$.

Uvešćemo nekoliko oznaka radi lakšeg snalaženja, koje ćemo u nastavku često upotrebljavati.

$\boldsymbol{\mu} = \mu(\boldsymbol{\theta}, \boldsymbol{x}) \equiv \big(\mu(\boldsymbol{\theta}, x_1), \mu(\boldsymbol{\theta}, x_2), \dots, \mu(\boldsymbol{\theta}, x_k)\big)$;

$\boldsymbol{\mu}^{(0)} = \mu\big(\boldsymbol{\theta}^{(0)}, \boldsymbol{x}\big)$; $\quad \widehat{\boldsymbol{\mu}} = \big(\widehat{\boldsymbol{\theta}}, \boldsymbol{x}\big)$;

$\nu \equiv \nu(\boldsymbol{\theta}, \boldsymbol{x}) \equiv \mu(\boldsymbol{\theta}, \boldsymbol{x}) - \mu\big(\boldsymbol{\theta}^{(0)}, \boldsymbol{x}\big) \equiv \mu(\boldsymbol{t}) - \mu_0(\boldsymbol{t}) \equiv \mu - \mu_0$;

$\nu_i = \nu(\boldsymbol{\theta}, x_i)$; $\quad \hat{\nu}_i = \nu\big(\widehat{\boldsymbol{\theta}}, x_i\big)$.

Dakle, sada ćemo prvo razmotriti konzistentnost $\widehat{\boldsymbol{\theta}}$, a potom ćemo proučiti asimtotsku raspodelu za $\widehat{\boldsymbol{\theta}}$, sa ograničenjem neprekidnosti u tačkama promene, $\boldsymbol{\xi}$.

5.1 Konzistentnost i red konvergencije

Razmatraćemo pitanje konzistentnosti i reda konvergencije $\widehat{\boldsymbol{\theta}}$ ka $\boldsymbol{\theta}^{(0)}$**.** Na početku, uvešćemo pojam *identifikovanja* regresione funkcije. To jest, pretpostavljajući da nema grešaka posmatranja, za koje vrednosti $\boldsymbol{x}$ se može posmatrati $\mu(\boldsymbol{\theta}, \boldsymbol{x})$ u cilju jedinstvenog određivanja te funkcije na celom intervalu [0,1].

Pokazaćemo da pod odgovarajućim pretpostavkama $\widehat{\boldsymbol{\beta}}$ konvergira ka $\boldsymbol{\beta}^{(0)}$ sa redom $O_p\left(n^{-\frac{1}{2}}(\log\log n)^{\frac{1}{2}}\right)$ i ξ_j konvergira ka $\xi_j^{(0)}$ sa redom koji je određen brojem izvoda po x, u kojim se funkcije $f_j\left(\beta_j^{(0)}, x\right)$ i $f_{j+1}\left(\beta_{j+1}^{(0)}, x\right)$ poklapaju u $x = \xi_j^{(0)}$, za $j = 1, 2, \dots, r-1$.

Pretpostavićemo da se za $m_j - 1$ izvoda po x funkcije $f_j\left(\beta_j^{(0)}, x\right)$ i $f_{j+1}\left(\beta_{j+1}^{(0)}, x\right)$ poklapaju u $x = \xi_j^{(0)}$, ali da se razlikuju u m_j-tom izvodu. Dalje, to će značiti da funkcije f_j i f_{j+1} imaju neprekidan levi i desni m_j-ti izvod u $x = \xi_j^{(0)}$, $j = 1, 2, \dots, r-1$.

Definicija 5.1.1. Parametar $\boldsymbol{\beta}$ je *identifikovan* u $\boldsymbol{\mu}^{(0)}$ po vektoru $\boldsymbol{x} = (x_1, \dots, x_k)$ ako sistem od k jednačina $\mu(\boldsymbol{\theta}, \boldsymbol{x}) = \boldsymbol{\mu}^{(0)}$ istovremeno određuje $\boldsymbol{\beta}^{(0)}$.

Lema 5.1.2. Ako je $\boldsymbol{\beta}$ identifikovano u $\boldsymbol{\mu}^{(0)}$ po $\boldsymbol{x}$, tada postoje okoline N i T, gde je N (k-dimenzionalna) okolina od $\boldsymbol{\mu}^{(0)}$ i T je (k-dimenzionalna) okolina od $\boldsymbol{x}$, takve da važi:

(i) za sve (k-dimenzionalne) vektore $\boldsymbol{\mu} \in N$ i $\boldsymbol{x}^* \in T$, takve da $\boldsymbol{\mu}$ može biti predstavljeno kao $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{x}^*)$ za neko $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\beta}$ je identifikovano u $\boldsymbol{\mu}^{(0)}$ po $\boldsymbol{x}^*$;

(ii) postoji konstanta C takva da transformacija $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\mu}, \boldsymbol{x}^*)$ zadovoljava Lipšicov uslov $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| \leq C\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ svaki put kad je $\boldsymbol{x}^* \in T$ i funkcije $\boldsymbol{\mu_1} \equiv \boldsymbol{\mu}(\boldsymbol{\theta_1}, \boldsymbol{x}^*)$, $\boldsymbol{\mu_2} \equiv \boldsymbol{\mu}(\boldsymbol{\theta_2}, \boldsymbol{x}^*)$ su obe u N.

Neka je $H_n(s)$ raspodela za promenljivu $x_i,\ i = 1, \dots, n$. Neka je $H_n(s_2) - H_n(s_1) = n^{-1}\{\text{broj posmatranja u } (s_1, s_2]\}$. Pretpostavimo da su $x_i,\ i = 1, \dots, n,$ izabrani tako da zadovoljavaju sledeću hipotezu:

Hipoteza. $H_n(s)$ konvergira u raspodeli ka $H(s)$, što zapisujemo $H_n(s) \xrightarrow{r} H(s)$, gde je $H(s)$ funkcija raspodele sa $H(0) = 0$, $H(1) = 1$.

Definicija 5.1.3. *Centar posmatranja* je tačka rasta funkcije H.

Lema 5.1.4. Pretpostavimo da postoji $\varepsilon > 0$ takvo da za svako $K > 0$ postoje $d(K)$, $n(K)$ takvi da je $d > d(K)$, $n > n(K)$ i sledi da je

$$\inf_{\{\boldsymbol{\beta}:\ \|\boldsymbol{\beta}-\boldsymbol{\beta}^{(0)}\|>d\}} H_n\{\boldsymbol{x}:\ |\mu(\boldsymbol{\theta}, \boldsymbol{x}) - \mu(\boldsymbol{\theta}^{(0)}, \boldsymbol{x})| > K\} > \varepsilon. \qquad (*)$$

Tada postoji d^* tako da je

$$\lim_{n\to\infty} P\left(\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}\right\| \le d^*\right) = 1.$$

Lema 5.1.5. Ako je x_0 centar posmatranja, $\delta > 0$, $\eta > 0$ i uslov $(*)$ iz Leme 5.1.4. zadovoljen, tada

$$P\{|\mu(\widehat{\boldsymbol{\theta}}, \boldsymbol{x}) - \mu(\boldsymbol{\theta}^{(0)}, \boldsymbol{x})| \ge \eta \text{ za svako } \boldsymbol{x} \text{ tako da } |\boldsymbol{x} - x_0| \le \delta\ \} \to 0.$$

Lema 5.1.6. Pretpostavimo da za svako $K > 0$, takvo da $Kp(n) < n$, za dovoljno veliko n, važi

$$\lim_{n\to\infty} \sum_{i=1}^{Kp(n)} \left(\frac{\vartheta_i^2 \log\log \sum_{j=1}^{i} \vartheta_j^2}{h(i) \sum_{j=1}^{i} \vartheta_j^2} \right)^{1+\delta} < \infty.$$

Tada

$$T_N \equiv \frac{\sum_{i=1}^{N} \vartheta_i e_i}{(\sum_{i=1}^{N} \vartheta_i^2)^{\frac{1}{2}}} = O_p\left((\log\log n)^{\frac{1}{2}}\right), \qquad \text{kada } n \to \infty.$$

Tvrđenje 5.1.7. (konzistentnost): Ako važi

(i) uslov $(*)$ iz Leme 5.1.4. :

$$\inf_{\{\boldsymbol{\beta}:\, \|\boldsymbol{\beta}-\boldsymbol{\beta}^{(0)}\|>d\}} H_n\{\boldsymbol{x}:\, |\mu(\boldsymbol{\theta},\boldsymbol{x}) - \mu(\boldsymbol{\theta}^{(0)},\boldsymbol{x})| > K\} > \varepsilon,$$

(ii) $\boldsymbol{\beta}$ je identifikovano u $\boldsymbol{\mu}^{(0)}$ po $\boldsymbol{x}$,

(iii) komponente $\boldsymbol{x}$ su centri posmatranja,

tada

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)} = o_p(1) \tag{5.1.1}$$

$$\hat{\sigma}^2 - \sigma_0^2 = o_p(1). \tag{5.1.2}$$

Dokaz.

Neka su N, T (k-dimenzionalne) okoline $\boldsymbol{\mu}^{(0)}$ i neka $\boldsymbol{x}$ zadovoljava uslove iz Leme 5.1.2.

Iz Leme 5.1.5. sledi da za dato $\varepsilon > 0$, kada $n \to \infty$, postoji $\boldsymbol{x}^* \in T$ tako da $\boldsymbol{\mu} = \mu(\widehat{\boldsymbol{\theta}}, \boldsymbol{x}^*) \in N$ i $\|\nu(\widehat{\boldsymbol{\theta}}, \boldsymbol{x})\| \leq \varepsilon$.

Iz Leme 5.1.2. , $\boldsymbol{\beta} = \boldsymbol{\beta}(\widehat{\boldsymbol{\mu}}, \boldsymbol{x}^*)$ je jedinstveno određeno i zadovoljava nejednakost:

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}\| \leq C\|\mu(\widehat{\boldsymbol{\theta}}, \boldsymbol{x}^*) - \mu(\boldsymbol{\theta}^{(0)}, \boldsymbol{x}^*)\| \leq C\varepsilon. \tag{5.1.3}$$

Pošto je ε proizvoljno, iz formule (5.1.3) sledi jednakost (5.1.1).

Jednakost (5.1.2) direktno sledi, jer je

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i - \hat{v}_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \sigma_0^2 + o_p(1).$$

Sa druge strane,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 + \frac{1}{n}\sum_{i=1}^{n} \hat{v}_i^2 - \frac{1}{n}\sum_{i=1}^{n} e_i \hat{v}_i.$$

Iz Leme 5.1.4. i uniformne konvergencije u verovatnoći od $\sum_{i=1}^{n} e_i v_i$ ka $\left\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\right\| \leq d^*$, sledi da je $\hat{\sigma}^2 \geq \sigma_0^2 + o_p(1)$. Dakle, $\hat{\sigma}^2 = \sigma_0^2 + o_p(1)$, što je i trebalo dokazati.

Tvrđenje 5.1.8. Pretpostavimo da je W podskup od [0,1] takav da je $H(W) > 0$. Tada je

$$\min_{x_i \in W} |\hat{v}(x_i)| = O_p\left(n^{-\frac{1}{2}} (\log\log n)^{\frac{1}{2}}\right). \tag{5.1.4}$$

Tvrđenje 5.1.9. (red konvergencije): Ako je

(i) $\boldsymbol{\beta}$ identifikovano u $\boldsymbol{\mu}^{(0)}$ po $\boldsymbol{x}$ i komponente $\boldsymbol{x}$ su centri posmatranja,

(ii) razmak posmatranja oko svake tačke promene je takav da su zadovoljeni uslovi Leme 5.1.6. tada

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)} = O_p\left(n^{-\frac{1}{2}} (\log\log n)^{\frac{1}{2}}\right).$$

Dokaz.

Iz Teoreme 5.1.8. sledi da u svakoj maloj okolini centra posmatranja postoji x_i tako da

$$\mu(x_i) - \mu_0(x_i) = O_p\left(n^{-\frac{1}{2}} (\log\log n)^{\frac{1}{2}}\right).$$

Iz Leme 5.1.2. sledi dokaz ovog tvrđenja.

Sada ćemo razmatrati brzinu kovergencije od $\hat{\xi}$ ka ξ. Pretpostavimo da $f_j\left(\beta_j^{(0)}, x\right)$ i $f_{j+1}\left(\beta_{j+1}^{(0)}, x\right)$ $m_j - 1$ izvoda po x u $\xi_j^{(0)}$, $j = 1, \dots, r-1$. Zatim pretpostavimo da funkcije f_j i f_{j+1} imaju neprekidan levi i desni m_j-ti izvod u $x = \xi_j^{(0)}$, $j = 1, 2, \dots, r-1$ i razlikuju se u oba ova izvoda. Označimo ove pretpostavke kraće kao *uslovi* (ξ).

Neka $D^-(h,j,k)$ i $D^+(h,j,k)$ označavaju k-ti levi i desni izvod po x, respektivno, za $f_h\left(\beta_h^{(0)},x\right)$ u $x=\xi_j^{(0)}$. Ako se oni poklapaju, onda ćemo zajedničku vrednost označiti sa $D(h,j,k)$.

Razvijamo $f_j(\beta_j,x)$ i $f_{j+1}(\beta_{j+1},x)$ u Tejlorov red oko $\beta_j^{(0)}$, $\xi_j^{(0)}$ i $\beta_{j+1}^{(0)}$, $\xi_j^{(0)}$, respektivno. Podsetimo da je $f_j\left(\beta_j^{(0)},\xi_j^{(0)}\right)=f_{j+1}\left(\beta_{j+1}^{(0)},\xi_j^{(0)}\right)$, $D(j,j,k)=D(j+1,j,k)$, $k=1,2,\dots,m_j-1$.

Za $\beta\in\mathcal{B}$ u okolini $\beta^{(0)}$, tačka promene ξ_j, za dva segmenta $f_j(\beta_j,x)$ i $f_{j+1}(\beta_{j+1},x)$, je dobijena rešavanjem jednačine:

$$f_{j+1}(\beta_{j+1},\xi_j)-f_j(\beta_j,\xi_j)=0.$$

Za β_j, β_{j+1}, ξ_j u okolini $\beta_j^{(0)}$, $\beta_{j+1}^{(0)}$, $\xi_j^{(0)}$,

$$\begin{aligned}0&=f_{j+1}(\beta_{j+1},\xi_j)-f_j(\beta_j,\xi_j)\\&=\left[\frac{\partial f_{j+1}^{(0)}}{\partial\beta_{j+1}}+o(1)\right]'\left(\beta_{j+1}-\beta_{j+1}^{(0)}\right)-\left[\frac{\partial f_j^{(0)}}{\partial\beta_j}+o(1)\right]'\left(\beta_j-\beta_j^{(0)}\right)\\&\quad+\frac{1}{m_j!}\left[D^{\pm}(j+1,j,m_j)-D^{\pm}(j,j,m_j)+o(1)\right]\left(\xi_j-\xi_j^{(0)}\right)^{m_j}\end{aligned}$$

gde $D^{\pm}$ je D^+ ako je $\xi_j>\xi_j^{(0)}$ i D^- ako je $\xi_j<\xi_j^{(0)}$. Tada

$$\begin{aligned}&\frac{1}{m_j!}\left[D^{\pm}(j+1,j,m_j)-D^{\pm}(j,j,m_j)+o(1)\right]\left(\xi_j-\xi_j^{(0)}\right)^{m_j}\\&=\left[\frac{\partial f_j^{(0)}}{\partial\beta_j}+o(1)\right]'\left(\beta_j-\beta_j^{(0)}\right)-\left[\frac{\partial f_{j+1}^{(0)}}{\partial\beta_{j+1}}+o(1)\right]'\left(\beta_{j+1}-\beta_{j+1}^{(0)}\right).\end{aligned}\qquad(5.1.5)$$

Iz jednačine (5.1.5) i Teoreme 5.1.9 sledi

$$\left(\hat{\xi}_j-\xi_j^{(0)}\right)^{m_j}=O_p\left(n^{-\frac{1}{2}}(\log\log n)^{\frac{1}{2}}\right),\qquad j=1,\dots,r-1.$$

Ovo je formalno navedeno u sledećoj teoremi.

Teorema 5.1.10. Ako je

(i) $\boldsymbol{\beta}$ dobro identifikovan u $\boldsymbol{\mu}^{(0)}$ po $\boldsymbol{x}$ i komponente $\boldsymbol{x}$ su centri posmatranja,

(ii) *uslovi* (ξ) su zadovoljeni,

(iii) rastojanje posmatranja oko svake tačke promene je takav da su zadovoljeni uslovi Leme 5.1.6. ,

tada

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)} = O_p\left(n^{-\frac{1}{2}}(\log\log n)^{\frac{1}{2}}\right)$$

i

$$\left(\hat{\xi}_j - \xi_j^{(0)}\right)^{m_j} = O_p\left(n^{-\frac{1}{2}}(\log\log n)^{\frac{1}{2}}\right), \quad j = 1, \dots, r-1.$$

Važan specijalan slučaj ove teoreme je sledeća posledica.

Posledica 5.1.11. Ako su zadovoljene pretpostavke Teoreme 5.1.10. i dodatno $m_1 = \cdots = m_{r-1} = 1$, tada

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)} = O_p\left(n^{-\frac{1}{2}}(\log\log n)^{\frac{1}{2}}\right).$$

Glavni rezultat ovog poglavlja je da je $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)} = O_p\left(n^{-\frac{1}{2}}(\log\log n)^{\frac{1}{2}}\right)$ ako postoji vektor $\boldsymbol{x}$ čije komponente su centri posmatranja i koji identifikuje $\boldsymbol{\beta}$ u $\boldsymbol{\mu}^{(0)}$.

Lema 5.1.5. implicira da $\mu(\boldsymbol{\theta}, \boldsymbol{x})$ mora biti blizu $\mu\left(\boldsymbol{\theta}^{(0)}, \boldsymbol{x}\right)$ za najmanje jednu vrednost $\boldsymbol{x}$ u blizini svakog centra posmatranja. Konzistentnost $\widehat{\boldsymbol{\beta}}$ je posledica toga.

Uslov $(*)$ iz Leme 5.1.4. obezbeđuje da se ocena $\widehat{\boldsymbol{\beta}}$ dobijena metodom najmanjih kvadrata nalazi u krugu sa centrom $\boldsymbol{\beta}^{(0)}$ i poluprečnikom d^*.

5.2 Asimptotska raspodela

Dosadašnja saznanja o konzistentnosti i redu konvergencije nam omogućavaju da sada razmatramo asimptotsku raspodelu za $\hat{\theta}$.

Pokazaćemo da

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)} = O_p\left(1/\sqrt{n}\right) \quad \text{i} \quad \left(\hat{\xi}_j - \xi_j^{(0)}\right)^{m_j} = O_p\left(1/\sqrt{n}\right).$$

Javljaju se različita asimptotska ponašanja u zavisnosti od toga da li je $\boldsymbol{\beta}^{(0)}$ unutrašnja ili rubna tačka skupa $\mathcal{B}$. Pokazaćemo u nastavku da ako je $\boldsymbol{\beta}^{(0)}$ unutrašnja tačka skupa $\mathcal{B}$ sledi da $\widehat{\boldsymbol{\beta}}$ ima asimptotski normalnu raspodelu [16].

Lema 5.2.1. Pretpostavimo da su jednaki sledeći izvodi: $D^+\left(j, j, m_j\right) = D^-\left(j, j, m_j\right)$ i $D^+\left(j+1, j, m_j\right) = D^-\left(j+1, j, m_j\right)$, $j = 1, \dots, r-1$. Ako su $m_1, \dots, m_{r-1}$ svi neparni, tada je $\boldsymbol{\beta}^{(0)}$ unutrašnja tačka od $\mathcal{B}$. Ako je bilo koji od m_j paran, tada je $\boldsymbol{\beta}^{(0)}$ granična (rubna) tačka skupa $\mathcal{B}$.

Važan specijalan slučaj ove Leme je kada su segmenti prave linije, odnosno kada su u formi linearnog modela, tada $m_1 = \cdots = m_{r-1} = 1$.

Teorema 5.2.2. U slučaju segmentirane regresije (kada je svaki segment prava linija), ako je $\boldsymbol{\beta}^{(0)}$ unutrašnja tačka od $\mathcal{B}$ tada $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}\right)$ konvergira u raspodeli ka $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{G}^{-1})$, tj.

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}\right) \xrightarrow{r} \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{G}^{-1}) \tag{5.2.1}$$

i

$$G_{jk} = \int_{\xi_0}^{\xi_r} \frac{\partial \mu\left(\boldsymbol{\theta}^{(0)}, \boldsymbol{x}\right)}{\partial \beta_j} \frac{\partial \mu\left(\boldsymbol{\theta}^{(0)}, \boldsymbol{x}\right)}{\partial \beta_k} dH(x)$$

gde je $\boldsymbol{G}$ matrica informacija dimenzije $q \times q$, gde je q broj nepoznatih parametara i $\boldsymbol{G}$ je strogo pozitivno definitna matrica.

Asimptotska raspodela od $\hat{\xi}_j$ zavisi od vrednosti m_j i od toga da li je $D_j^- = D_j^+$.

Teorema 5.2.3. Ako su $m_1, \dots, m_{r-1}$ svi neparni i $D_j^- = D_j^+ = D_j$, tada

$$\sqrt{n}\begin{bmatrix} \left(\hat{\xi}_1 - \xi_1^{(0)}\right)^{m_1} \\ \left(\hat{\xi}_2 - \xi_2^{(0)}\right)^{m_2} \\ \vdots \\ \left(\hat{\xi}_{r-1} - \xi_{r-1}^{(0)}\right)^{m_{r-1}} \end{bmatrix} \xrightarrow{r} \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{A}\, \boldsymbol{G}^{-1}\boldsymbol{A}^T) \tag{5.2.2}$$

gde je $\boldsymbol{A} \equiv \boldsymbol{A}(\boldsymbol{\beta}, \boldsymbol{\xi})$ matrica dimenzije $(r-1) \times q$. Ako je m_j, $j = 1, \dots r-1$ neparno, a $D_j^- \neq D_j^+$, tada asimptotska raspodela za $\sqrt{n}\left(\hat{\xi}_j - \xi_j^{(0)}\right)^{m_j}$ ne mora biti normalna.

Neka je funkcija f definisana na sledeći način:

$$f(x; \boldsymbol{\beta}, \boldsymbol{\xi}) = \begin{cases} \beta_1 \boldsymbol{g}_1(x), & \xi_0 \le x \le \xi_1 \\ \beta_2 \boldsymbol{g}_2(x), & \xi_1 < x \le \xi_2 \\ \vdots & \vdots \\ \beta_r \boldsymbol{g}_r(x), & \xi_{r-1} < x \le \xi_r. \end{cases} \tag{5.2.3}$$

Tada j-ti red matrice $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\beta}, \boldsymbol{\xi})$ je $\left(\boldsymbol{a}_{j1}^T, \boldsymbol{a}_{j2}^T, \dots, \boldsymbol{a}_{jr}^T\right)$, gde je $\boldsymbol{a}_{jk}$ vektor dat sa:

$$\boldsymbol{a}_{jk} = \begin{cases} c_j^{-1} \boldsymbol{g}_j(\xi_j), & k = j \\ -c_j^{-1} \boldsymbol{g}_{j+1}(\xi_j), & k = j+1 \\ \mathbf{0}, & \text{inače} \end{cases}$$

gde je

$$c_j = \frac{1}{m_j!}\left[\beta_{j+1} \boldsymbol{g}_{j+1}^{(m_j)}(\xi_j) - \beta_j \boldsymbol{g}_j^{(m_j)}(\xi_j)\right]$$

i

$$\boldsymbol{g}_j^{(m)}(x) = \frac{d^m \boldsymbol{g}_j(x)}{dx^m}.$$

Na osnovu (5.2.1) i (5.2.2) sledi da funkcija za svaki segment $\beta_j \boldsymbol{g}_j(x)$, $j = 1, \dots r$, ima m_j neprekidnih izvoda u ξ_j i m_{j-1} neprekidnih izvoda u ξ_{j-1}. Ovo nije realno ograničenje, ali funkcije koje mi najčešće koristimo za svaki segment, kao što su polinomi, su beskonačno diferencijabilne na intervalu $[\xi_0, \xi_r]$, odnosno na intervalu $[0,1]$.

Za uzorak $(x_1, y_1), \dots, (x_n, y_n)$, ocena za matricu $\boldsymbol{G}$ je:

$$\widehat{\boldsymbol{G}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \mu(\widehat{\boldsymbol{\theta}}, \boldsymbol{x})}{\partial \beta_j} \frac{\partial \mu(\widehat{\boldsymbol{\theta}}, \boldsymbol{x})}{\partial \beta_k}.$$

Koristeći (5.2.3), lako je pokazati da je $\widehat{\boldsymbol{G}}$ dijagonalna blok matrica sa r blokova, po jedan za svako β_j, $j = 1, \dots r$, a j-ti blok je matrica:

$$\widehat{\boldsymbol{G}}_j = \frac{1}{n} \sum_{i:\, x_i \in \hat{I}_j} \boldsymbol{g}_j(x_i) \boldsymbol{g}_j^T(x_i),$$

gde je

$$\hat{I}_j = \{x_i : \hat{\xi}_{j-1} < x_i \leq \hat{\xi}_j\}.$$

Tako, asimptotski, ako su svi m_j, $j = 1, \dots r-1$ neparni, možemo posmatrati parametre $\hat{\beta}_j, j = 1, \dots r$, kao nezavisne i normalno raspoređene, tj.

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 \left(\boldsymbol{X}_j^T \boldsymbol{X}_j\right)^{-1}\right) \tag{5.2.4}$$

gde

$$\boldsymbol{X}_j^T \boldsymbol{X}_j = \sum_{x_i \in \hat{I}_j} \boldsymbol{g}_j(x_i) \boldsymbol{g}_j^T(x_i).$$

Ako je svaki $m_j = 1$, $j = 1, \dots r-1$, parametar $\hat{\boldsymbol{\xi}}$ je asimptotski normalan sa očekivanjem $\boldsymbol{\xi}$ i varijansom ocenjenom varijansno-kovarijansnom matricom $\sigma^2 \widehat{\boldsymbol{A}} \left(n \widehat{\boldsymbol{G}}\right)^{-1} \widehat{\boldsymbol{A}}^T$, gde je $\widehat{\boldsymbol{A}} = \boldsymbol{A}(\widehat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$.

Budući da tokom rada posmatramo model segmentirane regresije sa dva segmenta i gde su segmenti u formi linearnog modela, sada ćemo prikazati asimptotske ocene za takav model.

Dakle, posmatramo sledeći model:

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i, & x_i \leq \xi \\ \alpha_2 + \beta_2 x_i + \varepsilon_i, & x_i > \xi \end{cases} \qquad i = 1, \dots, n.$$

Na osnovu (5.2.4), imamo sledeće asimptotsko ponašanje, za $d = 1,2$,

$$\hat{\alpha}_d \sim \mathcal{N}\left(\alpha_d, \frac{\sigma^2 \sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_d \sim \mathcal{N}\left(\beta_d, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$$

i

$$\operatorname{cov}(\hat{\alpha}_d, \hat{\beta}_d) = -\frac{\sigma^2 \bar{x}}{\sum_i (x_i - \bar{x})^2}.$$

Iz ograničenja neprekidnosti imamo da je

$$\hat{\xi} = \frac{\beta_1^{(0)} - \beta_2^{(0)}}{\beta_2^{(1)} - \beta_1^{(1)}}.$$

Kako je u ovom slučaju $\boldsymbol{g}_d(x) = (1, x)^T$ za $d = 1,2$, sledi da je

$$\boldsymbol{A} = \frac{1}{\beta_2^{(1)} - \beta_1^{(1)}}(1, \xi, -1, -\xi).$$

Stoga iz (5.2.2), jer je svako $m_j = 1$, $j = 1, \dots r - 1$, imamo da je $\hat{\xi}$ asimptotski normalno sa očekivanjem ξ i varijansom:

$$\operatorname{var}(\hat{\xi}) = \sigma^2 \boldsymbol{A} \begin{pmatrix} \boldsymbol{X}_1^T \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2^T \boldsymbol{X}_2 \end{pmatrix}^{-1} \boldsymbol{A}^T.$$

6 Uspešnost procenjivanja modela

6.1 Koeficijent determinacije

Važan faktor kao odlučujući kriterijum kod regresionih modela je koeficijent determinacije.

Kod linearne regresije koeficijent determinacije je jednak koeficijentu korelacije. Međutim, kod segmentirane regresije mogu biti različiti i koeficijent korelacije gubi deo svog značenja [7]. Ipak kod segmentirane regresije je potrebno proveriti da segmentacija ne daje manji koeficijent determinacije od koeficijenta korelacije (koeficijent determinacije treba da bude veći od koeficijenta korelacije - ovo inicira da segmentirana regresija bolje fituje podatke od linearne regresije).

Ukupno odstupanje jedne registrovane vrednosti promenljive Y_i od srednje vrednosti $\bar{Y}$ se može podeliti na: odstupanje objašnjeno modelom, $\hat{Y}_i - \bar{Y}$, i odstupanje registrovane vrednosti od vrednosti određene modelom (ocenjene vrednosti), $Y_i - \hat{Y}_i$. Ovo razlaganje važi i za kvadrate ovih odstupanja, odnosno važi

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

(SST)	(SSR)	(SSE)
ukupna suma kvadrata	suma kvadrata regresije	suma kvadrata greške

Rastavljanje varijacija promenljive Y iz uzorka dovodi do mere uspešnosti prilagođavanja, koja se naziva koeficijent determinacije i označava sa R^2. On zapravo predstavlja deo varijacija promenljive Y koje se mogu pripisati varijacijama promenljive X.

Koeficijent determinacije se definiše na sledeći način:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}.$$

Koeficijent determinacije predstavlja zapravo meru koliko se regresiona linija uzorka dobro prilagođava posmatranim podacima. R^2 ne može biti negativan ili veći od 1, tj.

$$0 \leq R^2 \leq 1.$$

Regresiona linija najbolje opisuje podatke kada je koeficijent determinacije jednak jedinici. Ukoliko je vrednost od R^2 blizu nule, to znači da se regresiona linija vrlo slabo prilagođava posmatranim podacima. Jedno od mogućih razloga za to je da varijacije promenljive X ne utiču na promenljivu Y, zatim je moguće da je uticaj nezavisne promenljive X slab u odnosu na uticaj slučajnog odstupanja ili je moguće da je regresioni model pogrešno postavljen.

6.2 Otkrivanje uticajnih podataka

Tokom istraživanja moguće je uočiti podatke koji dovode do različitih rezultata. Razlikujemo dve vrste „uticajnih podataka“, to su:

- autlajeri (*eng. outlier*)
- uticajna opažanja

Ukoliko se pojavi autlajer potrebno je prvo proveriti da li su podaci ispravno uneti. Relativno često se dešava da se pogreši prilikom unosa podataka.

Zatim potrebno je ispitati zašto se to desilo. Ponekad otkriće autlajera može biti od velikog značaja. Neka naučna otkrića potiču od pojave neočekivanih odstupanja ili nepravilnosti. Jedan primer važnosti autlajera je, u statističkoj analizi, transakcija kredinih kartica. Autlajer u ovom slučaju može predstavljati zloupotrebu kartice.

Treba isključiti taj podatak iz analize, ali ga ponovo uključiti ukoliko se model menja. Isključivanje jedne ili više tačaka može da dovede do različitih statističkih rezultata (da li je nešto statistički značajno ili ne) ili do neobjavljenih istraživanja. To može dovesti do teške odluke šta je razumno isključenje. Da bi se izbegla bilo kakva sugestija o neiskrenosti, uvek treba prijaviti postojanje autlajera, čak i ako ih ne uključujemo u konačan model.

Autlajere nije jednostavno ukloniti, jer oni leže izvan opsega drugih podataka, ali je važno znati kako ove tačke utiču na model i onda proceniti da li ih treba zadržati.

7 Primena segmentirane regresije

Segmentirana regresija je veoma često korišćen metod u medicini. Kada se analizira učestalost pojave raka i stope smrtnosti, zdravstveni i medicinski istraživači su posebno zainteresovani da znaju da li je bilo promena u trendu tokom vremena, i ako je došlo do promena kada se to desilo. Ovakva pitanja igraju važnu ulogu u merenju napretka u borbi protiv raka i uticaja intervencije na ishod bolesti. U statističkim terminima, promena u trendu se može definisati kao promena nagiba u regresiji. Moguće je identifikovati i faktore rizika za određeni karcinom, na osnovu kliničkih i demografskih promenljivih. Na primer, može se predvideti da li će pacijent, hospitalizovan zbog srčanog udara, imati drugi infarkt. Predviđanje se zasniva na demografiji, ishrani i kliničkim merenjima za tog pacijenta. Takođe, može se procenjivati nivo glukoze u krvi dijabetičara, na osnovu infracrvene apsorpcije spektra krvi te osobe.

Mnogi naučnici posmatraju model segmentirane regresije u različitim realnim situacijama. Na primer, u radu Yeh et al. se razmatra ideja „anaerobnog praga“. Pretpostavlja se da ako obim posla neke osobe dostigne određeni prag, gde mišići te osobe ne mogu dobiti dovoljno kiseonika, tada aerobni metabolički procesi postaju anaerobni metabolički procesi. Taj prag se naziva „anaerobni prag“. U ovom slučaju predložen je model sa dva segmenta. U radu McGee i Carleton (1970) se posmatra primer, gde struktura obima prodaje akcija na regionalnoj berzi, na kojoj učestvuju i njujorška i američka berza, zavisi od promene propisa vlade. Model sa četiri segmenta se smatra odgovarajućim u ovoj analizi. Primeri ove vrste u različitim kontekstima su prikazani u radovima Sprent (1961), Dunicz (1969), Schulze (1984) i mnogim drugim [11].

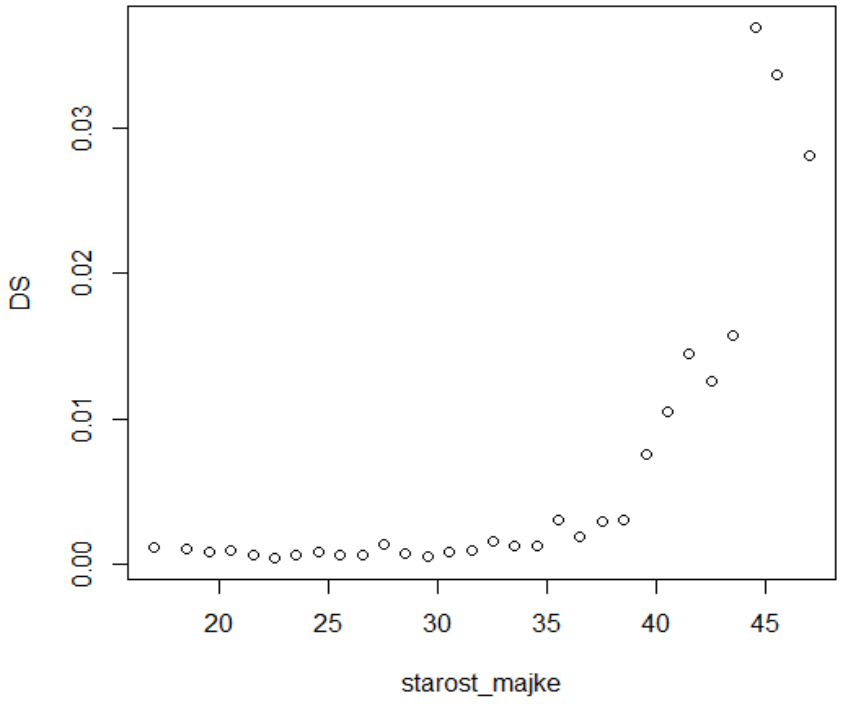
U nekim situacijama, iako se model segmentirane regresije smatra pogodnim, ne može se odrediti odgovarajući broj segmenata, kao što je navedeno za prethodne primere. Osim toga, u slučaju višestruke regresije, može da ne bude jasno koja nezavisna promenljiva se odnosi na promenu zavisne promenljive ili koja nezavisna promenljiva se može najbolje koristiti kao segmentirana promenljiva. U nekim problemima gde su nezavisne promenljive malih dimenzija, grafički prikaz može biti efikasan za određivanje broja segmenata i koje nezavisne promenljive je najbolje izabrati kao segmentirane promenljive. Međutim, ukoliko su nezavisne promenljive velikih dimenzija, međusobno nezavisne promenljive mogu biti u suprotnosti sa takvim pristupom.

7.1 Pojava Daunovog sindroma kod novorođenčadi

Sada ćemo detaljnije predstaviti jedan primer koristeći stvarne podatke i softverski program R. Primer je vezan za pojavu Daunovog sindroma kod novorođenčadi.

Daunov sindrom je genetski poremećaj izazvan dodatnim hromozomom 21 ili delom hromozoma 21 koji se translocira drugom hromozomu. Učestalost Daunovog sindroma veoma zavisi od starosti majke i naglo raste nakon 30. godine. 1960. godine je sprovedeno istraživanje uticaja starosti majke na učestalost Daunovog sindroma u Britanskoj Kolumbiji, jednoj od najgušće naseljenih kanadskih pokrajina, pod pokroviteljstvom registra za zdravstveni nadzor. Koristićemo podatke koji su prikupljeni u toj studiji. Majke su klasifikovane po starosti. Većina grupa odgovara starosti majke izražene u godinama, ali prva grupa obuhvata sve majke starosti 15 do 17 godina, a poslednju grupu čine majke starosti 46 do 49 godina. Nisu prikupljeni podaci za majke starije od 50 godina i mlađe od 15 godina.

Posmatrani podaci se mogu grafički prikazati preko dijagrama rasipanja (*Scatter plot*) na sledeći način:



Grafik 7.1.1: *Dijagram rasipanja koji pokazuje vezu između starosti majke i procenta beba rođenih sa Daunovim sindromom*

Kružići na grafiku prikazuju procenat beba rođenih sa Daunovim sindromom za različite godine starosti majki. Dobro je poznato da rizik od Daunovog sindroma raste sa majčinim godinama, ali važno je proceniti gde i kako se menja taj rizik u odnosu na starost majke. Postavljaju se sledeća pitanja, na koje je veoma važno dati odgovor:

(*i*) da li starost žene povećava rizik od Daunovog sindroma?

(*ii*) da li je rizik konstantan tokom celog perioda starosti?

(*iii*) ukoliko rizik zavisi od starosti, da li postoji prag vrednost?

U opštem slučaju, problem je proceniti segmentirani model, odnosno procena tačke promene i relevantne mere nesigurnosti svih parametara modela. Pre svega, neophodno je jasno segmentirati model.

Mi ćemo procenu modela vršiti u softverskom programu R. Prvo se procenjuje standardni linearni model i potom se dodaje segmentirani model, pa se ponovo procenjuje opšti (ukupan) model. Dakle, fituje se novi model uzimajući u obzir linearnu vezu po delovima.

Procenjujemo model segmentirane regresije sa jednom tačkom promene:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 (X_i - \xi) + \varepsilon_i, \quad i = 1, \ldots, 30,$$

gde su ε_i nezavisne i identično raspodeljene slučajne promenljive, sa očekivanjem nula i disperzijom σ^2.

U sledećoj tabeli prikazani su rezultati ocene tačke promene:

tačka promene	**ocenjena vrednost**	**standardna greška**
$\boldsymbol{\xi}$	38.2000	0.6867

Tabela 7.1.1: *Ocena tačke promene*

Zatim dati su rezultati ocene ostalih parametara:

parametri	**ocenjena vrednost**	**standardna greška**	***t*-vrednost**	***p*-vrednost**
$\boldsymbol{\alpha}$	-0.0007812	0.003231	-0.242	0.811
$\boldsymbol{\beta_1}$	0.00007192	0.0001148	0.626	0.536
$\boldsymbol{\beta_2}$	0.003623	0.0004163	8.703	NA

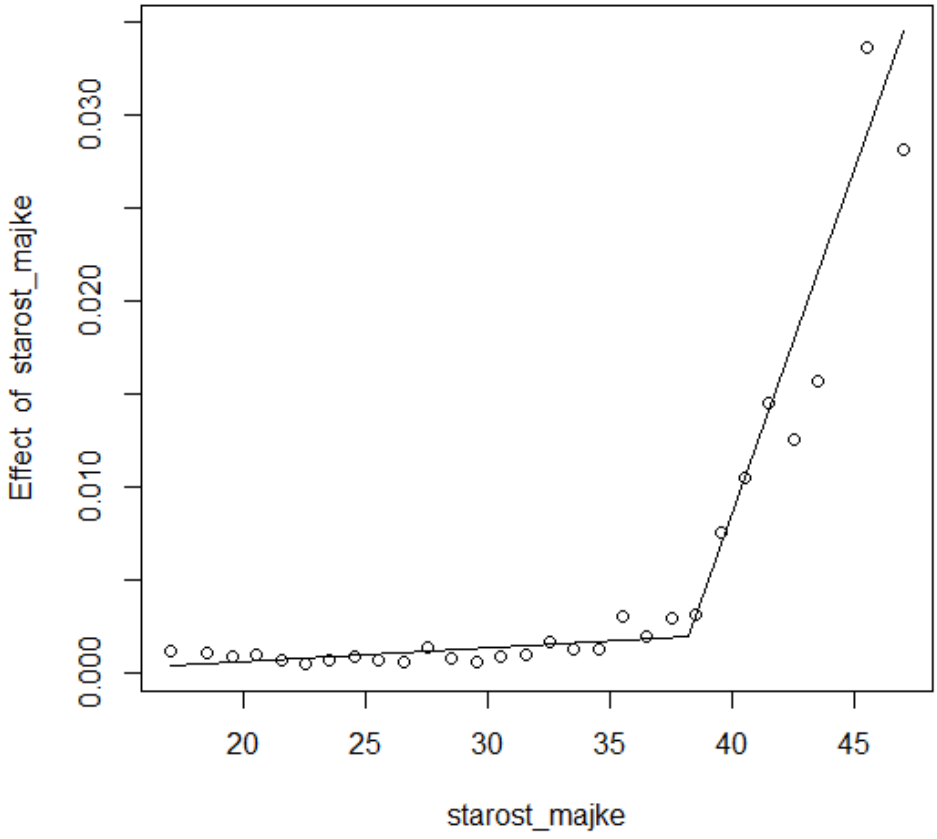
Tabela 7.1.2: *Ocena parametara*

Na osnovu ovih rezultata, nagib u drugom segmentu (nakon tačke promene) ocenjen je sa $\hat{\beta}_1 + \hat{\beta}_2 = 0.003695$, što znači da kako se povećava starost majke povećava se procenat beba rođenih sa Daunovim sindromom.

Dakle, ocenjeni model segmentirane regresije izgleda:

$$\hat{Y}_i = \underset{(0.003231)}{-0.0007812} + \underset{(0.0001148)}{0.00007192}X_i + \underset{(0.000416)}{0.003623}(X_i - \underset{(0.6867)}{38.2}).$$

Grafički prikaz procenjenog modela dat je na sledećem grafiku:



Grafik 7.1.2: *Ocenjeni model segmentirane regresije*

Ocenjeni model za posmatrani problem može se prikazati i jednačinama za svaki segment na sledeći način:

$$\hat{Y}_i = \begin{cases} -0.0007812 + 0.00007192X_i, & X_i \leq 38.2 \\ -0.1419302 + 0.003695X_i, & X_i > 38.2 \end{cases}.$$

Kako bi se proverila značajnost razlike u nagibu koristi se Dejvisov test. To vršimo pozivanjem funkcije *davies.test()* u programu R. Korišćenje ovog testa je pouzdano i zahteva da se navede regresioni model, promenljiva čiji segmentirani odnos se testira i broj tačaka za ocenjivanje. U našem primeru, to je segmentirani model, gde segmentiranu promenljivu predstavlja starost majke, a broj tačaka za ocenjivanje je 5.

Dejvisov test koristi samo Wald-ovu test statistiku, tj. $S(\xi_k) = \hat{\beta}_2 / SE(\hat{\beta}_2)$, za svako fiksirano ξ_k, iako se mogu koristiti alternativne test statistike. Oznaka SE je za standardnu grešku.

Ako tačka promene postoji, raspodela za $\hat{\beta}_2$ je Gausova, stoga ocene (i standardne greške) za nagibe možemo lako izračunati preko funkcije *slope()*, gde je interval poverenja 95%.

U sledećoj tabeli su prikazane ocene za nagibe za promenljivu koja predstavlja starost majke:

parametri	**ocenjena vrednost**	**standardna greška**	***t*-vrednost**	**Int.pov.(95%)_**	**Int.pov.(95%)+**
$\boldsymbol{\beta_1}$	0.00007192	0.0001148	0.6265	-0.0001641	0.0003079
$\boldsymbol{\beta_1 + \beta_2}$	0.003695	0.0004001	9.2340	0.0028720	0.0045170

Tabela 7.1.3 *Ocene za parametre koji predstavljaju nagib nezavisne promenljive (starost majke)*

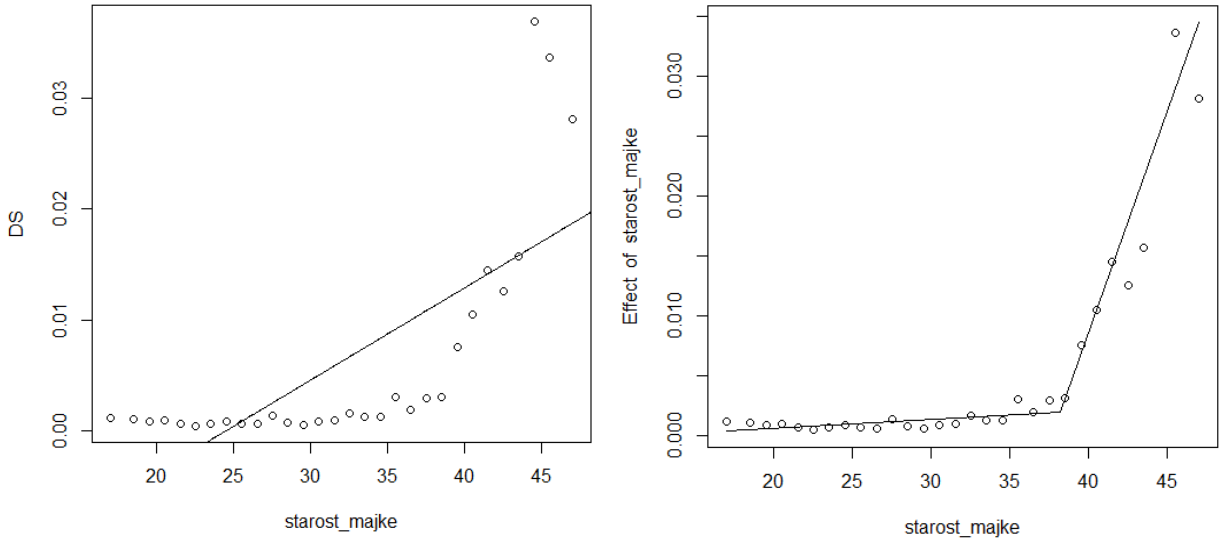
Pod nultom hipotezom segmentirani model može biti fitovan izostavljanjem segmentirane promenljive iz početnog modela. Na taj način dobijamo sledeće rezultate:

tačka promene	**početna vrednost**	**ocena**	**standardna greška**
$\boldsymbol{\xi}$	25	38.19659	0.5536572

Tabela 7.1.4 *Ocena tačke promene ako se isključi segmentirana promenljiva*

Iako fit nije suštinski promenjen, standardna greška tačke promene je primetno smanjena.

Sada ćemo uporediti model linearne regresije i model segmentirane linearne regresije. Koeficijent prilagođavanja kod linearne regresije iznosi samo $R^2 = 0.5369$, dok kod segmentirane regresije je $R^2 = 0.7968$.



Grafik 7.1.4 *Linearna i segmentirana regresija*

Vizelni pregled daje utisak da model segmentirane regresije bolje reprezentuje podatke za ovaj problem, što i jeste slučaj ako uporedimo njihove koeficijente prilagođavanja.

Dakle, rizik od pojave Daunovog sindroma kod novorođenčadi se povećava kako se povećava starost majke. Rizik naglo raste nakon 38. godine starosti majke i to predstavlja prag vrednost ovog problema.

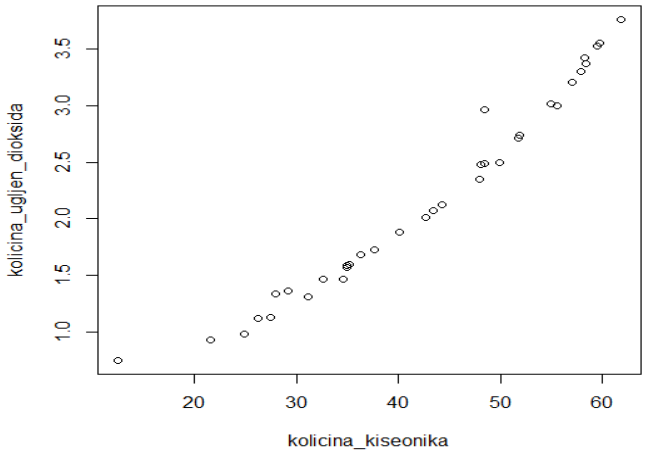
7.2 Metabolički procesi

Sada ćemo prikazati primer modela segmentirane regresije, koji predstavlja takozvani prag model.

Kada ljudi treniraju moraju da proizvedu energiju i postoje različiti metabolički putevi kojima se dobija ta energija (aerobni i anaerobni). Za datog pojedinca je važno da zna da li se dati put menja tokom vežbanja i ukoliko se menja, kada se to dešava. Jedan od načina da se ovo detektuje je putem ispitivanja veze između dve metaboličke promenljive tokom vremena, dok osoba trenira. U ovom konkretnom primeru posmatrana je osoba koja vesla, veslač je bio povezan sa opremom za merenje, koja očitava određene fizičke reakcije tokom vremena. Opterećenje je povećano tokom vremena, tj. povećana je otpornost veslača na veslanje [4].

Promenljive koje posmatramo u ovom primeru su količina udahnutog kiseonika (litara u minuti), što je nezavisna promenljiva, i količina izdahnutog ugljen-dioksida (litara u minuti), koja predstavlja ishodnu (zavisnu) promenljivu. Merenja su uzimana na svakih 30 sekundi do maksimalno 17.5 minuta. Ono što nas interesuje jeste da li postoji približno linearna veza između ove dve promenljive ili da li postoji promena u nagibu kada se dostigne kritičan nivo udisanja kiseonika. Tačka promene predstavlja tačku u kojoj se smenjuju metabolički putevi, iz aerobnog u anaerobni.

Na sledećem grafiku su prikazani podaci koje posmatramo (dijagram rasipanja):



Grafik 7.2.1: *Količina izdahnutog ugljen-dioksida (litar po minuti) u odnosu na količinu udahnutog kiseonika (litar po minuti)*

Procenjujemo model segmentirane regresije sa jednom tačkom promene:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 (X_i - \xi) + \varepsilon_i, \quad i = 1, \ldots, 35,$$

gde su ε_i nezavisne i identično raspodeljene slučajne promenljive, sa očekivanjem nula i disperzijom σ^2.

Na osnovu podataka za Y_i i X_i, koji su dati u drugoj tabeli u Dodatku, dobijeni su sledeći rezultati:

parametri	**ocenjene vrednosti**	**stand. greška**	***t*-vrednost**	***p*-vrednost**
$\boldsymbol{\alpha}$	0.074496	0.135824	0.548	0.587
$\boldsymbol{\beta_1}$	0.042350	0.004456	9.504	1.07e-10
$\boldsymbol{\beta_2}$	0.043990	0.005971	7.368	NA
$\boldsymbol{\xi}$	39.520	1.731		

Tabela 7.2.1: *Ocenjeni parametri*

U datoj tabeli su prikazane ocene parametara. Ocenjeni model izgleda:

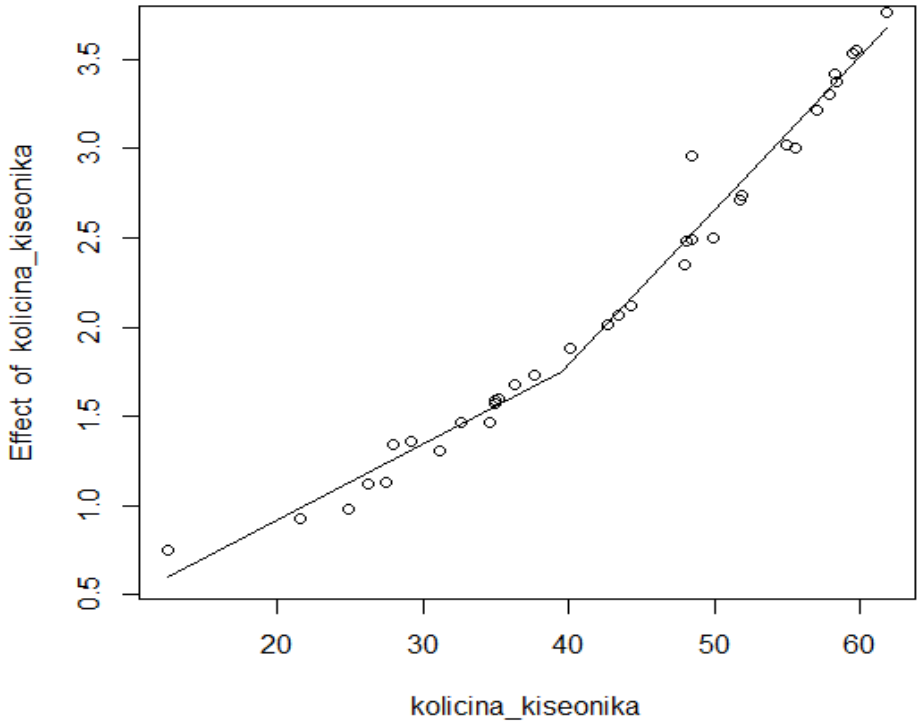
$$\hat{Y}_i = \underset{(0.1358)}{0.0745} + \underset{(0.0044)}{0.04235} X_i + \underset{(0.00597)}{0.04399} (X_i - \underset{(1.731)}{39.52}).$$

Regresione funkcije za svaki segment su sledeće:

$$\hat{Y}_i = \begin{cases} 0.074496 + 0.04235 X_i, & X_i \leq 39.52 \\ -1.66399 + 0.08635 X_i, & X_i > 39.52 \end{cases}.$$

Dakle, vidimo da kako se povećava količina udahnutog kiseonika povećava se količina ugljen-dioksida koju veslač izdiše.

Grafički se ocenjena regresiona linija ovog modela prikazuje na sledeći način:



Grafik 7.2.2: *Ocenjeni model segmentirane regresije*

Budući da postoji tačka promene, sada ćemo oceniti nagibe za nezavisnu promenljivu pre i posle tačke promene. Rezultati su prikazani u sledećoj tabeli:

parametri	**ocenjene vr.**	**standardna greška**	***t* vrednost**	**Int.pov.(95%)_**	**Int.pov.(95%)+**
$\boldsymbol{\beta_1}$	0.04235	0.004456	9.505	0.03327	0.05144
$\boldsymbol{\beta_1 + \beta_2}$	0.08635	0.003974	21.73	0.07824	0.09445

Tabela 7.2.2: *Ocene za nagibe nezavisne promenljive (količina udahnutog kiseonika)*

Vizelni pregled daje utisak da model segmentirane regresije dobro reprezentuje podatke za ovaj problem. Iako postoji više od dvostrukog povećanja u nagibu između dva segmenta modela, standardizovana razlika $\left(\frac{\widehat{\beta}_2-\widehat{\beta}_1}{\widehat{\sigma}}\right)$ je prilično mala.

Linearna veza između količine izdahnutog ugljen-dioksida i količine udahnutog kiseonika se menja kada količina kiseonika prelazi 39.52 litara po minuti. To može biti zbog činjenice da na početku vežbanja, tokom aerobne proizvodnje kiseonika, koristi se kiseonik, ali kako vežbanje postaje teže, veslačeve energetske potrebe prevazilaze količinu koja može da se proizvodi samo aerobnim putem. U ovom trenutku veslač počinje da koristi anaerobno proizvedenu energiju i to izaziva naglu promenu u linearnoj vezi između količine ugljen-dioksida i kiseonika.

U praksi bi ovaj model mogao biti korišćen na zdravim pojedincima u ranoj fazi uzimanja leka, čime bi se ispitivala potencijalna farmakološka aktivnost novog hemijskog jedinjenja. Više ljudi bi učestvovalo u istraživanju u kojem bi bili izloženi nizu različitih režima ili doza leka. Novi hemijska jedinjenja koja bi se mogla istražiti u ovom modelu su farmakološke terapije koje povećavaju glikogenolizu, povećavaju glikogen mišića i jetru ili terapije koje smanjuju proizvodnju mlečne kiseline, kao i kreatinin. Od ove vrste terapije bi se očekivalo da odlože tačku promene od aerobne do anaerobne proizvodnje.

Zaključak

U ovom radu ilustrovana je kroz primere ključna ideja segmentirane regresije i takav model može biti ocenjen u R softverskom programu kroz paket „segmented“. Iako se mogu primeniti alternativni pristupi za nelinearni model, na primer splajnovi, glavna prednost segmentirane regresije leži u interpretaciji parametara. Ponekad segmentirana regresija može da obezbedi razumnu aproksimaciju osnovnog oblika regresije, i prag vrednost i nagib mogu biti veoma informativni i značajni.

Zbog jednostavnosti, ograničili smo pažnju na slučaj segmentirane regresije u kojem su svi segmenti regresione funkcije u formi linearnog modela. Međutim, tehnike koje se upotrebljavaju trebalo bi da budu dovoljne, recimo aproksimacija Tejlorovim razvojem, da se obrade mnogi slučajevi u kojima su segmenti nelinearni. Isto tako posmatrali smo samo neprekidan slučaj, ali celokupna priča se može primeniti i na metod segmentirane regresije sa prekidima. Na osnovu rezultata dobijenih primenom segmentirane regresije na konkretne primere, metod segmentirane regresije daje veoma precizne procene parametara, kada je u pitanju neprekidni slučaj, dok u slučaju segmentirane regresije sa prekidima možda su više odgovarajuće neke druge metode koje se koriste, recimo Bejzov metod.

U ovom radu predložili smo procedure za detektovanje tačke promene, a nismo se bavili procenom broja tačaka promene, ali postoje neke metode koje bi mogle da pomognu da se odredi broj tačaka promene među nekoliko konkurentskih modela sa drugačijim brojem tačaka. Zatim, proučavali smo asimptotska ponašanja ocena parametara kod segmentirane regresije sa jednom segmentiranom promenljivom, gde smo pokazali da te ocene zadovoljavaju asimptotske osobine i da konvergiraju ka normalnoj raspodeli.

Ostala su neka otvorena pitanja, kao što su kako da se podele podaci koristeći više od jedne nezavisne promenljive. Zatim, u mnogim ekonomskim problemima, ishodna promenljiva pokazuje određene vrste zavisnosti tokom vremena, pa ukoliko nezavisna promenljiva predstavlja vremenski niz, odnosno uređena je u odnosu na vreme, tada model segmentirane regresije postaje prag autoregresivni model (*eng. threshold autoregressive model*). Ovaj interesantan nelinearni model vremenskih serija danas proučavaju mnogi autori, kao i model segmentirane regresije. Model segmetirane regresije ima sve veću primenu u realnom životu.

Dodatak

Ovde su prikazane tabele sa podacima koji su korišćeni u primerima primene segmentirane regresije, u sedmom poglavlju.

Podaci za primer u Poglavlju 7.1

U tabeli su prikazani prosečna starost majke, ukupan broj rođenih beba, kao i broj beba koje su rođene sa Daunovim sindromom.

Podaci su preuzeti iz:
C. J. Geyer, *Constrained maximum likelihood exemplified by isotonic convex logistic regression*, Journal of the American Statistical Association 86: 717–724, 1991

k	**prosečna starost majke**	**br. rođenih beba**	**br. slučajeva sa Daunovim sindromom**
1	**17**	**13555**	**16**
2	**18.5**	**13675**	**15**
3	**19.5**	**18752**	**16**
4	**20.5**	**22005**	**22**
5	**21.5**	**23896**	**16**
6	**22.5**	**24667**	**12**
7	**23.5**	**24807**	**17**
8	**24.5**	**23986**	**22**
9	**25.5**	**22860**	**15**
10	**26.5**	**21450**	**14**
11	**27.5**	**19202**	**27**
12	**28.5**	**17450**	**14**
13	**29.5**	**15685**	**9**
14	**30.5**	**13954**	**12**
15	**31.5**	**11987**	**12**

16	**32.5**	**10983**	**18**
17	**33.5**	**9825**	**13**
18	**34.5**	**8483**	**11**
19	**35.5**	**7448**	**23**
20	**36.5**	**6628**	**13**
21	**37.5**	**5780**	**17**
22	**38.5**	**4834**	**15**
23	**39.5**	**3961**	**30**
24	**40.5**	**2952**	**31**
25	**41.5**	**2276**	**33**
26	**42.4**	**1589**	**20**
27	**43.5**	**1018**	**16**
28	**44.5**	**596**	**22**
29	**45.5**	**327**	**11**
30	**47**	**249**	**7**

Podaci za primer u Poglavlju 7.2

U tabeli su prikazani količina udahnutog kiseonika (u litrima po minuti), količina izdahnutog ugljen-dioksida (u litrima po minuti), kao i vreme merenja.

Vreme	**Količina kiseonika (X) (l/min)**	**Količina ugljen-dioksida (Y) (l/min)**
1	**12.5**	**0.75**
2	**26.2**	**1.12**
3	**24.8**	**0.98**
4	**27.4**	**1.13**
5	**31.1**	**1.31**
6	**34.6**	**1.47**
7	**21.5**	**0.93**
8	**27.9**	**1.34**
9	**29.2**	**1.36**
10	**35.2**	**1.60**

11	**32.6**	**1.47**
12	**34.9**	**1.57**
13	**34.9**	**1.59**
14	**37.6**	**1.73**
15	**36.3**	**1.68**
16	**40.1**	**1.88**
17	**42.7**	**2.01**
18	**43.4**	**2.07**
19	**44.2**	**2.12**
20	**47.9**	**2.35**
21	**49.9**	**2.50**
22	**48.1**	**2.48**
23	**48.4**	**2.49**
24	**51.7**	**2.71**
25	**51.8**	**2.74**
26	**55.5**	**3.00**
27	**54.9**	**3.02**
28	**57.0**	**3.21**
29	**57.9**	**3.30**
30	**58.3**	**3.37**
31	**58.2**	**3.42**
32	**59.5**	**3.53**
33	**59.7**	**3.55**
34	**61.8**	**3.76**
35	**48.4**	**2.96**

Literatura

[1] B. Baltagi, *Econometrics*, Fifth Edition, New York: Springer-Verlag , 2011

[2] C. Diniz, L. Brochi, *Robustness of two-phase regression tests*, REVSTAT-Statistical Journal, Volume 3, Number 1, 1-18, 2005

[3] C. Chen, J. Chan, R. Gerlach, W. Hsieh, *A comparison of estimators for regression models with change points* 21: 395-414, 2011

[4] S. Julious, *Inference and Estimation in a Changepoint Regression Problem*, The Statistician 50: 51-61, 2001

[5] F. Osorio, M. Galea, *Detection of a Change-point in Student-t Linear Regression Models*, Departamento de Estadistica, Universidad de Valparaiso, Chile, 2004

[6] J. Kmenta, *Počela ekonometrije*, drugo izdanje, MATE d.o.o. , Zagreb, 1997

[7] Liquid Gold team, *Drainage research in farmer's fields: analysis of data*, Part of project Liquid Gold" of the International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands, July 2002

[8] Peter Filzmoser, *Linear and Nonlinear Methods for Regression and Classification and applications in R*, Department of Statistics and Probability Theory, Vienna University of Technology, 2008

[9] R. Berk, *Statistical Learning from a Regression Perspective*, Springer Science and Business Media, LLC, 2008

[10] R. Quandt, *The estimation of the parameters of a linear regression system obeying two separate regimes,* Journal of the American Statistical Association, Vol. 53, No. 284, str. 873-880, 1958

[11] W. Shiying, *Asimptotic inference for segmented regression models*, The University of British Columbia, 1993

[12] V. Muggeo, *Segmented: an R package to fit regression models with broken-line relationships*, The Newsletter of the R project, 8/1, 20-25, 2008

[13] V. Muggeo, *Estimating regression models with unknown break-points*, Statistics in Medicine 22: 3055–3071, 2003

[14] Z. Hualing, H. Chen, *Detecting Change Points in Segmented Linear Regression Heteroscedastic Models by Empirical Likelihood Methods*, International Journal of Intelligent Technologies and Applied Statistics 5: 75-85, 2012

[15] Z. Liu, L. Qian, *Changepoint estimation in a segmented linear regression via empirical likelihood*, Communications in Statistics-Simulation and Computation 39: 85-100, 2010

[16] G. Seber, C. Wild, *Nonlinear Regression*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2003

[17] P. Feder, *On asymptotic distribution theory in segmented regression problems-identified case*, The Annals of Statistics 3: 49–83, 1975

Kratka biografija

Suzana Vidić je rođena 13. marta 1989. godine u Šapcu. Završila je osnovnu školu „Nata Jeličić“ u Šapcu, kao nosilac *Vukove diplome*, a potom društveno-jezički smer „Šabačke gimnazije“ u Šapcu, sa odličnim uspehom.

Po završetku srednje škole, 2008. godine, upisuje osnovne akademske studije na Prirodno-matematičkom fakultetu u Novom Sadu, smer primenjena matematika (modul: matematika finansija), koje završava u julu 2011. godine.

Iste godine, u oktobru, upisuje master akademske studije, na istom fakultetu, takođe smer primenjena matematika. Položila je sve ispite predviđene planom i programom, zaključno sa junskim ispitnim rokom 2013. godine, i time stekla uslov za odbranu master rada.

UNIVERZITET U NOVOM SADU

PRIRODNO-MATEMATIČKI FAKULTET

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:	
RBR	
Identifikacioni broj:	
IBR	
Tip dokumentacije:	Monografska dokumentacija
TD	
Tip zapisa:	Tekstualni štampani materijal
TZ	
Vrsta rada:	Master rad
VR	
Autor:	Suzana Vidić
AU	
Mentor:	dr Zorana Lužanin
MN	
Naslov rada:	Segmentirana regresija sa primenom
NR	
Jezik publikacije:	srpski (latinica)
JP	
Jezik izvoda:	srpski/engleski
JI	
Zemlja publikovanja:	Republika Srbija
ZP	

Uže geografsko područje:	Vojvodina
UGP	
Godina:	2014.
GO	
Izdavač:	Autorski reprint
IZ	
Mesto i adresa:	Prirodno-matematički fakultet, Departman za matematiku i informatiku, Trg Dositeja Obradovića 4, Novi Sad
MA	
Fizički opis rada:	(7/77/17/8/8/1)
FO	(broj poglavlja/strana/literatura/tabela/grafika/dodataka)
Naučna oblast:	Matematika
NO	
Naučna disciplina:	Ekonometrija
ND	
Predmetna odrednica/ ključne reči:	segmentirana regresija, tačka promene, konzistentnost, asimptotska raspodela, testiranje hipoteza
PO	
UDK	
Čuva se:	Biblioteka Departmana za matematiku i informatiku, Prirodno matematički fakultet, Univerzitet u Novom Sadu
ČU	
Važna napomena:	nema
VN	
Izvod:	Na početku rada su navedeni osnovni pojmovi linearne regresije i predstavljen je model segmentirane regresije. Potom sledi teorija ocena, gde su ocenjeni parametri ovog modela i testirane hipoteze. Predstavljeni su testovi kojima se detektuje tačka promene. Nakon toga se govori o asimptotskom ponašanju parametara modela, zatim o
IZ	

	uspešnosti procenjivanja modela i uticaju autlajera. Na kraju rada model je primenjen na dva konkretna primera, koristeći podatke iz medicinskih istraživanja i softverski program R.
Datum prihvatanja teme od NN veća:	11.06.2013.
DP	
Datum odbrane:	maj 2014.
DO	
Članovi komisije:	
KO	
Predsednik:	dr Andreja Tepavčević, redovni profesor Prirodno-matematičkog fakulteta, Univerziteta u Novom Sadu
Član:	dr Dora Seleši, vanredni profesor Prirodno-matematičkog fakulteta, Univerziteta u Novom Sadu
Mentor:	dr Zorana Lužanin, redovni profesor Prirodno-matematičkog fakulteta, Univerziteta u Novom Sadu

UNIVERSITY OF NOVI SAD

FACULTY OF SCIENCE

KEY WORDS DOCUMENTATION

Accession number:	
ANO	
Identification number:	
INO	
Document type:	Monograph publication
DT	
Type of record:	Textual printed material
TR	
Content code:	Master's thesis
CC	
Author:	Suzana Vidić
AU	
Mentor/comentor:	Zorana Lužanin, Ph. D.
MN	
Title:	Segmented regression and its application
TI	
Language of text:	Serbian (Latin)
LT	
Language of abstract:	Serbian/English
LA	
Country of publication:	Republic of Serbia
CP	

Locality of publication: **LP**	Vojvodina
Publication year: **PY**	2014.
Publisher: **PU**	Author's reprint
Publication place: **PP**	Department of Mathematics and Informatics, Faculty of Science, Trg Dositeja Obradovića 4, Novi Sad
Physical description: **PD**	(7/77/17/8/8/1) (chapters/pages/references/tables/graphs/add lists)
Scientific field: **SF**	Mathematics
Scientific discipline: **SD**	Econometrics
Subject/ Key words: **SKW**	segmented regression, change point, consistency, asymptotic distribution, hypothesis testing
UC	
Holding data: **HD**	Library of Department of Mathematics and Informatics, Faculty of Science, University of Novi Sad
Note: **N**	none
Abstract: **AB**	At the very beginning of the paper, the basic terms of linear regression are defined and the model of segmented regression is presented. Then follows evaluation theory, where parameters of this model are evaluated and hypothesis tested.

Tests for detection of the change points are presented. After that there is elaboration on asymptotic behavior of parameters, that is consistency, convergency and asymptotic distribution. The next is story on model goodness of fit and the impact of outliers. The end of the paper shows the application of two concrete examples, by using the data from medical research and software package R.

Accepted by the Scientific Board: 11.06.2013.

ASB

Defended on: may 2014.

DE

Thesis defend board:

DB

President: Andreja Tepavčević, Ph. D. , full professor, Faculty of Science, University of Novi Sad

Member: Dora Seleši, Ph. D. , associate professor, Faculty of Science, University of Novi Sad

Member: Zorana Lužanin, Ph. D. , full professor, Faculty of Science, University of Novi Sad