Kvantilna regresija

Rumenjak, Andrea

Master's thesis / Diplomski rad

2014

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:168519

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-03-24**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Andrea Rumenjak

KVANTILNA REGRESIJA

Diplomski rad

Voditelj rada:
prof. dr. sc. Miljenko Huzak

Zagreb, srpanj,2014.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem svom mentoru prof. dr. sc. Miljenku Huzaku na pomoći i strpljenju prilikom izrade ovog rada.
Zahvaljujem se i svojoj obitelji i svim prijateljima koji su mi pružali podršku i pomoć tijekom cijelog mog studija.

Sadržaj

Sadržaj **iv**

Uvod **1**

1 Prva regresija **3**
1.1 Boškovićev primjer 3
1.2 Povezanost s medianom 7
1.3 Regresija i ostali kvantili 8
1.4 Kratki uvid u kvantilnu regresiju 10

2 Svojstva kvantilne regresije **14**
2.1 Uvod u kvantilnu regresiju 14
2.2 Uvjet podgradijenta 16
2.3 Ekvivarijanca 19
2.4 Robustnost 20
2.5 Ukratko o Waldovu testu 23

3 Interpretacija kvantilne regresije i primjeri **24**
3.1 Efekt tretmana 24
3.2 Interpretacija modela kvantilne regresije 27
3.3 Interpretacija pogrešno postavljenog modela kvantilne regresije 28
3.4 Primjeri 30

Bibliografija **38**

Uvod

Kvantilna regresija, statistička je metoda koja je našla svoju primjenu u ekonometriji, ekologiji i biometrici. Zanimljiva je činjenica da se kvantilna regresija koristi kod modela određivanja cijene uloženoga kapitala (CAPM) i kod *Fama-French three factor modela*(Vidjeti [1]). Ako promatramo robusne mjere rizika kao što su vrijednost pod rizikom (VaR) ili uvjetna vrijednost pod rizikom (CVaR), tada je naglasak stavljen na lijevi rep distribucije povrata. Metoda najmanjih kvadrata konstruirana je tako da se fokusira na očekivanu vrijednost te zbog toga se ne mogu proučavati granične vrijednosti ili kvantili distribucije. Isto tako, metoda najmanjih kvadrata (OLS) je tehnika koja zahtjeva pretpostavku normalnosti podataka što nije slučaj kod teških repova i kod asimetrije koja se susreće u distribucijama povrata financijske imovine. Kvantilna regresija je stekla popularnost u financijama kao alternativa metodi najmanjih kvadrata zato što je to robusna tehnika koja može objasniti lijeve i desne repove najčešće nesimetrične distribucije povrata, objašnjava outliere ili ekstreme distribucije te je time učinkovitija za ispitivanje rizika.

U radu su opisana najvažnija obilježja kvantilne regresije te kroz primjere vidimo njenu upotrebu.

U prvom poglavlju opisana je prva regresija koju je proveo Ruđer Bošković. Razrađen je geometrijski postupak kojim je izračunao eliptičnost Zemlje te analitički dio postupka koji je povezan sa kvantilnom regresijom. Također su uvedeni osnovni pojmovi vezani uz kvantilnu regresiju kao što su medijan, kvantili, funkcija gubitka te je pomoću osnovnih pojmova napravljen kratki uvid u kvantilnu regresiju i opisan njezin postupak.

Sljedeće obrađeno u radu su osnovna svojstva kvantilne regresije koja su opisana u drugom poglavlju. Najvažnija svojstva su uvjet podgradijenta, ekvivarijanca i robustnost kvantilne regresije gdje je opisana funkcija utjecaja te je objašnjena važnost tih svojstava. Nakon toga spomenut je Waldov test koji je razrađen za model dva uzorka.

Zadnje poglavlje posvećeno je primjeni kvantilne regresije. Opisan je efekt tretmana, objašnjena je interpretacija modela kvantilne regresije, krivo postavljenog modela kvan-

tilne regresije i provedena je konkretna analiza podataka. U primjerima je kvantilna regresija primjenjena na AR(1) model te je provedena analiza kako količina prihoda u kućanstvo belgijske radničke klase utječe na potrošnju za hranu.

Grafikoni i rezultati u ovom radu su dobiveni pomoću programskih jezika **R** i **SAS**.

Poglavlje 1

Prva regresija

1.1 Boškovićev primjer

Tijekom 18. stoljeća među znanstvenicima se vodila diskusija o pitanju oblika Zemlje. Postavljalo se pitanje je li zemlja spljoštena na polovima ili nije. Krajem 17. stoljeća Newton je dokazao spljoštenost zemlje na polovima i to pomoću njezine rotacije. Domenico Cassini je smatrao da je zemlja spljoštena u području ekvatora, a do tog zaključka je došao računavši duljinu luka stupnja meridijana (luk koji na meridijanskoj kružnici omeđuju dva polumjera Zemlje i pri tome u točki središta Zemlje zatvaraju kut od 1°). U tom periodu dva glavna načina određivanja oblika Zemlje su bila: eksperimenti pomoću njihala i određivanje duljine luka meridijana.
Ruđer Bošković je smatrao da je najbolji način određivanja nepravilnosti oblika Zemlje točno određivanje dva stupnja meridijana na različitim geografskim dužinama, ali na istim geografskim širinama. Kako bi što preciznije odredio oblik Zemlje, Bošković je usporedio pet vrijednosti duljine luka stupnja meridijana. Mjerenja su bila provedena u Quito, Rtu Dobre Nade, Parizu, Laponiji i Rimu (podaci su prikazani u tablici 1.1).

Tablica 1.1: Podaci duljine luka stupnja meridijana

Lokacija	Geografska širina	duljina luka	$\sin^2$(širina)
Quito	$0^\circ 0'$	56751	0
Rt Dobre Nade	$33^\circ 18'$	57037	0.2987
Rim	$42^\circ 59'$	56979	0.4648
Pariz	$49^\circ 23'$	57074	0.5762
Laponija	$66^\circ 19'$	57422	0.8386

U svome radu Bošković nije dao analitičko rješenje, već je po uzoru na Newtonovu tradi-

ciju dao samo geometrijski opis. Mi ćemo provesti i jednu i drugu analizu paralelno. U analitičkom smislu, Bošković je imao 5 odgovarajućih jednadžbi:

$$y_i = a + b \sin^2 \lambda_i \tag{1.1}$$

gdje y predstavlja duljinu luka, a λ geografsku širinu. Nepoznanice a i b redom predstavljaju duljinu stupnja luka na ekvatoru i odstupanje duljine luka na polu od njegove vrijednosti na ekvatoru. Budući da bilo koje dvije lokacije mogu poslužiti kako bi se izračunao a odnosno b, Bošković je to učinio za svih 10 kombinacija te je još izračunao i eliptičnost po formuli $\frac{1}{\text{elipticnost}} = \frac{3a}{b}$ (podaci su prikazani u tablici 1.2 i nešto se razlikuju od Boškovićevih rezultata). Time je dobio 10 različitih rješenja.

Tablica 1.2: Rezultati od a i b dobiveni za svih 10 kombinacija i izračunata eliptičnost Zemlje

Par	a	b	eliptičnost	Par	a	b	eliptičnost
(1,2)	56751	957	$\frac{1}{178}$	(1,3)	56751	491	$\frac{1}{347}$
(1,4)	56751	561	$\frac{1}{304}$	(1,5)	56751	800	$\frac{1}{213}$
(2,3)	57141	-349	$\frac{-1}{491}$	(2,4)	56997	133	$\frac{1}{1285}$
(2,5)	56824	713	$\frac{1}{239}$	(3,4)	56583	853	$\frac{1}{199}$
(3,5)	56428	1185	$\frac{1}{143}$	(4,5)	56310	1326	$\frac{1}{127}$

Bošković je pokušao sjediniti dobivene rezultate te je izračunao aritmetičku sredinu varijable b i koristeći duljinu luka na ekvatoru na mjestu Quito ($z = 56751$), dobio je vrijednost eliptičnosti 1/256. No, s time nije bio zadovoljan pa je u drugom pokušaju izbacio parove (2,3) i (2,4) jer je smatrao da se najviše razlikuju. Tada je dobio vrijednost eliptičnosti 1/198. Ni ta vrijednost nije bila zadovoljavajuća te je Bošković odlučio promatrati odstupanja između dobivene prosječne vrijednosti i ostalih 10 odnosno 8 vrijednosti u tablici.

Nakon par godina, Bošković je odlučio generalizirati problem te je napravio listu pretpostavki koje mora zadovoljavati srednja vrijednost dobivena pomoću duljine luka.
Neka y_1, y_2, ... predstavljaju vrijednosti duljine luka, a λ_1, λ_2, ... geografsku širinu. Te neka δy_i predstavlja vrijednost "korekcije" koju bi Bošković napravio stupnju y_i. Prva pretpostavka je da "iskorigirane" vrijednosti luka $y_i + \delta y_i$ zadovoljavaju

$$y_i + \delta y_i = a + b \sin^2 \lambda_i \tag{1.2}$$

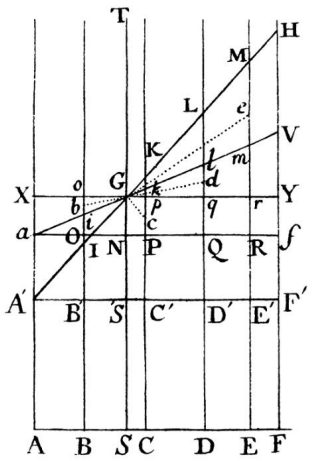
za neki izbor a i b. Druga i treća pretpostavka odnose se na izbor a i b. Druga pretpostavka kaže da se zbroj pozitivnih i zbroj negativnih korekcija mora podudarati po apsolutnoj vrijednosti, odnosno da je zadovoljena sljedeća pretpostavka

$$\sum_i \delta y_i = 0. \tag{1.3}$$

Treća pretpostavka je da je suma apsolutnih vrijednosti korekcija minimalna, tj.

$$\sum_i |\delta y_i| \longrightarrow min. \tag{1.4}$$

Bošković je svoju ideju samo izrekao i okrenuo se geometrijskom rješavanju problema (slika 1.1).



Slika 1.1: Boškovićev alogoritam. Horizontalni pravac AF prikazuje $\sin^2 \lambda$, gdje je λ geografska širina na sredini luka. Vertikalni pravci AX prikazuju duljinu luka po stupnju. Pet vrijednosti duljine luka predstavljene su pomoću a, b, c, d i e. G predstavlja centar ravnoteže.

Neka je AF jedna jedinica duljine te neka je A početak, a A, B, C, D i E neka predstavljaju pet vrijednosti $\sin^2(\lambda_i)$ prikazanih na intervalu. Duljine Aa, Bb, Cc Dd i Ee predstavljaju duljinu odgovarajućeg luka. Problem je bio kako pronaći pravac $A^{'}H$, takav da korekcije $aA^{'}$, bO, cK, dL i eM zadovoljavaju uvjet 1.3 i 1.4. Prvo što je Bošković napravio je da je promatrao drugi uvjet, koji opisuje stanje mehaničke ravnoteže. Došao je do zaključka da pravac $A^{'}H$ mora prolaziti kroz ravnotežnu točku (G). Time je smanjio problem na traženje pravca kroz točku G koja zadovoljava treći uvjet.

Kako bi pronašao rješenje, Bošković je zamislio pravac SGT koji je okretao u smjeru kazaljke na satu u točki G. Prilikom rotiranja pravca, suma apsolutnih vrijednosti korekcija bi se smanjivala sve dok minimum nije postignut, te bi onda opet počela rasti. Budući da bi se rotiranjem pravca vrijednosti luka mijenjale proporcionalno udaljenostima AS, BS, SC, SD i SE, dovoljno bi bilo rotirati pravac sve dok ne prođe dovoljan broj točaka tako da barem polovina sume udaljenosti AS + BS + SC + SD + SE ne dostigne vrijednost koja odgovara točkama koje je pravac prošao. Vrijednost sume korekcija do tog trenutka pada, a nakon toga raste. Bošković je uz još jednu pretpostavku pojednostavio problem. Zaključio je da će rotirajući pravac SGT susresti svih pet točaka u obrnutom redoslijedu na obroncima pet pravaca koji idu iz točke G do točaka aX/AS, bo/BS, cp/SC i tako dalje.

Bošković je opravdao svoj geometrijski algoritam tako da je proveo analizu tablice 1.2. Prvo je našao točku G tako da je izračunao aritmetičku sredinu od $\sin^2 \lambda_i$. (gdje je dobio vrijednost AS = 0.4357) i aritmetičku sredinu od y_i (gdje je dobio vrijednost SG=57053). Zatim je računao $\sin^2 \lambda_i - AS$ (gdje je dobio vrijednosti za AS, BS, SC, SD i SE), razliku $y_i - SG$ (gdje je dobio vrijednosti za aX, bo, cp, qd i re) te je na kraju još izračunao omjer $(\sin^2 \lambda_i - AS)/(y_i - SG)$. (Podaci su prikazani u tablici 1.3)

Tablica 1.3: Opravdanje geometrijskog algoritma

Luk	$\sin^2 \lambda_i - AS$	$y_i - SG$	$(\sin^2 \lambda_i - AS)/(y_i - SG)$
a	-0.4357	-302	0.0014
b	-0.1370	-16	0.0086
c	0.0291	-74	-0.0004
d	0.1405	21	0.0067
e	0.4029	369	0.0011

Sada je Bošković mogao krenuti na proučavanje svog algoritma. Prvo što je uočio je da je omjer zapravo inverz vrijednosti koeficjenata smjera pravaca koji prolazi ranije opisanim točkama te je zatim iz vrijednosti omjera zaključio da je rotirajući pravac prolazio točkama e, a, d, b i na kraju kroz točku c. Zatim si je Bošković postavio pitanje koliko daleko mora ići. Suma apsolutnih vrijednosti udaljenosti dobivenih u tablici 1.3 prikazanih u drugom stupcu je 1.1452 i kad računamo jednu polovinu sume (kako je navedeno u algoritmu) dobijemo vrijednost 0.5726. Po algoritmu vidimo da bi Bošković trebao nastaviti rotirati pravac sve dok ne dostigne točku gdje je suma jednaka 0.5726. Budući da je SE = 0.4029 < 0.5726 a SE + SA = 0.8386 > 0.5726, možemo zaključiti da bi se pravac trebao moći rotirati sve dok ne dođe do točke a. Dok se pravac rotirao između točaka a i e, vrijednost korekcije točke e se povećala proporcionalno SE = 0.4029, a vrijednost ostalih korekcija se smanjila proporcionalno vrijednosti SA+SD+SB+SC = 0.7468. Kada je rotirajući pravac

prošao točku a, ukupno povećanje je proporcionalno SE + SA = 0.8386, dok je ukupno smanjenje proporcionalno vrijednosti SD + SB + SC = 0.3066. Iz toga svega možemo zaključiti da se rotirajući pravac SGT mora zaustaviti kada dostigne točku a, što odgovara vrijednostima $a = 56751$, $b = 692$ i eliptičnost = $1/248$ (gdje je koristio malo drugačiju formulu $\frac{1}{\text{elipticnost}} = \frac{3a}{b} + 2$).

1.2 Povezanost s medianom

Tijekom svog istraživanja, Ruđer Bošković nije bio svjestan da se svojim pokušajem prve regresije približava pojmu kvantilne regresije i to pola stoljeća prije nego što je prvi puta objavljen rad o metodi najmanjih kvadrata (1805., Adrien-Marie Legendre). Zanimljivo je da je pojam metode najmanjih kvadrata povezan sa prvim djelom istraživanja. Procjenitelj dobiven metodom najmanjih kvadrata može se povezati s pojmom težinskog prosjeka. Neka je h indeks jednog od naših 10 parova pravaca te neka je $\beta(h) = (a(h), b(h))'$ vektor njegovih koeficijenata tj.:

$$\beta(h) = X(h)^{-1}y(h). \tag{1.5}$$

Za naš bivarijantni model i za $h = (i, j)$ imamo:

$$X(h) = \begin{pmatrix} 1 & x_i \\ 1 & x_j \end{pmatrix} \tag{1.6}$$

$$y(h) = \begin{pmatrix} y_i \\ y_j \end{pmatrix} \tag{1.7}$$

Može se pokazati da je procjenitelj za vektor parametara dobiven metodom najmanjih kvadrata

$$\hat{\beta} = \sum_h w(h)\beta(h), \tag{1.8}$$

gdje je uz oznaku $|X(h)| = det(X(h))$,

$$w(h) = \frac{|X(h)|^2}{\sum_h |X(h)|^2}. \tag{1.9}$$

Ovakav prikaz procjenitelja za β dobiven metodom najmanjih kvadrata lako se može proširiti na model linearne regresije sa p parametara. U primjeru bivarijatnog modela, težine w su proporcionalne udaljenostima između dizajniranih točaka, što nam ukazuje na to da je metoda najmanjih kvadrata osjetljiva na outliere. Upravo je to najveća mana metode najmanjih kvadrata te se nadamo da će se metodom kvantilne regresije taj problem smanjiti.

Tijekom drugog djela istraživanja, Bošković se sve više približava pojmu kvantilne regresije i to pomoću treće pretpostavke 1.4. Njegova ideja je objašnjena u poglavlju 1.1., a kasnije je Laplace taj postupak nazvao izračunavanje težinskog medijana. Možemo staviti da je

$$b_i = \frac{y_i - \bar{y}}{x_i - \bar{x}} \tag{1.10}$$

te svaki nagib pravaca b_i možemo povezati sa težinom $w_i = |x_i - \bar{x}|$. Poredamo nagibe po veličini i njihove uređene vrijednosti označimo s $b_{(i)}$ (i-te po veličini), te njihove pripadne težine označimo sa $w_{(i)}$. Želimo naći najmanji j tako da vrijedi:

$$\sum_{i=1}^{j} w_{(i)} > \frac{1}{2} \sum_{i=1}^{n} w_{(i)}. \tag{1.11}$$

Označimo traženi j sa j^*. Tada je Boškovićev procjenitelj za nagib pravca b jednak $\hat{b} = b_{(j^*)}$. Kada pogledamo sve što je Bošković predložio, možemo zaključiti da je zapravo napravio spoj između medijana i aritmetičke sredine (parametar b koji predstavlja nagib pravca je procijenio pomoću medijana, dok je parametar a slobodni član procijenjen pomoću aritmetičke sredine). Prirodno se postavlja pitanje: "postoji li povezanost regresije sa ostalim kvantilima?"

1.3 Regresija i ostali kvantili

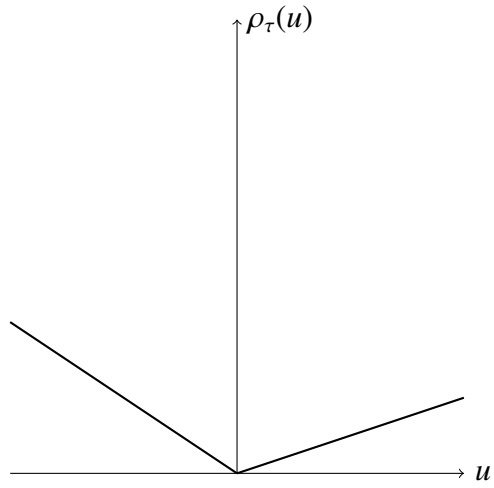
Prvo želimo vidjeti kako se definiraju kvantili.

Definicija 1.3.1. *Neka je X slučajna varijaba koja ima funkciju distribucije $F(x) = \mathbb{P}(X \leq x)$. Tada za zadani $\tau \epsilon \langle 0, 1 \rangle$ kažemo da je $Q(\tau)$ $\tau - ti$ kvantil slučajne varijable X ako vrijedi*

$$Q(\tau) = F^{-1}(\tau) = \inf \{x : F(x) \geq \tau\}. \tag{1.12}$$

Pretpostavimo da nam je potrebna procjena točkom za neku slučajnu varijablu, koja ima a posteriori funkciju distribucije F. Neka je funkcija gubitka za neki $\tau \epsilon \langle 0, 1 \rangle$ zadana s:

$$\rho_\tau(u) = u(\tau - \mathbb{1}_{u<0}). \tag{1.13}$$



Slika 1.2: Funkcija gubitka

Tada želimo naći $\hat{x}$ tako da minimiziramo sljedeću funkciju gubitka:

$$\begin{aligned} \mathbb{E}(\rho_\tau(X-\hat{x})) &= \int_{-\infty}^{\infty} \rho_\tau(x-\hat{x})dF(x) \\ &= (\tau-1)\int_{-\infty}^{\hat{x}} (x-\hat{x})dF(x) + \tau\int_{\hat{x}}^{\infty} (x-\hat{x})dF(x) \end{aligned} \tag{1.14}$$

Nakon derivacije po varijabli $\hat{x}$ dobijemo:

$$\begin{aligned} 0 &= (\tau-1)\int_{-\infty}^{\hat{x}} (-1)dF(x) + \tau\int_{\hat{x}}^{\infty} (-1)dF(x) \\ &= (1-\tau)\int_{-\infty}^{\hat{x}} (-1)dF(x) - \tau\int_{\hat{x}}^{\infty} dF(x) \\ &= F(\hat{x}) - \tau \end{aligned} \tag{1.15}$$

tj, $F(\hat{x}) = \tau$. Zbog monotonosti funkcije F vrijedi da bilo koji element iz $\{x : F(x) = \tau\}$ može minimizirati gubitak. Iz toga slijedi da imamo interval τ-tog kvantila te onda iz tog intervala izaberemo najmanji element.

Kako interpretativno možemo povezati funkciju gubitka s kvantilima? Prirodno je povezati kvantile s procjenom optimalne točke kod asimetričnog linearnog gubitka. U slučaju simetrije očito je da imamo medijan. Kada je linearni gubitak asimetričan, onda biramo procijenjenu točku tako da se ona nalazi na monotonijoj strani marginalnog gubitka. Na primjer, ako je podcijenjeni dio marginalno 4 puta skuplji od precjenjenog dijela, izabrat

ćemo takav $\hat{x}$ da vrijedi da je $F(\hat{x}) = 4 \cdot (1 - F(\hat{x}))$. Tada ćemo dobiti da je $\hat{x}$ 80-ti percentil od F.

Zamjenimo sada F sa empirijskom funkcijom distribucije. Tada dobivamo

$$F_n(x) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}. \tag{1.16}$$

I u ovom slučaju tražimo $\hat{x}$ tako da minimiziramo očekivani gubitak i time dobijemo τ-ti uzorački kvantil:

$$\int \rho_\tau(x - \hat{x}) dF_n(x) = n^{-1} \sum_{i=1}^{n} \rho_\tau(x_i - \hat{x}). \tag{1.17}$$

Ponovno dobivamo interval rješenja za τ_n, $\{x : F_n(x) = \tau\}$. Time smo opisali problem traženja τ-tog uzoračkog kvantila.

1.4 Kratki uvid u kvantilnu regresiju

Problem traženja τ-tog uzoračkog kvantila može se zapisati i kao:

$$\min_{\xi \epsilon \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi). \tag{1.18}$$

Iz metode najmanjih kvadrata znamo da aritmetička sredina rješava problem

$$\min_{\mu \epsilon \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2, \tag{1.19}$$

pa nam to sugerira da ako izrazimo uvjetno očekivanje od y uz uvjet x kao $\mu(x) = x'\beta$, da tada možemo procijeniti β rješavanjem:

$$\min_{\beta \epsilon \mathbb{R}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2. \tag{1.20}$$

Na sličan način imamo da za τ-ti uzorački kvantil, $\hat{\alpha}(\tau)$ rješava problem:

$$\min_{\alpha \epsilon \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha). \tag{1.21}$$

Sada možemo definirati funkciju τ-tog uvjetnog kvantila sa $Q_y(\tau|x) = x'\beta(\tau)$, uzimajući u obzir da $\hat{\beta}(\tau)$ rješava problem:

$$\min_{\beta \epsilon \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x'\beta). \tag{1.22}$$

Problem kvantilne regresije 1.22, možemo zapisati i kao problem linearnog programiranja:

$$\min_{(\beta,u,v)\epsilon\mathbb{R}^p x\mathbb{R}_+^{2n}} \{\tau 1_n' u + (1-\tau) 1_n' v | X\beta + u - v = y\} \tag{1.23}$$

gdje $1_n'$ predstavlja n-dimenzionalni vektor jedinica. Iz činjenice da minimiziranjem 1.23 zapravo minimiziramo linearnu funkciju na poliedarskom skupu koji je opisan sa $(2n+1)$-dimenzionalnih hiperravnina, slijede neka lijepa svojstva. Jedno svojstvo je da $\min\{u_i, v_i\}$ mora biti nula za svaki i. Isto tako slijedi da za neki i možemo gledati rješenje oblika $\xi = y_i$.

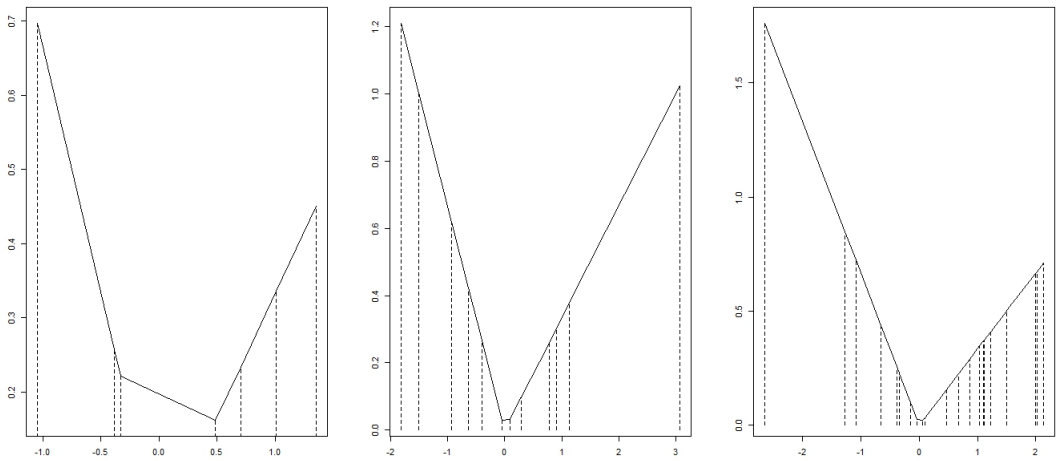
Označimo sa $R(\xi) = \sum_{i=1}^n \rho_\tau(y_i - \xi)$ kriterijsku funkciju. Optimalnost u ξ je zadovoljena ako vrijedi da je lijeva i desna derivacija od $R(\xi)$ nenegativna, tj. vrijedi:

$$\begin{aligned} R'(\xi+) &= \lim_{h\to 0} \frac{R(\xi+h) - R(\xi)}{h} \\ &= \sum_{i=1}^n \lim_{h\to 0} \frac{\rho_\tau(y_i - \xi - h) - \rho_\tau(y_i - \xi)}{h} \\ &= \sum_{i=1}^n (\mathbb{1}_{y_i \le \xi} - \tau) \ge 0 \end{aligned} \tag{1.24}$$

i

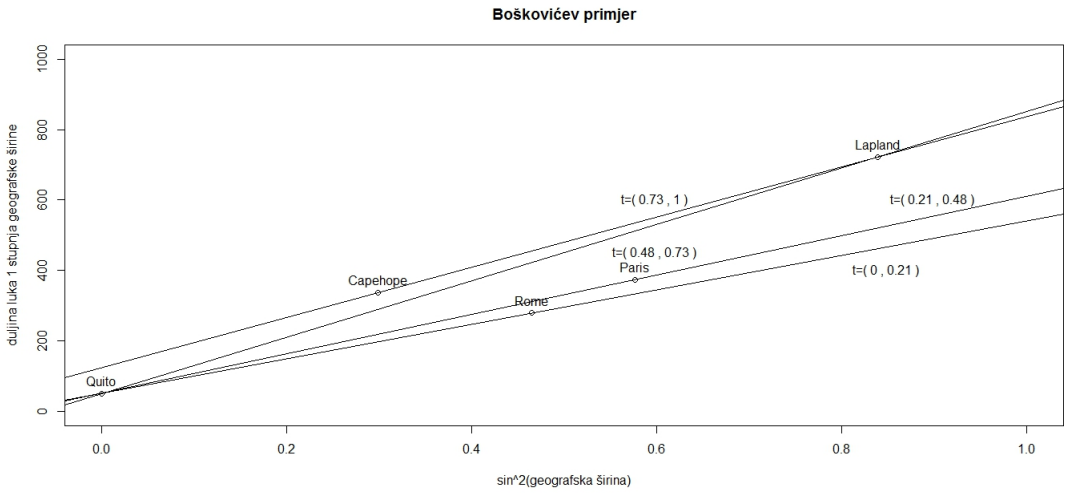
$$\begin{aligned} R'(\xi-) &= \lim_{h\to 0} \frac{R(\xi-h) - R(\xi)}{h} \\ &= \sum_{i=1}^n \lim_{h\to 0} \frac{\rho_\tau(y_i - \xi + h) - \rho_\tau(y_i - \xi)}{h} \\ &= \sum_{i=1}^n (\tau - \mathbb{1}_{y_i < \xi}) \ge 0. \end{aligned} \tag{1.25}$$

Vidimo da će to biti zadovoljeno ako $n\tau$ leži u intervalu $[N^-, N^+]$, gdje je $N^-(\xi) = \#\{y_i < \xi\}$ i $N^+(\xi) = \#\{y_i \le \xi\}$.



Slika 1.3: Kvantilna funkcija. Slika predstavlja funkciju gubitka za $\tau = 1/3$ na trima slučajnim uzorcima koji dolaze iz normalne razdiobe. Uzorci su veličine 7,12 i 23. Vertikalne linije predstavljaju točke opservacija. Na drugoj slici gdje je $n = 12$ vidimo da je minimum funkcije zapravo interval.

Iskoristimo sada Boškovićev primjer s početka kako bismo ilustrirali ideju kvantilne regresije.



Slika 1.4: Kvantilna regresija za Boškovićev primjer

Na slici 1.4 (dobivena pomoću R-a) prikazana su rješenja kvantilne regresije. Funkcija "rq" nam je dala 4 para od mogućih 10 kojih smo imali na početku problema. Rješavajući (1.22) za bilo koji τ na intervalu od (0,0.21) (koji je na slici označen kao t) dobivamo pravac koji prolazi kroz Quito i Rim. Zatim, za $\tau = 0.21$ naše rješenje skače u interval (0.21,0.48) i tada dobivamo pravac kroz Quito i Pariz. Analogno se dobivaju pravci kroz Quito i Laponiju i Quito i Rt Dobre Nade.

Tablica 1.4: U tablici su prikazani procijenjeni parametri za kvantilne pravce $y_i = \beta_0 + \beta_1 x_i$ gdje smo promatrali intervale kvantila

Interval kvantila	β_0	β_1
$[0, 0.21]$	51.0000	490.5336
$[0.21, 0.48]$	51.0000	560.5692
$[0.48, 0.73]$	51.0000	800.1431
$[0.73, 1]$	123.9985	713.0950

Poglavlje 2

Svojstva kvantilne regresije

2.1 Uvod u kvantilnu regresiju

Glavno pitanje na koje moramo dati odgovor je kako uopće kvantilna regresija funkcionira. Prisjetimo se prvo metode najmanjih kvadrata. Njena geometrijska interpretacija nam sugerira ideju minimiziranja euklidske udaljenosti $\|y - \hat{y}\|$ nad svim točkama $\hat{y}$, u linearnoj ljusci razapetim stupcima od X. Tada dobivamo elegantan oblik rješenja $\hat{y} = X(X'X)^{-1}X'y$ ukoliko je X punog ranga. Želeći zatim minimizirati po β:

$$\|y - \hat{y}(\beta)\|^2 = (y - X\beta)'(y - X\beta)$$

dobivamo izraz za stacionarnu točku:

$$\nabla_\beta \|y - \hat{y}(\beta)\|^2 = 2X'(y - X\beta) = 0.$$

pomoću kojeg dobijemo $\hat{\beta}$.
Sličan postupak koristimo i za kvantilnu regresiju. Euklidsku udaljenost zamijenimo sa:

$$d_\tau(y, \hat{y}) = \sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}_i). \tag{2.1}$$

Zatim želimo derivirati po $\beta \in \mathbb{R}^p$ kriterijsku funkciju:

$$R(\beta) = d_\tau(y, \hat{y}(\beta)) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta) \tag{2.2}$$

gdje je $\hat{y}(\beta) := (x_1'\beta, \ldots, x_n'\beta)$.

Vidimo da je funkcija derivabilna svugdje osim u onim točkama β za koje je $y_i = x_i'\beta$. Deriviranjem u smjeru vektora w tamo gdje je moguće, dobijemo sljedeći izraz:

$$\begin{aligned} \nabla R(\beta, w) &= \frac{d}{dt} R(\beta + tw)\,|_{t=0} \\ &= \frac{d}{dt} \sum_{i=1}^{n} (y_i - x_i'\beta - x_i' tw)[\tau - \mathbb{1}_{(y_i - x_i'\beta - x_i' tw)<0}]\,|_{t=0} \\ &= -\sum_{i=1}^{n} \psi^*(y_i - x_i'\beta, -x_i' w) x_i' w, \end{aligned} \tag{2.3}$$

gdje je

$$\psi^*(u, v) = \begin{cases} \tau - \mathbb{1}_{(u<0)} & \text{za} \quad u \neq 0 \\ \tau - \mathbb{1}_{(v<0)} & \text{za} \quad u = 0. \end{cases}$$

Ako je $\nabla R(\hat{\beta}, w) \geq 0$ za sve $w \epsilon \mathbb{R}^p$ koji zadovoljavaju $\|w\| = 1$, onda $\hat{\beta}$ minimizira $R(\beta)$.

Iz geometrijske interpretacije problema traženja točke $\hat{y} = X\hat{\beta}(\tau)$ koja je najbliža točki y uzimajući u obzir udaljenost d_τ, možemo uočiti važnu značajku od $\hat{\beta}(\tau)$, a to je da bi izbor optimalne točke $\hat{y}$ trebao sačuvati što više koordinata vektora reziduala $u(\beta) = y - X\beta$ jednakih nuli. Ako na primjer, procjenjujemo p parametara tj. $\beta \in \mathbb{R}^p$, tada se ne možemo nadati da će više od p koordinata vektora reziduala $u(\beta)$ biti jednakih nuli, ali nema razloga zašto bi tolerirali da bude i manje od p parametara jednakih nuli.

U kvantilnoj regresiji nam je cilj naći podskupove od p-opservacija koji će nam karakterizirati rješenje. Kako bismo si olakšali razumijevanje tih p-članih podskupova uvest ćemo neke oznake. Neka je $\mathcal{H}$ skup svih indeksa, tj. p-članih podskupova skupa $\mathcal{N} = \{1, 2, \ldots, n\}$ te neka za $h \in \mathcal{H}$, $X(h)$ predstavlja podmatricu od X čiji reci sadrže $\{x_i : i \in h\}$. Također, neka je $y(h)$ p-dimenzionalni vektor čije su koordinate $\{y_i : i \in h\}$. Komplement od h s obzirom na $\mathcal{N}$ označimo sa $\bar{h}$.

Uzimajući u obzir ove oznake, možemo zapisati rješenje koje prolazi točkama (x_i, y_i), $i \in h$ kao:

$$\beta(h) = X(h)^{-1} y(h), \tag{2.4}$$

gdje je matrica $X(h)$ regularna. Sada smo pripremili teren kako bismo mogli opisati glavna svojstva kvantilne regresije.

2.2 Uvjet podgradijenta

Prvo što moramo napraviti je proučiti rješenje

$$\beta(h) = X(h)^{-1}y(h). \tag{2.5}$$

Postavljamo si pitanje: što ako je za neki h, $X(h)$ singularna matrica. U tom slučaju možemo restringirati rješenje $\beta(h)$ na $h \in \mathcal{H}^* = \{h \in \mathcal{H} : |X(h)| \neq 0\}$, gdje je $|X(h)| = det(X(h))$. Vidimo i da uvjet optimalnosti povlači da su usmjerene derivacije nenegativne u svim smjerovima (zbog 1.24 i 1.25). Želimo provjeriti da $\beta(h)$ to zadovoljava, pa proučavamo:

$$\nabla R(\beta(h), w) = -\sum_{i=1}^{n} \psi_\tau^*(y_i - x_i\beta(h), -x_i w)x_i' w. \tag{2.6}$$

Stavimo da je $v = X(h)w$, pa imamo da je zadovoljen uvjet optimalnosti ako i samo ako vrijedi:

$$0 \leq -\sum_{i=1}^{n} \psi_\tau^*(y_i - x_i'\beta(h), -x_i'X(h)^{-1}v)x_i'X(h)^{-1}v, \text{ za sve } v \in \mathbb{R}^p. \tag{2.7}$$

Primijetimo da za $i \in h$ imamo i-ti jedinični vektor u $\mathbb{R}^p$ koji označimo sa $e_i' = x_i'X(h)^{-1}$ pa možemo gornju relaciju zapisati kao:

$$0 \leq -\sum_{i\epsilon h} \psi_\tau^*(0, -v_i)v_i - \xi' v = -\sum_{i\epsilon h}(\tau - \mathbb{1}_{(v_i<0)})v_i - \xi' v \tag{2.8}$$

gdje je:

$$\xi' = -\sum_{i\in \bar{h}} \psi_\tau^*(y_i - x_i'\beta(h), -x_i'X(h)^{-1}v)x_i'X(h)^{-1}. \tag{2.9}$$

Pod uvjetom da je $y_i - x_i'\beta(h) \neq 0$, za bilo koji $i \notin h$, uvjet usmjerene derivacije vrijedi za svaki $v \in \mathbb{R}^p$ ako i samo ako vrijedi za $v = \pm e_k$, $k = 1, 2, \ldots, p$. Prema tome, imamo p nejednakosti za $v = e_i$

$$0 < -(\tau - 1) + \xi_i(e_i) \text{ za } i = 1, 2, \ldots, p \tag{2.10}$$

dok za $v_i = -e_i$ imamo

$$0 < \tau - \xi_i(-e_i) \text{ za } i = 1, 2, \ldots, p. \tag{2.11}$$

Definicija 2.2.1. *Kažemo da su regresijske opservacije (y, X) u općem položaju ako za svaki p imamo jedinstvenu preciznu prilagodbu, tj. ako za bilo koji $h \in \mathcal{H}^*$ vrijedi*

$$y_i - x_i\beta(h) \neq 0 \textit{ za bilo koji } i \notin h.$$

Teorem 2.2.2. *Ako su (y,X) u općem položaju, onda postoji rješenje problema kvantilne regresije 1.22 oblika $\beta(h) = X(h)^{-1}y(h)$ ako i samo ako za neki $h \in \mathcal{H}^*$ vrijedi*

$$(\tau - 1)1_p \leq \xi_h \leq \tau 1_p, \tag{2.12}$$

gdje je $\xi_h = \sum_{i \in \bar{h}} \psi_\tau(y_i - x_i'\beta(h))x_i'X(h)^{-1}$ i $\psi_\tau(u) = \tau - \mathbb{1}_{(u<0)}$. Također, $\beta(h)$ je jedinstveno rješenje ako i samo ako vrijede stroge nejednakosti, inače je rješenje konveksna ljuska od nekoliko rješenja oblika $\beta(h)$.

Teorem 2.2.3. *Neka N^+, N^-, N^0 predstavljaju broj elemenata vektora reziduala $y - X'\hat{\beta}(\tau)$ koji su pozitivni, negativni i jednaki nula. Ako X sadrži konstantni član u regresijskoj jednadžbi tj. ako postoji $\alpha \in \mathbb{R}^p$ tako da je $X\alpha = 1_n$, onda za bilo koji $\hat{\beta}(\tau)$ koji zadovoljava 1.22 vrijedi*

$$N^- \leq n\tau \leq N^- + N^0 \tag{2.13}$$

i

$$N^+ \leq n(1 - \tau) \leq N^+ + N^0 \tag{2.14}$$

Dokaz. Znamo da je uvjet optimalnosti za $\hat{\beta}(\tau)$ zadovoljen ako i samo ako vrijedi (2.7) tj. akko vrijedi

$$-\sum_{i=1}^{n} \psi_\tau^*(y_i - x_i'\hat{\beta}(\tau), -x_i'w)x_i'w \geq 0, \ \forall w \in \mathbb{R}^p. \tag{2.15}$$

Sada stavimo da je $w = \alpha$, tako da je $X\alpha = 1_n$. Tada imamo

$$-\sum_{i=1}^{n} \psi_\tau^*(y_i - x_i'\hat{\beta}(\tau), -1) \geq 0. \tag{2.16}$$

To povlači da je

$$-\tau N^+ - (\tau - 1)N^- - (\tau - 1)N^0 \geq 0$$

$$\tau N^+ + (\tau - 1)N^- + (\tau - 1)N^0 \leq 0.$$

Budući da je $n = N^+ + N^+ + N^0$ imamo

$$n\tau \leq N^- + N^0,$$

tj. imamo

$$n(1 - \tau) \geq N^+.$$

Isto napravimo i za slučaj $w = -\alpha$. Tada vrijedi:

$$\sum_{i=1}^{n} \psi_\tau^*(y_i - x_i'\hat{\beta}(\tau), 1) \geq 0. \tag{2.17}$$

To povlači da je

$$\tau N^+ + (\tau - 1)N^- + \tau N^0 \geq 0$$
$$-\tau N^+ + (1 - \tau)N^- - \tau N^0 \leq 0$$
$$N^- \leq n\tau,$$

tj. imamo

$$n(1 - \tau) \leq N^+ + N^0.$$

Kombiniranjem dobivenih nejednakosti dobivamo (2.14), odnosno (2.13). □

Kao posljedicu teorema dobivamo sljedeći korolar:

Korolar 2.2.4. *Ako je $N^0 = p$, tada je omjer negativnih reziduala približno τ*

$$\frac{N^-}{n} \leq \tau \leq \frac{N^- + p}{n},$$

a omjer pozitivnih reziduala je približno $(1 - \tau)$

$$\frac{N^+}{n} \leq 1 - \tau \leq \frac{N^+ + p}{n}.$$

Napomena 2.2.5. *Kada imamo specijalni slučaj da je $X \equiv 1_n$, onda ovaj rezultat opisuje τ-ti uzorački kvantil. Ako je $n\tau$ cijeli broj, tada imamo interval τ-tih uzoračkih kvantila, inače imamo jedinstveni uzorački kvantil.*

Korolar 2.2.6. *Pretpostavimo da imamo dvodimenzionalni uzorak gdje je X oblika*

$$X(h) = \begin{pmatrix} 1_{n_1} & 0 \\ 0 & 1_{n_2} \end{pmatrix} \tag{2.18}$$

i stavimo da je $y = (y_1', y_2')$ u skladu sa X. Označimo sa $\hat{\beta}_i(\tau)$ bilo koji τ-ti uzorački kvantil poduzorka y_i. Tada bilo koje rješenje kvantilne regresije za ovaj problem ima oblik

$$\hat{\beta}(\tau) = (\hat{\beta}_1(\tau), \hat{\beta}_2(\tau))', \tag{2.19}$$

tj. pravac koji opisuje τ-to rješenje regresijskog kvantila, kod problema s dva uzorka, jednostavno spaja dva odgovarajuća uobičajena uzoračka kvantila za ta dva uzorka.

Dokaz. Dokaz slijedi direktno iz činjenice da se uvjet optimalnosti

$$-\sum_{i=1}^{n} \psi^*(y_i - b, -x_i'w)x_i'w \leq 0,$$

za $b \in \mathbb{R}^2$ i $w \in \mathbb{R}^2$ može razdvojiti u dva zasebna uvjeta,

$$-\sum_{i=1}^{n_j} \psi^*(y_{i_j} - b_j, -w_j)w_j \leq 0, \;\; j = 1, 2$$

gdje je y_{i_j} j-ti element od y_i. □

2.3 Ekvivarijanca

Neka svojstva metode najmanjih kvadrata uzimamo zdravo za gotovo, ali ona imaju važnu ulogu u interpretaciji rezultata regresije. Na primjer, pretpostavimo da imamo podatke koji su izmjereni u miljama po satu, te ih želimo prilagoditi SI sustavu tj. želimo ih pretvoriti u metre po sekundi. Očekujemo da takve promjene neće imati utjecaja na naše procijene, tj. očekujemo da će se i naše procijenjene vrijednosti ponašati na isti način. Drugim riječima očekujemo da ćemo na kraju analize podataka transformacijom moći doći do milja po satu i da će interpretacija rezultata biti ista. Procjenitelji kvantilne regresije imaju neka takva svojstva, koja ćemo nazivati ekvivarijancama. Ta svojstva će nam biti važna prilikom interpretiranja statističkih rezultata. Kako bismo si olakšali postupak, označit ćemo τ-ti regresijski kvantil koji se odnosi na opservaciju (y, X) sa $\hat{\beta}(\tau; y, X)$. Glavna svojstva ekvivarijance su opisana u sljedećem teoremu:

Teorem 2.3.1. *(Koenker i Basset) Neka je A p-dimenzionalna regularna matrica, $\gamma \in \mathbb{R}^p$ i $a > 0$. Tada za svaki $\tau \in [0, 1]$ vrijedi:*

(i) $\hat{\beta}(\tau; ay, X) = a\hat{\beta}(\tau; y, X)$ *(skaliranje ekvivarijance)*

(ii) $\hat{\beta}(\tau; -ay, X) = a\hat{\beta}(1 - \tau; y, X)$ *(skaliranje ekvivarijance)*

(iii) $\hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y, X) + \gamma$ *(regresijska ekvivarijanca)*

(iv) $\hat{\beta}(\tau; y, XA) = A^{-1}\hat{\beta}(\tau; y, X)$ *(ekvivarijanca reparametrizirane matrice dizajna)*

Postoji još jedno svojstvo ekvivarijance, koje nazivamo ekvivarijanca na monotone transformacije. Neka je h strogo rastuća funkcija na $\mathbb{R}$. Tada za neku slučajnu varijablu Y vrijedi

$$Q_{h(Y)}(\tau) = h(Q_Y(\tau)), \tag{2.20}$$

tj. kvantili transformirane slučajne varijable $h(Y)$ su transformirani kvantili originalne slučajne varijable. Matematičko očekivanje ne dijeli to svojstvo $\mathbb{E}[h(Y)] \neq h(\mathbb{E}[Y])$, osim za afine funkcije h. Svojstvo (2.20) direktno slijedi iz činjenice da za bilo koju monotonu strogo rastuću funkciju h, vrijedi:

$$\mathbb{P}(Y \leq y) = \mathbb{P}(h(Y) \leq h(y)). \tag{2.21}$$

Transformacije u kvantilnoj regresiji su jednostavnije za protumačiti nego u metodi najmanjih kvadrata. Ako označimo sa $x'\hat{\beta}$ procjenitelj uvjetnog medijana od $h(y)$ uz uvjet x, tada zbog svojstva ekvivarijance možemo staviti da je $h^{-1}(x'\hat{\beta})$ odgovarajući procjenitelj uvjetnog medijana od y uz uvjet x.

2.4 Robustnost

Najbitnije svojstvo kvantilne regresije je svojstvo robustnosti. Funkcija utjecaja mjeri kako na procjenitelj $\hat{\theta}$, procijenjen pomoću funkcije distribucije F, utječu neke perturbacije od F. To jest, možemo gledati na $\hat{\theta}$ kao funkcional od F, $\hat{\theta}(F)$. Kontaminiranu funkciju distribucije možemo zapisati kao

$$F_\varepsilon = \varepsilon\delta_y + (1-\varepsilon)F. \tag{2.22}$$

Sada možemo izraziti funkciju utjecaja od $\hat{\theta}$ sa

$$IF_{\hat{\theta}}(y, F) = \lim_{\varepsilon\to 0} \frac{\hat{\theta}(F_\varepsilon) - \hat{\theta}(F)}{\varepsilon}. \tag{2.23}$$

Za matematičko očekivanje imamo da je

$$\hat{\theta}(F_\varepsilon) = \int y dF_\varepsilon = \varepsilon y + (1-\varepsilon)\hat{\theta}(F). \tag{2.24}$$

Kada to uvrstimo u 2.23 dobijemo da je funkcija utjecaja za matematičko očekivanje jednaka

$$\begin{aligned} IF_{\hat{\theta}}(y, F) &= \lim_{\varepsilon\to 0} \frac{\varepsilon y + (1-\varepsilon)\hat{\theta}(F) - \hat{\theta}(F)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{\varepsilon(y + \hat{\theta}(F)}{\varepsilon} \\ &= y - \hat{\theta}(F). \end{aligned} \tag{2.25}$$

Za medijan imamo da je

$$m_\varepsilon = \widetilde{\theta}(F_\varepsilon) = F_\varepsilon^{-1}(1/2),$$

tj. imamo da je

$$\frac{1}{2} = \varepsilon\delta_y(m_\varepsilon) + (1-\varepsilon)F(m_\varepsilon) \tag{2.26}$$

Kada to raspišemo za slučaj kada je $m_\varepsilon < x$ i za slučaj kada je $m_\varepsilon \geq x$ dobijemo da je

$$\begin{aligned} \frac{F_\varepsilon^{-1}(1/2) - F^{-1}(1/2)}{\varepsilon} &= \frac{F^{-1}(1/2(1-\varepsilon)) - F^{-1}(1/2)}{\varepsilon}\mathbb{1}_{\{m_\varepsilon < x\}} \\ &+ \frac{F^{-1}((1-2\varepsilon)/2(1-\varepsilon)) - F^{-1}(1/2)}{\varepsilon}\mathbb{1}_{\{m_\varepsilon \geq x\}}. \end{aligned}$$

Kada pustimo ε u 0 dobijemo da je

$$IF_{\widetilde{\theta}}(y, F) = \frac{sgn(y - \widetilde{\theta}(F))}{f(F^{-1}(1/2))}. \tag{2.27}$$

uz pretpostavku da funkcija gustoće f postoji i da je pozitivna. Funkcija utjecaja za τ-ti kvantil se dobije tako da se u formulu (2.27) umjesto $1/2$ uvrsti τ. Ograničenost funkcije utjecaja je trivijalno zadovoljeno činjenicom da je τ konačan. Prije nego zapišemo kako izgleda funkcija IF moramo zapisati F tako da predstavlja distribuciju od (x, y). Dobivamo da je

$$dF = dG(x)f(y|x)dy.$$

Uz pretpostavku da je funkcija gustoće f neprekidna i strogo pozitivna imamo da je

$$IF_{\hat{\beta}_F(\tau)}((y, x), F) = Q^{-1}x \cdot sgn(y - x'\hat{\beta}_F(\tau)) \tag{2.28}$$

gdje je

$$Q = \int xx' f(x'\hat{\beta}_F(x))dG(x). \tag{2.29}$$

Robustnost procijenitelja kvantilne regresije na outliere od y se može opisati na sljedeći način. Zamislimo podatke kroz koje prolazi τ-ti regresijski pravac. Sada pretpostavimo da uzmemo neki y_i s pravca i krenemo ga micati sve dalje i dalje od pravca u smjeru y osi. Postavlja se pitanje kako takvo ponašanje točke utječe na prilagođeni pravac. Uvjet podgradijenta nam govori da možemo micati točku y_i gore dolje uz uvjet da ne prijeđemo prilagođeni pravac (bez mijenjanja strane). To objašnjava sljedeća svojstva:

(i) funkcija utjecaja je konstantna iznad prilagođenog kvantila

(ii) Opservacije se ne zanemaruju, sve opservacije jednako sudjeluju u izboru reprezentativnih točaka.

Teorem 2.4.1. *Neka je D dijagonalna matrica s nenegativnim elementima d_i, za $i = 1, \ldots, n$. Tada je*

$$\hat{\beta}(\tau; y, X) = \hat{\beta}(\tau; X\hat{\beta}(\tau; y, X) + D(y - X\hat{\beta}(\tau; y, X)), X) \tag{2.30}$$

Teorem nam zapravo kaže da dokle god se ne mijenja predznak reziduala možemo mijenjati bilo koju opservaciju y bez da to ima utjecaja na početno rješenje. To je važno svojstvo kvantilne regresije. Funkcija utjecaja je nezamjenjiv alat, koji je dizajniran za mjerenje osjetljivosti procjenitelja na beskonačne perturbacije nominalnog modela. No, procedure mogu biti beskonačno robusne, ali i dalje vrlo osjetljive na male konačne perturbacije.

Promotrimo još asimptotsko ponašanje procijenitelja metoda najmanjih apsolutnih odstupanja. Za

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^2} \sum |y_i - \theta_1 x_i - \theta_2| \tag{2.31}$$

uz neke uvjete (vidjeti [3]) vrijedi

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} \mathcal{N}(0, \omega^2 D^{-1}), \tag{2.32}$$

gdje je $\theta_0 = (\theta_1, \theta_2)'$, $\omega^2 = 1/(4f^2(0))$ (za funkciju gustoće f koja je neprekidna i strogo pozitivna u okolini točke 0) i

$$\lim_{n\to\infty} n^{-1}X'X \to D, \text{ za } X = (x_i, 1)_{i=1}^n \text{ i } D \text{ pozitivnu definitnu matricu.} \tag{2.33}$$

Na sličan način možemo naći i asimptotsko ponašanje za Boškovićev procjenitelj. Promotrimo klasičan model linearne regresije

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + u_i = x_i\beta + u_i, \tag{2.34}$$

gdje su u_i, $i = 1, \ldots, n, \ldots$ nezavisne slučajne varijable sa zajedničkom funkcijom distribucije F, za koju vrijedi $F(1/2) = 0$, $\mathbb{E}[u] = \mu$ i sa funkcijom gustoće f koja je neprekidna i pozitivna u okolini točaka 0 i μ. Pretpostavimo još da vrijedi

$$\sigma^2 = \mathbb{E}(u - \mu)^2 < \infty, \tag{2.35}$$

da dizajn matrica sadrži slobodni član ($x_{1j} = 1, \forall j$) i da zadovoljava uvjet

$$\lim_{n\to\infty} \frac{1}{n} X'X \to D, \tag{2.36}$$

gdje je D pozitivno definitna matrica.
Može se pokazati da za $\mu = 0$ vrijedi (vidjeti: [2])

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} \mathcal{N}(0, \omega_0^{-2}(D^{-1} - E_1) + \sigma^2 E_1), \tag{2.37}$$

gdje je $\omega_0 = 2f(0)$ i E_1 je $p \times p$ matrica sa 1 na $(1, 1)$ elementu i 0 ostalo.

2.5 Ukratko o Waldovu testu

Kao i u linearnoj regresiji tako i u kvantilnoj regresiji postoje neki testovi koji se koriste. Ukratko ćemo opisati Waldov test. On se koristi za ispitivanje paralelnosti kvantilnih regresijskih pravaca. Nulta hipoteza je da su koeficijenti smjera pravaca jednaki. Odnosno, ako imamo model od dva uzorka $Y_i = \alpha_1 + \alpha_2 x_i + u_i$, gdje je $x_i = 0$ za n_1 opservacija prvog uzorka i $x_i = 1$ za n_2 opservacija drugog uzorka, tada je procjena koeficjenta smjera α_2 τ-tog kvantilnog regresijskog pravca zapravo razlika između τ-tog uzoračkog kvantila oba uzorka. Prema tome, test za jednakost koeficjenta smjera τ_1 i τ_2 kvantilnog regresijskog pravca se može zapisati kao

$$\begin{aligned} \alpha_2(\tau_2) - \alpha_1(\tau_1) &= (Q_2(\tau_2) - Q_1(\tau_2)) - (Q_2(\tau_1) - Q_1(\tau_1)) \\ &= (Q_2(\tau_2) - Q_2(\tau_1)) - (Q_1(\tau_2) - Q_1(\tau_1)). \end{aligned} \tag{2.38}$$

Sada nultu hipotezu možemo zapisati kao $H_0 : \alpha_2(\tau_2) - \alpha_1(\tau_1) = 0$. Ako zapišemo asimptotsku varijancu od $\hat{\alpha_2}(\tau_2) - \hat{\alpha_1}(\tau_1)$ kao

$$\sigma^2(\tau_1, \tau_2) = \left[\frac{\tau_1(1-\tau_1)}{f^2(F^{-1}(\tau_1))} - 2\frac{\tau_1(1-\tau_2)}{f(F^{-1}(\tau_1))f(F^{-1}(\tau_2))} + \frac{\tau_2(1-\tau_2)}{f^2(F^{-1}(\tau_2))} \right] \left[\frac{n}{nn_1 - n_1^2} \right], \tag{2.39}$$

tada je testna statistika

$$T_n = (\hat{\alpha}_2(\tau_2) - \hat{\alpha}_1(\tau_1))/(\hat{\sigma}(\tau_1, \tau_2)). \tag{2.40}$$

Pokazuje se da je nul-distribucija od T_n (tj. razdioba od T_n uz istinitost od H_0) jednaka jediničnoj normalnoj distribuciji $N(0, 1)$.

Poglavlje 3

Interpretacija kvantilne regresije i primjeri

3.1 Efekt tretmana

Najjednostavniji primjer regresije je model dva uzorka za kontrolu i tretman. Počet ćemo sa modelom koji su uveli Lehmann i Doksum 1970-tih godina, a poslužit će kao dobar uvod za interpretaciju općenitih modela kvantilne regresije. Lehmann (1974) je predložio sljedeći model za kontrolu tretmana:

> Pretpostavimo da tretman dodaje iznos $\Delta(x)$ u slučaju kada je objekt koji nije tretiran x. Tada je funkcija distribucije tretmana G jednaka funkciji distribucije slučajne varijable $X + \Delta(X)$, gdje X ima funkciju distribucije F.

Specijalni slučajevi koriste model pomaka lokacije $\Delta(X) = \Delta_0$ i model pomaka skale $\Delta(X) = \Delta_0 \cdot X$. Ako vrijedi da je

$$\Delta(x) \geq 0 \ \forall x, \tag{3.1}$$

tada je funkcija distribucije tretmana G stohastički veća od funkcije distribucije kontrole F. Doksum (1974) je pokazao da ako definiramo $\Delta(x)$ kao horizontalnu udaljenost između F i G u točki x tako da vrijedi

$$F(x) = G(x + \Delta(x)) \tag{3.2}$$

tada je $\Delta(x)$ jedinstveno određena i možemo ju zapisati kao

$$\Delta(x) = G^{-1}(F(x)) - x. \tag{3.3}$$

Ako zamijenimo $F(x)$ sa τ, tada dobijemo kvantilni efekt tretmana, tj. imamo

$$\delta(\tau) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau). \tag{3.4}$$

Prirodno nam je sada zapisati kvantilni efekt tretmana kao

$$\hat{\delta}(\tau) = \hat{G_n}^{-1}(\tau) - \hat{F_m}^{-1}(\tau). \tag{3.5}$$

gdje G_n i F_m predstavljaju empirijske funkcije distribucije za tretman i kontrolu, koji se temelji na n odnosno m opservacija. Ako zapišemo model kvantilne regresije za problem tretmana dva uzorka kao

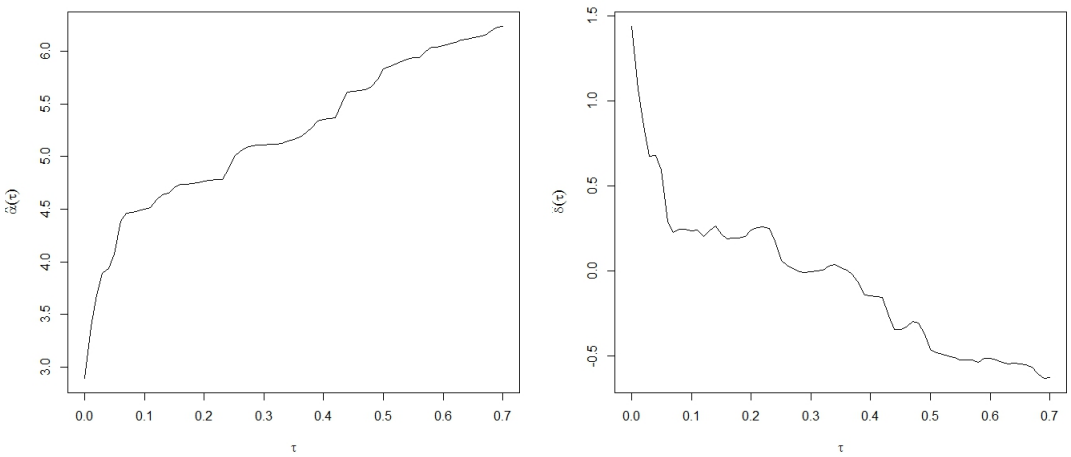
$$Q_{Y_i}(\tau|D_i) = \alpha(\tau)(1 - D_i) + \beta(\tau)D_i \tag{3.6}$$

gdje D_i predstavlja indikator tretmana ($D_i = 1$ označuje tretman, a $D_i = 0$ označuje kontrolu), tada dobivamo procjenitelje $\hat{\alpha}(\tau) = \hat{F_m^{-1}}(\tau)$ i $\hat{\beta}(\tau) = \hat{G_n^{-1}}(\tau)$. Ako zapišemo model kao

$$Q_{Y_i}(\tau|D_i) = \alpha(\tau) + \delta(\tau)D_i \tag{3.7}$$

tada možemo direktno procijeniti kvantilni efekt tretmana.

Primjer 3.1.1. *Kako bismo ilustrirali kvantilni efekt tretmana, koristimo podatke iz 1960. od Bjerkedala koji je promatrao utjecaj "tubercle bacilli" na zamorcima. Bjerkedal je zabilježio vrijeme doživljenja zamoraca u danima i to tijekom 736 dana. Postojala je kontrolna grupa koja se sastojala od 107 subjekata i grupa koja je primala tretman koja se sastojala od 60 subjekata. U kontrolnoj grupi 42 subjekata je živjelo dulje od predviđenog eksperimenta (736 dana).*



Slika 3.1: Lijevi graf prikazuje funkciju $\hat{\alpha}(\tau)$ koja predstavlja empirijsku funkciju kvantila za logaritam vremena doživljenja za kontrolnu grupu. Desni graf prikazuje $\hat{\delta}(\tau)$ koja predstavlja kvantilni efekt tretmana. Odnosno to prikazuje horizontalnu razliku između empirijskih funkcija kontrolne grupe i grupe koja je primila tretman.

Na grafovima su prikazani necenzurirani podaci (uklonjeni su oni zamorci koji su živjeli duže od eksperimenta). Na desnom grafu (3.1) možemo iščitati iz lijevog repa distribucije da tretman ima pozitivan učinak na zamorce, dok iz desnog repa distribucije iščitavamo da tretman ima negativan učinak na vrijeme doživljenja zamoraca. Možemo zaključiti da tretman ima pozitivan učinak na kraćem roku, ali se čini vrlo nepovoljan za duži rok.

Možemo interpretirati kontrolne subjekte u smislu skrivenih (latentnih) karakteristika. Možemo reći da su kontrolni subjekti koji su skloni umiranju u ranoj dobi krhki, a oni koji su skloni umiranju kasnije robusni. Ova karakteristika implicitno je indeksirana pomoću τ odnosno pomoću kvantila distribucije doživljenja na kojem se pojavljuje subjekt koji nije tretiran tj. $\alpha(\tau)$. Smatramo da odaziv tretmana u odnosu na kontrolnu grupu $\alpha(\tau)$ je jednak $\alpha(\tau) + \delta(\tau)$. Ako je skrivena karakteristika, recimo sklonost dugovječnosti, vidljiva "ex ante", tada možemo razmatrati efekt tretmana $\delta(\tau)$ kao eksplicitnu interakciju sa vidljivom varijablom. Međutim, u odsustvu takve vidljive varijable, kvantilni efekt tretmana može se smatrati kao prirodna mjera odaziva tretmana. Naravno, ne postoji način na koji bi mogli saznati djeluje li tretman stvarno na način propisan pomoću $\delta(\tau)$. Zapravo, tretman može utjecati na krhke subjekte tako da postanu izrazito robustni, a na jake subjekte tako da postanu krhki. Iz eksperimenata možemo jedino vidjeti razliku između distribucija doživljenja, i zato je prirodno povezati efekt tretmana s razlikom između kvantila tih dviju distribucija. I to je upravo ono što radi kvantilni efekt tretmana.

Ako imamo više tretmana tada trebamo prilagoditi interpretaciju. U slučaju kada imamo p različitih tretmana model kvantilne regresije za problem tretmana možemo zapisati kao:

$$Q_{Y_i}(\tau|D_{ij}) = \alpha(\tau) + \sum_{j=1}^{p} \delta_j(\tau) D_{ij}, \tag{3.8}$$

gdje je $D_{ij} = 1$ ako je i-ta opservacija dobila j-ti tretman, a $D_{ij} = 0$ inače. U ovom slučaju $\delta_j(\tau)$ predstavlja kvantilni efekt tretmana koji povezuje distribuciju odaziva kontrole s odazivom varijable koje je pod utjecajem tretmana j. U slučaju kada imamo kontinuirani tretman (npr. studije doza-odgovor) prirodno je pretpostaviti da je i efekt (odnosno utjecaj tretmana) linearan. To možemo zapisati kao:

$$Q_{Y_i}(\tau|x_i) = \alpha(\tau) + \beta(\tau) x_i. \tag{3.9}$$

Pretpostavljamo da je tretman efekta $\beta(\tau)$ koji predstavlja promjenu varijable x iz x_0 u x_0+1 jednak tretmanu efekta koji predstavlja promjenu varijable x iz x_1 u $x_1 + 1$.

Zanimljivo je to da se kvantilni efekt tretmana (3.3) može povezati sa tradicionalnim QQ-grafom za dva uzorka. Odnosno, funkcija $\hat{\Delta}(x) = G_n^{-1}(F_m(x)) - x$ predstavlja točno ono što prikazuje tradicionalni QQ-graf za dva uzorka.

3.2 Interpretacija modela kvantilne regresije

Poznat nam je postupak traženja parametra β u linearnom regresijskom modelu. U modelu

$$\mathbb{E}(Y|X = x) = x'\beta, \tag{3.10}$$

koeficijent β možemo interpretirati u terminima parcijalne derivacije, odnosno imamo

$$\frac{\partial \mathbb{E}(Y|X = x)}{\partial x_j} = \beta_j. \tag{3.11}$$

Postoje i neki problemi u interpretaciji. Na primjer, možemo imati više koeficijenata koji su vezani uz jednu kovarijancu u kvadratnom modelu ili u modelu sa interakcijom. U tom slučaju promjene kovarijance utječu na promjene nekoliko koordinata vektora x, pa i derivacije moraju biti u skladu s tim promjenama. Pretpostavimo da imamo model

$$\mathbb{E}(Y|Z = z) = \beta_0 + \beta_1 z + \beta_2 z^2. \tag{3.12}$$

Očito je

$$\frac{\partial \mathbb{E}(Y|Z = z)}{\partial z} = \beta_1 + 2\beta_2 z, \tag{3.13}$$

što povlači da utjecaj promjene varijable z u uvjetnom očekivanju od Y ovisi o β_1 i β_2, ali ovisi i o vrijednosti varijable z. U slučaju transformacijskog modela

$$\mathbb{E}(h(Y)|X = x) = x'\beta, \tag{3.14}$$

želja bi nam bila zapisati derivaciju kao

$$\frac{\partial \mathbb{E}(Y|X = x)}{\partial x_j} = \frac{\partial h^{-1}(x'\beta)}{\partial x_j}. \tag{3.15}$$

No, kako bi Nixon rekao: "Možeš to napraviti, ali bi bilo krivo" (iako se to često radi u slučaju logaritamske funkcije gdje je $h(Y) = log(Y)$). Problem je u tome što je $\mathbb{E}(h(Y)) \neq h(\mathbb{E}(Y))$ osim za afine funkcije h. To čini interpretaciju regresije mnogo težom u praksi.

Situacija je nešto jednostavnija u slučaju kvantilne regresije zato što je $Q_{h(Y)}(\tau|X = x) = h(Q_Y(\tau|X = x))$ za bilo koju strogo rastuću monotonu transformaciju h. Iz toga direktno slijedi da ako imamo

$$Q_{h(Y)}(\tau|X = x) = x'\beta(\tau), \tag{3.16}$$

tada imamo i

$$\frac{\partial Q_Y(\tau|X = x)}{\partial x_j} = \frac{\partial h^{-1}(x'\beta)}{\partial x_j}. \tag{3.17}$$

Na primjer, ako uzmemo da je

$$Q_{log(Y)}(\tau|X = x) = x'\beta(\tau), \tag{3.18}$$

tada imamo da je

$$\frac{\partial Q_Y(\tau|X = x)}{\partial x_j} = \exp(x'\beta)\beta_j. \tag{3.19}$$

3.3 Interpretacija pogrešno postavljenog modela kvantilne regresije

Kod klasične metode najmanjih kvadrata ako promatramo model

$$y_i = \theta(z_i) + u_i \tag{3.20}$$

gdje su $\{u_i\}$ nezavisne jednako distribuirane varijable sa funkcijom distribucije F i $\mathbb{E}(u_i) = 0$, ali pogrešno procijenimo linearni model

$$y_i = \beta_0 + \beta_1 z_i + v_i, \tag{3.21}$$

tada procjenitelj metode najmanjih kvadrata $\hat{\beta} = (X'X)^{-1}X'y$, gdje je $X = (x_i) = (1, z_i)$ ima svojstvo da je

$$\begin{aligned} \tilde{\beta} &= \mathbb{E}\hat{\beta} \\ &= \mathbb{E}\left[(X'X)^{-1}X'\left(\begin{bmatrix} \theta(z_1) \\ \vdots \\ \theta(z_n) \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}\right)\right] \\ &= (X'X)^{-1}X'\begin{bmatrix} \theta(z_1) \\ \vdots \\ \theta(z_n) \end{bmatrix} + 0. \end{aligned} \tag{3.22}$$

Iz toga slijedi da je $X\tilde{\beta}$ ortogonalna projekcija od Θ, gdje je $\Theta = \begin{bmatrix} \theta(z_1) \dots \theta(z_n) \end{bmatrix}'$. Očito je $\tilde{\beta}$ ovisna o točkama dizajna $\{x_i\}$ budući da minimizira $\sum(\theta(z_i) - x_i'\beta)^2$.

U kvantilnoj regresiji analiza posljedica pogrešno postavljenog modela je puno kompliciranija. Treba imati na umu da

$$\hat{\beta}(\tau) = \arg\min \sum \rho_\tau(y_i - x_i'\beta), \tag{3.23}$$

aproksimativno rješava jednadžbu $\Psi(\beta) = 0$, gdje je

$$\Psi(\beta) = n^{-1} \sum \psi_\tau(y_i - x_i'\beta)x_i. \tag{3.24}$$

To možemo zapisati kao

$$\begin{aligned} \Psi(\beta) &= n^{-1} \sum \psi_\tau(u_i + \theta(x_i) - x_i'\beta)x_i & (3.25) \\ &= n^{-1} \sum (\tau - \mathbb{1}_{(u_i+\theta(x_i)-x_i'\beta)<0})x_i. & (3.26) \end{aligned}$$

Može se pokazati da je tada

$$\mathbb{E}(\Psi(\beta)) = n^{-1} \sum (\tau - F(x_i'\beta - \theta(x_i)))x_i. \tag{3.27}$$

Prema tome, vidimo da rješenje $\tilde{\beta}(\tau)$ jednadžbe $\mathbb{E}(\Psi(\beta)) = 0$ ne ovisi samo o funkciji θ i o $\{x_e\}$ kao što je u slučaju metode najmanjih kvadrata, nego ovisi i o funkciji distribucije F.

Zanimljiv je slučaj kada su $\{u_i\}$ uniformno distribuirane. Tada imamo da je

$$\mathbb{E}(\Psi(\beta)) = n^{-1} \sum (\tau + \theta(x_i) - x_i'\beta)x_i. \tag{3.28}$$

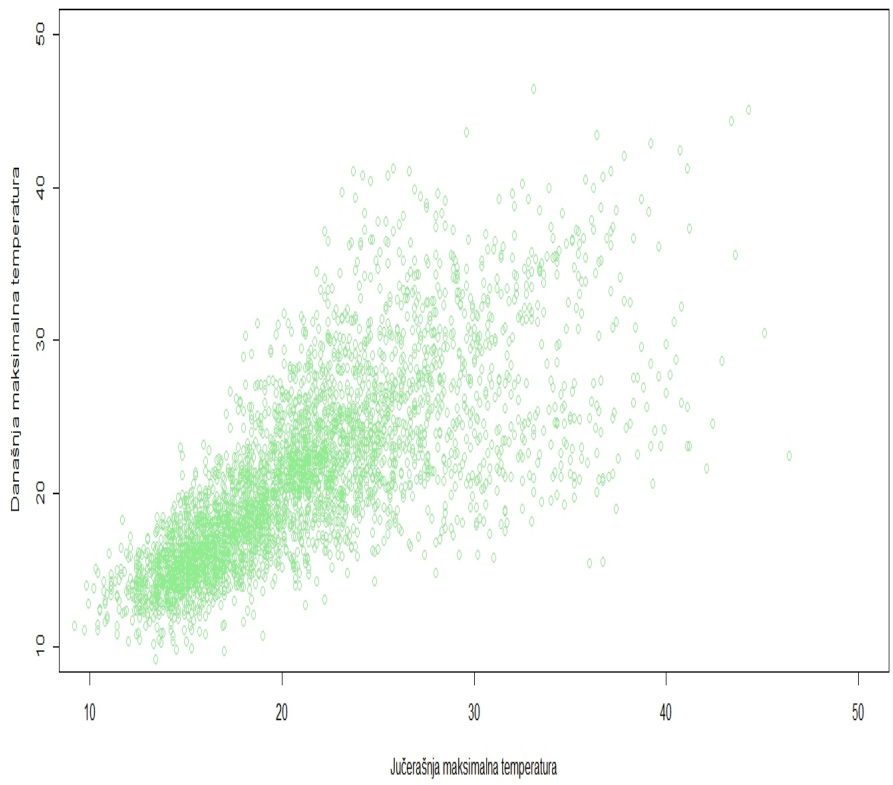
Rješenje $\beta(\tau)$ možemo eksplicitno zapisati kao

$$\beta(\tau) = (\sum x_i x_i')^{-1} \sum x_i'(\theta(x_i) + \tau) = (X'X)^{-1}X'(\theta + \mathbb{1}\tau). \tag{3.29}$$

Budući da X eksplicitno sadrži slobodni član, efekt od τ se pojavljuje samo kao slobodni član od $\beta(\tau)$, dok je koeficijent nagiba isti kao kod metode najmanjih kvadrata.

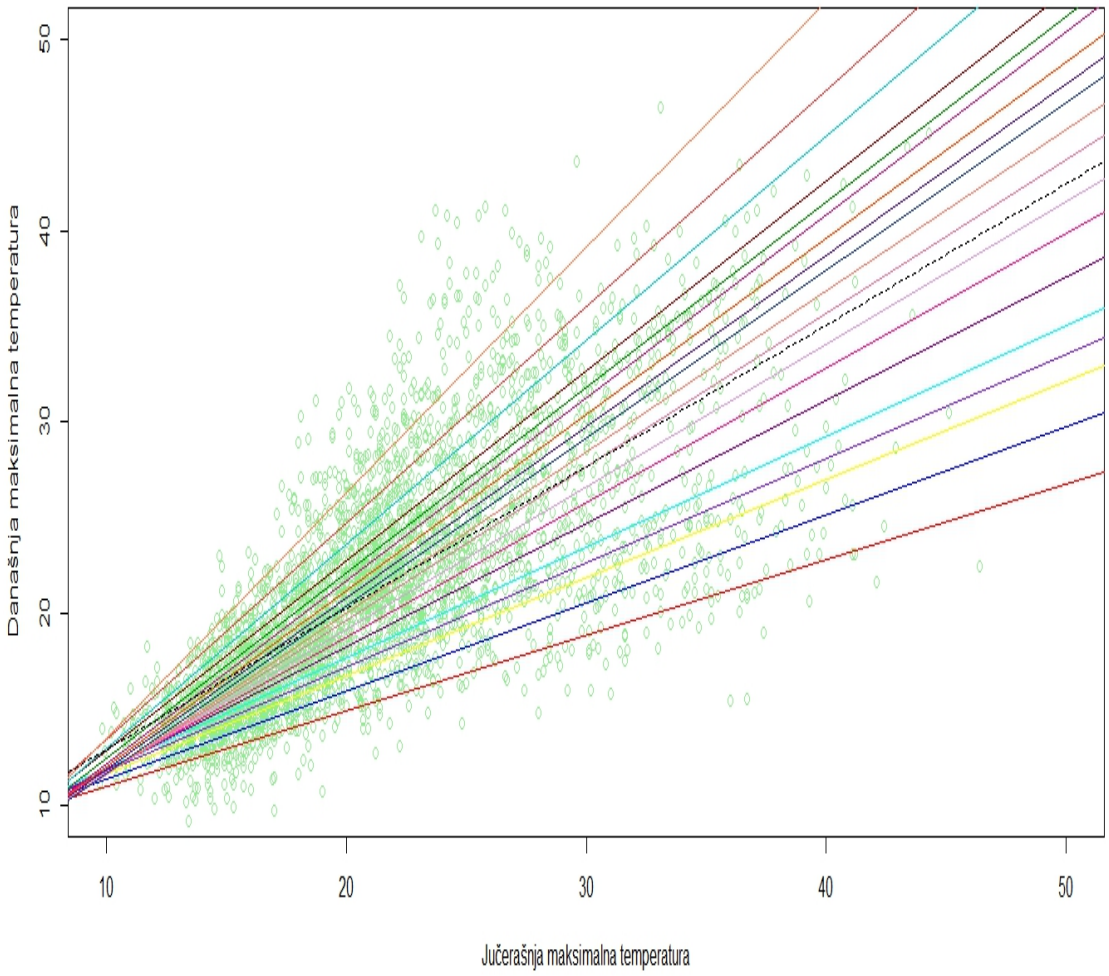
3.4 Primjeri

Primjer 3.4.1. *U ovom primjeru ćemo proučavati AR(1) model za dnevnu temperaturu u Melbournu, Australija (podaci su od 1.1.2004 do 1.1.2014 godine). Prvo što moramo pogledati je graf današnje maksimalne temperature nasuprot jučerašnje maksimalne temperature.*



Na prvi pogled izgleda kao da je današnja maksimalna temperatura jednaka jučerašnjoj. Taj dojam se stekne zbog lijevog djela grafa, koji predstavlja niže temperature. Tamo se podaci grupiraju oko pravca $x = y$. No, ako pogledamo desni dio grafa, koji odgovara ljetnom razdoblju, vidjet ćemo da tamo nije ista situacija. Kod visokih temperatura možemo primijetiti da će sljedeći dan temperatura biti ili jednaka (tj. prati pravac $x = y$) ili znatno manja (malo manja temperatura se događa u rijetkim situacijama). U terminima uvjetne funkcije gustoće to bi značilo da ako je danas vruće, sutrašnja temperatura će pratiti bi-

modalnu distribuciju gdje gdje je jedan vrh funkcije centriran oko današnjeg maksimuma, a drugi vrh je centriran oko $20°$.



Na grafu iznad vidimo procijenjene kvantilne funckije za $\tau = (0.05, 0.1, \ldots, 0.9, 0.95)$ *i procijenjenu linearnu funkciju (crna isprekidana linija). Možemo uočiti da su razmaci između procijenjenih funkcija manji kod nižih temperatura te kako se temperatura povećava tako se povećavaju i razmaci. Još se može uočiti da su razmaci najveći kada je današnja temperatura otprilike* $20°$.

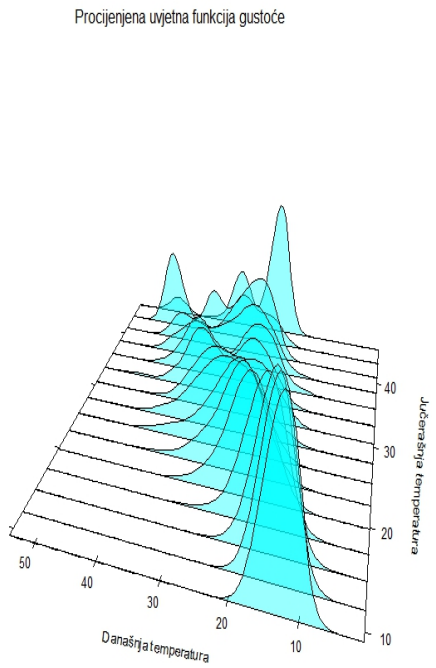
Tablica 3.1: U tablici su prikazani procijenjeni parametri za kvantilne pravce $y_i = \beta_0 + \beta_1 x_i$, njihova standardna greška i p-vrijednost dobivena Waldovim testom

Kvantili	β_0	Std.Error	β_1	Std. Error	p-vrijednost Waldovog testa
0.05	7.08485	0.31526	0.39394	0.01681	<0.0001
0.10	6.78852	0.29132	0.45902	0.01665	<0.0001
0.15	6.48348	0.29517	0.51304	0.01633	<0.0001
0.20	6.34966	0.24367	0.54362	0.01323	<0.0001
0.25	6.13090	0.24512	0.57865	0.01526	<0.0001
0.30	5.37111	0.22674	0.64444	0.01531	<0.0001
0.35	4.70661	0.15543	0.70248	0.01166	<0.0001
0.40	4.34091	0.19906	0.74242	0.01325	<0.0001
0.45	3.66000	0.22753	0.80000	0.01458	<0.0001
0.50	3.24493	0.20186	0.84058	0.01275	<0.0001
0.55	2.93750	0.23661	0.87500	0.01322	<0.0001
0.60	2.92105	0.24349	0.89474	0.01337	<0.0001
0.65	2.80536	0.25500	0.91964	0.01416	<0.0001
0.70	2.56765	0.27858	0.95588	0.01462	<0.0001
0.75	2.83401	0.29314	0.96599	0.01483	<0.0001
0.80	2.92683	0.36179	0.99187	0.02013	<0.0001
0.85	2.36667	0.40258	1.06410	0.02215	<0.0001
0.90	2.09836	0.45491	1.13115	0.02620	<0.0001
0.95	0.73125	0.56584	1.28125	0.03685	<0.0001

Tablica 3.2: u Tablici su prikazani 95% pouzdani intervali za procijenjene β_0 i β_1

Kvantili	95% pouzdani interval za β_0	95% pouzdani interval za
0.05	[6.6594, 7.5328]	[0.3680, 0.4202]
0.10	[6.4492, 7.2824]	[0.4241, 0.4785]
0.15	[5.9519, 7.0847]	[0.4724, 0.5373]
0.20	[5.7878, 6.8448]	[0.5198, 0.5675]
0.25	[5.5148, 6.6161]	[0.5533, 0.6138]
0.30	[4.7371, 6.0758]	[0.6012, 0.6828]
0.35	[4.0686, 5.5459]	[0.6612, 0.7427]
0.40	[3.4628, 4.8489]	[0.7072, 0.7905]
0.45	[2.9289, 4.3603]	[0.7598, 0.8409]
0.50	[2.3669, 3.8176]	[0.8025, 0.8915]
0.55	[2.3949, 3.6728]	[0.8384, 0.9139]
0.60	[2.2340, 3.4405]	[0.8617, 0.9360]
0.65	[2.1920, 3.2552]	[0.8846, 0.9622]
0.70	[2.2584, 3.1104]	[0.9182, 0.9763]
0.75	[2.1074, 3.3820]	[0.9339, 1.0050]
0.80	[1.6684, 3.5676]	[0.9556, 1.0812]
0.85	[1.6183, 3.2987]	[0.9928, 1.1255]
0.90	[1.0565, 3.1568]	[1.0644, 1.1889]
0.95	[-0.5730, 2.9793]	[1.1534, 1.3707]

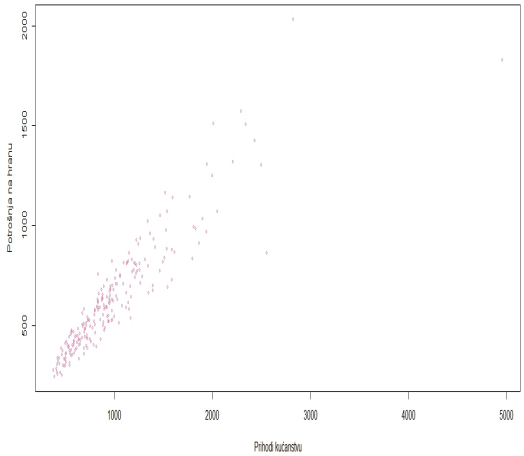
Sljedeći graf koji bi nas mogao zanimati je upravo graf procijenjene uvjetne funkcije gustoće.



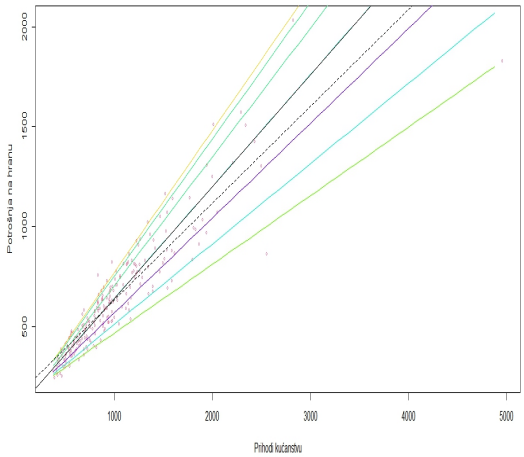
Ako dobro promotrimo graf možemo vidjeti da graf uvjetne funkcije gustoće kod većih temperatura postaje bimodalna funkcija gustoće, a to je upravo ono što smo zaključili iz prvog grafa.

Ovakvo ponašanje ima i meteorološko objašnjenje. Naime, povišeni tlak zraka iz unutrašnjosti kontinenta koji nosi vruće vrijeme mora u konačnici sresti hladnu frontu nastalu iznad Tasmanijskog mora, što uzrokuje nagli pad temperature. I u financijama i ekonomiji postoji mogućnost za ovakvim ponašanjem vremenskog niza, ali modeli koji se tamo koriste nisu u mogućnosti uočiti takvo ponašanje.

Primjer 3.4.2. *Drugi primjer koji ćemo promatrati je kako količina prihoda u kućanstvo belgijske radničke klase utječe na potrošnju za hranu. Prvo pogledajmo graf na kojem su prikazani podaci o potrošnji na hranu u odnosu na prihode kućanstva.*



Iz grafa možemo vidjeti da se podaci donekle linearno ponašaju i možemo primijetiti da ima puno više stanovnika sa manjim prihodima nego onih sa većim prihodima. Sada procijenimo kvantilne funkcije.



Na grafu su prikazane kvantilne regresijske funkcije za $\tau = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ *i linearna funkcija dobivena metodom najmanjih kvadrata koja je prikazana crnom isprekidanom linijom. Pogledamo li funkciju za medijan i usporedimo ju sa funkcijom dobivenom linearnom regresijom možemo vidjeti da se razlikuju.*

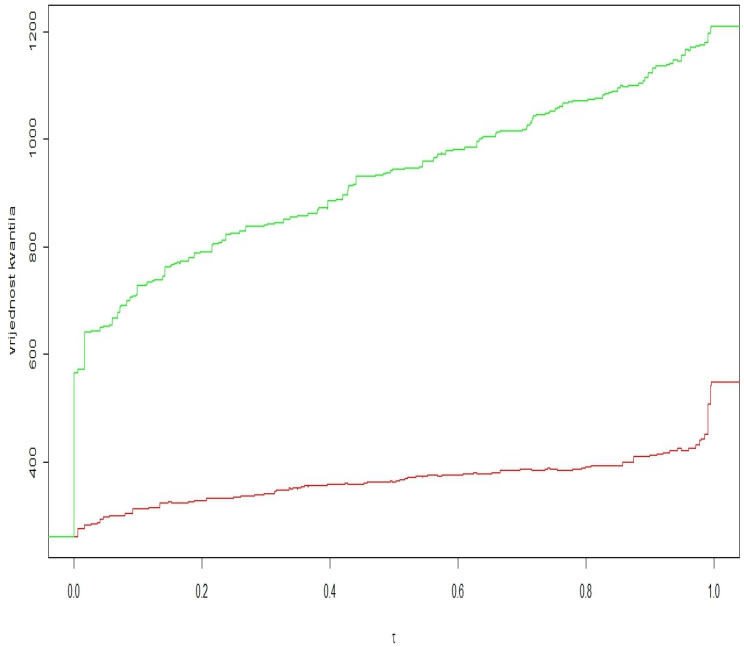
Tablica 3.3: U tablici su prikazani procijenjeni parametri za kvantilne pravce $y_i = \beta_0 + \beta_1 x_i$, njihova standardna greška i p-vrijednost dobivena Waldovim testom

Kvantili	β_0	Std.Error	β_1	Std. Error	p-vrijednost Waldovog testa
0.05	124.88004	16.6480	0.34336	0.0188	<0.0001
0.10	110.14157	25.2550	0.40177	0.0308	<0.0001
0.25	95.48354	20.5106	0.47410	0.0235	<0.0001
0.50	81.48225	28.0843	0.56018	0.0313	<0.0001
0.75	62.39659	29.7268	0.64401	0.0339	<0.0001
0.90	67.35087	26.4473	0.68630	0.0285	<0.0001
0.95	64.10396	22.6856	0.70907	0.0246	<0.0001

Tablica 3.4: u Tablici su prikazani 95% pouzdani intervali za procijenjene β_0 i β_1

Kvantili	95% pouzdani interval za β_0	95% pouzdani interval za
0.05	[92.0802, 157.6798]	[0.3063, 0.3804]
0.10	[60.3843, 159.8990]	[0.3410, 0.4625]
0.25	[55.0735, 135.8934]	[0.4277, 0.5205]
0.50	[26.1508, 136.8139]	[0.4985, 0.6218]
0.75	[3.8287, 120.9642]	[0.5772, 0.7109]
0.90	[15.2445, 119.4573]	[0.6302, 0.7424]
0.95	[19.4089, 108.7993]	[0.6606, 0.7576]

Ako razmislimo o podacima koji su prikupljeni, moglo bi nas zanimati postoji li razlika između procijenjene kvantilne funkcije bogatog stanovništva i procijenjene kvantilne funkcije siromašnog stanovništva.



Zelena linija prikazuje kvantilnu funkciju bogatog stanovništva (koji je dobiven pomoću 0.9*-tog kvantila podataka tako da su uzeti gornjih 10% podataka) dok crvena linija prikazuje siromašno stanovništvo (koje predstavlja donjih 10% podataka). Kao što je bilo i za očekivati bogato stanovništvo ima veće kvantile nego siromašno, odnosno kvantilna funkcija bogatih se nalazi iznad kvantilne funkcije siromašnih. Još jedna stvar koja bi nas mogla zanimati je da li su kvantilni regresijski pravci za različite vrijednosti od* τ *paralelni. To možemo ispitati pomoću funkcije* `anova` *u R-u, gdje zbog malih p-vrijednosti odbacujemo pretpostavku o jednakosti koeficijenata smjera kvantilnih regresijskih pravaca.*

Bibliografija

[1] Allen, David E. i Abhay Kumar Singh: *Minimizing Loss at Times of Financial Crisis: Quantile Regression as a Tool for Portfolio Investment Decisions*. School of Accounting, Finance and Economics & FEMARC Working Paper Series, 2009.

[2] Koencker, Roger i Gilbert Bassett: *On Boscovich's estimator*. The Annals of Statistics, 13:1625–1628, 1978.

[3] Koencker, Roger i Gilbert Bassett: *Regression Quantiles*. Econometrica, 46:33–50, 1978.

[4] Koenker, R.: *Quantile Regression*. Cambridge University Press, 2005.

[5] Stigler, Stephen M.: *The History of Statistics, The Measurment of Uncertainty befor 1900*. The Belknap press of Harvard University press, 1986.

Sažetak

Linearna regresija je postupak u kojem pokušavamo opisati zavisnu varijablu kao linearnu kombinaciju jedne ili više nezavisnih varijabli. Ona procjenjuje uvjetno očekivanje zavisne varijable uz danu nezavisnu varijablu te ju prikazuje kao afinu funkciju od zavisne varijable. Za taj tip regresije metoda najmanjih kvadrata je dovoljno učinkovita, ali problemi nastaju kada se žele proučavati ekstremi nekih podataka ili njihovi kvantili. Tu do izražaja dolazi kvantilna regresija, gdje kvantilni regresijski parametar predstavlja promjenu zavisne varijable (varijable odaziva) u određenom kvantilu za jednu jedinicu promjene nezavisne varijable (varijable poticaja).

U kvantilnoj regresiji procjenitelj za uvjetni medijan je procijenjen minimiziranjem simetrične težinske sume apsolutnih pogrešaka (gdje je težina jednaka 0.5), a procjenitelj za druge uvjetne kvantilne funkcije se procjenjuje minimiziranjem asimetrične težinske sume apsolutnih pogrešaka (gdje su težine jednake funkcijama od kvantila). Zbog toga je kvantilna regresija robustna na outliere što smo objasnili u radu.

Još jedna prednost kvantilne regresije koju smo prikazali u radu je svojstvo ekvivarijance koje nam omogućuje da imamo da je kvantil transformirane slučajne varijable (transformirana pomoću strogo rastuće afine funkcije) jednak transformiranom kvantilu originalne slučajne varijable, što nije slučaj kod matematičkog očekivanja. Zbog toga je interpretacija modela kvantilne regresije jednostavnija nego ona kod metode najmanjih kvadrata.

Summary

Linear regression is a process in which we try to describe the dependent variable as a linear combination of one or more independent variables. It estimates the conditional expectation of the dependent variable for a given independent variable and it is shown as an affine function of the dependent variable. For this type of regression least squares method is efficient, but problems arise when we want to study the extreme values of some data or their quintiles. This is where quantile regression is useful. The quantile regression parameter represents the change in the dependent variable (response variable) in a specified quantile of a unit change in the independent variable (predictor variable).

In quantile regression, the conditional median estimator is estimated by minimizing the symetrically weighted sum of absolute error (where the weight is equal to 0.5). The estimator for other conditional quantile function is estimated by minimizing the asymmetric weighted sum of absolute errors (where the weights are the functions of quintiles). Therefore, the quantile regression is robust to the presence of outliers as explained in this paper.

Another advantage of quantile regression, shown in this paper, is equivariance that allows the quantile transformed random variables (transformed by strictly increasing affine functions) to be equal to the transformed quantiles of the original random variables, which is not the case with mathematical expectation. Therefore the interpretation of the model of quantile regression is simpler than in the method of least squares.

Životopis

Rođena sam 16. rujna 1990. godine u Zagrebu. U školu sam krenula 1995. godine u Londonu gdje sam se školovala do 1999. godine. Nakon povratka u Zagreb nastavila sam svoje školovanje u osnovnoj školi Dr. Ante Starčevića. Nakon završene osnovne škole upisujem XV. gimnaziju. Godine 2009. upisujem Preddiplomski sveučilišni studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Nakon završenog preddiplomskog studija, 2012. godine, upisujem Diplomski sveučilišni studij Matematička statistika na istom fakultetu.